AS DISTRIBUIÇÕES DO ACASO

F. G. Brieger
Seção de Genética
Escola Superior de Agricultura
"Luiz de Queiroz", Universidade
de S. Paulo

ÍNDICE

PREFACIO

O presente trabalho trata das bases fundamentais da teoria da variação de acaso, e assim tornou-se inevitável entrar em discussões que requerem algumas noções de matemática. Porem para que o trabalho seja acessivel para um maior número de leitores, limitei a aplicação de termos matemáticos ao mínimo necessário.

Um conhecimento das definições matemáticas e das relações entre as diferentes distribuições do acaso é indispensável para os técnicos que querem especializar-se profundamente na matéria e idealizar novos métodos da análise. Mas devemos acentuar que êstes conhecimentos profundos são dispensáveis para a grande maioria daqueles que pretendem apenas aplicar os métodos da análise estatística em seus estudos experimentais. Para isto basta de modo geral seguir as normas estabelecidas e os testes recomendados por técnicos tão reconhecidos como KARL PEARSON, STUDENT, R. A. FISHER e outros.

Uma situação análoga encontramos em quase todos os trabalhos experimentais. Continuamente o especialista em determinado assunto precisa aplicar métodos de outros ramos, também especializados. Assim, o citologista, colorindo as suas lâminas com violeta genciana, com hematoxilina ou outros corantes, não precisa saber a fórmula ou a origem dêsses compostos, nem tampouco detalhes químicos do processo da coloração, detalhes êstes muitas vezes desconhecidos. Um químico, que quer apenas determinar a porcentagem de açúcar em extratos ou valor do pH em soluções, aplica em geral métodos químicos ou físicos, dispensando os conhecimentos da teoria ótica da polarização da luz, da teoria físico-química da mudança de côr dos indicadores e dos problemas fundamentais da teoria da condutibilidade elétrica e da ionização.

De um modo geral, um especialista deve saber tôdas as teorias da sua especialidade, mas quando precisa aplicar métodos de outras especialidades, êle pode e deve confiar na competência daqueles que elaboraram os respectivos métodos.

O presente trabalho é portanto destinado para os estatísticos e não ara os técnicos que querem apenas aplicar os métodos de análise na experimentação geral.

1 — INTRODUÇÃO (1)

Nos últimos quarenta anos a aplicação de novos métodos de análise estatística recebeu um grande impulso e também uma extensão pelas descobertas da escola londrina de KARL PEARSON, STUDENT e R. A. FISHER. Notamos, nos trabalhos mais recentes de estatística, nitidamente dois grupos: um que continua a dar quase que exclusivamente os métodos anteriores às descobertas de PEARSON (1900) e outro que apresenta os métodos modernos sem porém explicar a história de sua descoberta e a sua relação com os métodos da estatística anterior, nem mesmo as suas relações entre si. É o fim desta publicação mostrar a relação entre alguns dos diferentes métodos, modernos e velhos, e dos testes dêles derivados. Estes últimos, especialmente, são em geral explicados numa forma matemática tão alterada, que não é fácil entender a sua derivação nem as suas relações lógicas com os demais testes.

O fim de tôda a análise estatística é mostrar se valores observados correspondem aos valores que podem ser calculados de acôrdo com uma teoria ou hipótese. Devemos de um modo geral decidir estatisticamente se a diferença entre êstes

---

(1) Parece-me indicado no início do trabalho dar algumas explicações sôbre os símbolos que serão usados, pois até hoje não existe nenhum acôrdo científico sôbre o assunto.

Valores variáveis indicaremos com qualquer letra: v ou a, etc.. As médias serão caracterizadas por traços horizontais sôbre as letras: $\bar{v}$ $\bar{a}$, etc.. Quando forem determinadas as médias parciais e gerais, as primeiras serão indicadas por um traço: $\bar{v}$ e as médias gerais por dois traços $\bar{\bar{v}}$.

A dificuldade maior encontramos quando queremos definir os símbolos relativos à frequência das variáveis. Temos em geral que distinguir o número total (N) de variáveis e o número de variáveis livres, que em outras publicações indiquei sempre pelas letras duplas *nf*. Mas para simplificar a impressão, especialmente quando temos ainda que usar sufixos com nf_1, nf_a, etc., usarei em geral neste trabalho apenas a letra *n* para indicar o grau de liberdade, recorrendo apenas à forma *nf* quando há perigo de mal entendidos.

Nas referências às fórmulas da estatística clássica, isto é, especialmente na introdução, usarei a letra *n*, sem distinguir claramente entre o número total de variáveis e o seu grau de liberdade. Uma vez que estas discussões se referem apenas a amostras grandes, êstes dois conceitos podem, sem prejuizo, ser confundidos.

dois valores, observado e esperado, medido pelo valor do respectivo êrro standard, pode ou não ser atribuida ao acaso.

Denominando o valor observado por **v**, o valor ideal por v^o, ou quando êle é representado por uma média aritmética por **v**, e usando finalmente a letra grega σ para o êrro standard da variável v, temos assim a fórmula do desvio relativo simples:

$$D = \frac{v - v_o}{\sigma} \quad ou \quad = \frac{v - \bar{v}}{\sigma} \qquad \text{---------- 1—1}$$

Supoz-se que a variação dêste têrmo seguiria sempre a fórmula matemática que caracteriza a distribuição normal ou de GAUSS, descoberta independentemente por DE MOIVRE, LA PLACE e GAUSS. É o mérito dos autores já mencionados ter demonstrado que a chamada distribuição "normal" é apenas um caso, muito especial, entre um número grande de outras distribuições do acaso. Para melhor demonstrar em que consiste esta particularidade da distribuição de GAUSS, vamos dar em primeiro lugar ràpidamente a derivação da fórmula matemática da mesma.

Supomos que temos um aparêlho de acaso de GALTON, bem construido, no qual um número de bolas de aço, caindo verticalmente, encontra no seu caminho **n** pregos arranjados de tal modo que a probabilidade de ricochetear para a direita é **p** e para a esquerda **q**. Podemos então determinar as frequências de tôdas as combinações de desvios para a direita ou para a esquerda, calculando os têrmos de um binômio:

$$(p+q)^n = q^n + \frac{n!}{(n-1)!} \cdot p \cdot q^{n-1} + \frac{n!}{2!(n-2)!} \cdot p^2 \cdot q^{n-2} \cdots\cdots$$

$$\text{---------- 1—2}$$

De um modo geral podemos exprimir a frequência y de um têrmo **m** nesta série, pela equação:

$$y = \frac{n!}{m!(n-m)!} \cdot p^m \cdot q^{n-m} \qquad \text{---------- 1—3}$$

sendo a frequência máxima dos termos centrais das séries binomiais:

$$y(max) = \frac{n!}{np!\,nq!} \cdot p^{np} \cdot q^{nq} \quad \text{---------} \quad \text{— — — —} \quad 1—4$$

Podemos, de outro lado, dar a frequência de qualquer têrmo **m**, usando para a sua caracterização a sua distância **x** do têrmo central, pelas fórmulas:

$$m = np - x$$

$$n - m = n - (np - x) = nq + x$$

sendo consequentemente

$$y = \frac{n!}{(np-x)(nq+x)} \cdot p^{np-x} \cdot q^{nq+x} \quad \text{---------} \quad \text{— — — —} \quad 1—5$$

Aplicando agora alguns princípios do cálculo infinitesimal, podemos transformar esta equação geral, para obter a solução do caso especial no qual n atinge o valor infinito. Os gráficos na fig. 1 explicam porem ainda melhor o que acontece quando o valor do expoente **n** do binômio cresce. Estão reproduzidos quatro histogramas, que correspondem a valores escolhidos de n = 4, 16, 36 e 64. O êrro standard, calculado pela fórmula.

$$\sigma = \pm \sqrt{p \cdot (1-p) n}$$

tem então os valores: ± 1, ± 2, ± 3 e ± 4. As frequências de cada classe são sempre indicadas acima de cada coluna do histograma. Nota-se que com o crescimento de valor de **n**, aumenta o número de classes e no mesmo tempo diminui a diferença da altura de colunas seguidas.

Superpostas aos histogramas dos binômios foram desenhadas curvas "normais" ou de GAUSS. Quando n = 64, o histograma já acompanha a curva razoavelmente bem.

Em baixo, na figura 1, são dadas as frequências (ordenadas) da curva de GAUSS com classes de 0,25 do êrro standard. Se nós compararmos estes valores com as frequências do binômio $(1/2 + 1/2)^{64}$, podemos constatar uma concordância bem acentuada.

A transformação matemática do binômio para a curva da distribuição de GAUSS é dada em detalhe no tratado geral de Estatística de YULES AND KENDALL (1940, pg. 177-180) e chega-se finalmente à seguinte fórmula simples para as ordenadas da distribuição de GAUSS:

$$y = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\frac{x^2}{p.q.n}} \qquad \text{------.---} \quad 1\text{—}6$$

Podemos ainda substituir alguns dos têrmos nesta equação. Lembrando que definimos acima x como a distância de m do têrmo central, podemos tambem escrever em têrmos da variação em relação a uma média $\bar{v}$.

$$x = v - \bar{v}$$

O têrmo p. q. n. é a variança de um binômio, e devemos substituí-lo pela letra σ que caracteriza em geral o êrro standard de uma distribuição. Assim chegamos finalmente à fórmula:

$$y = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{v-\bar{v}}{\sigma}\right)^2} = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}D^2} \qquad \text{---.---} \quad 1\text{—}7$$

Esta é a equação que define a curva normal ou de GAUSS, a qual pode ser construida fàcilmente, calculando-se as ordenadas y para tôdas as abcissas D pela fórmula 1-7.

A condição principal que permitirá chegar, da equação das frequências do termo binomial (fórmula 1-5) à equação das ordenadas da curva de GAUSS (fórmula 1-7) consiste em supor que o número n atingiu o valor infinito.

Dissemos acima que o número n é o número de pregos do aparêlho de GALTON que uma bolinha encontra no seu caminho e que alteram a sua marcha. Êstes pregos representam os fatores de acaso que podem afetar um acontecimento qualquer. Não é suficiente porém que o número de pregos seja grande, se nós tambem não usarmos ao mesmo tempo um número elevado de bolinhas. Se houvesse poucas bolinhas, elas apenas encontrariam alguns pregos no seu caminho, mas para que todos os pregos entrem em jôgo, o número de bolinhas deverá

também ser grande. Assim, os pregos representam os fatores de acaso que atuam sôbre um acontecimento, neste caso a marcha das bolinhas, e chegamos à conclusão que ambos, os pregos e bolinhas, fatores de acaso e acontecimentos, devem ser muito numerosos, aproximando-se o seu número ao infinito.

O têrmo "acontecimento" também devemos ainda definir melhor em sua significação geral. Se por exemplo estudarmos a altura de plantas, ou o pêso de animais, etc.,a altura ou o pêso de cada indivíduo é um tal "acontecimento" que depende de um grande número de causas incontroláveis ou de acaso.

Para poder aplicar a fórmula da distribuição de GAUSS devemos estar certos de que as nossas observações abrangem tôdas as combinações possíveis dos fatores de acaso e que nenhuma delas ficou sem ser registrada, em consequência de uma limitação do número de observações. Podemos ter esta certeza quando o número de observações é tão grande de modo que não foi feita uma seleção qualquer, dirigida ou não, e que poderia resultar numa omissão de alguns acontecimentos. Para ter certeza que o jôgo do número infinitamente grande de agentes de acaso era absolutamente livre e ilimitado e que tudo foi também registrado, o número de observações também deverá ser infinitamente grande.

Como se pode deduzir da fórmula 1-6 e 1-7, o êrro standard é o único têrmo na equação, onde o valor n ainda aparece. Assim, podemos frizar que uma variação de acaso segue a norma estabelecida pela distribuição de GAUSS sòmente quando, no têrmo $D = (v - \bar{v}) : \sigma$ o valor do êrro standard for derivado de um número ilimitado ou infinitamente grande de observações.

Em geral, em nossos experimentos, o número de observações ou de varáveis é pequeno, e além disso introduzimos frequentemente durante o processo da análise limitações que restringem a liberdade da variação. Na análise de séries de observações sôbre variáveis em número limitado e restringido, isto é, onde não sòmente o número total, mas também o número de variáveis livres ou o seu **grau de liberdade** é reduzido, não podemos esperar que a variação siga a distribuição de GAUSS, e as frequências em tais séries não corresponderão às ordenadas desta distribuição. Devemos procurar a definição de outras distribuições do acaso nas quais o jôgo do acaso não é mais completamente livre, mas onde há uma limitação da variação e do número de variáveis livres.

2 — A ANÁLSE DO ÊRRO STANDARD E AS CONCLUSÕES DE LEXIS E DE PEARSON

Pelo que dissemos acima e de acôrdo com a equação 1-6 e 1-7, devemos em primeiro lugar discutir a variação de um êrro standard e a sua análise.

Sempre que o valor do êrro standard de uma distribuição é calculado de um número limitado de observações, temos que admitir que o valor calculado não é uma constante, mas que êle mesmo é variável no sentido de que determinações repetidas darão valores diversos, todos êles sendo diferentes do valor ideal desconhecido. Assim, é necessário achar uma médida da variação do êrro standard. Desde ha muito se conhece uma fórmula correspondente ao "êrro do êrro" e a sua derivação matemática é dada no tratado de **Yules and Kendall**(1940) pags. 395-401), sendo êle definido pela equação:

$$\sigma[\sigma] = \frac{\sigma}{\sqrt{2n}} \sqrt{\frac{\beta_2 - 1}{2}} \quad \text{---- ---- 2—1}$$

sendo o índice da excessividade ou curtosis:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\sum(v-\bar{v})^4}{\left(\sum(v-\bar{v})^2\right)^2}$$

O valor dêste índice é igual a 3 em certos casos, e especialmente no caso da distribuição de **Gauss**. Sòmente nêstes casos temos :

$$\frac{\beta_2 - 1}{2} = \frac{3-1}{2} = 1$$

Desaparecendo a raiz na equação (2-1), podemos escrever:

$$\sigma\,[\sigma] = \pm \frac{\sigma}{\sqrt{2n}} \quad \text{---------- 2—2}$$

para: $\beta_2 = 3$, o que se obtem na distribuição de **Gauss** e quan-

do n aproxima-se ao infinito. Nêste caso especial, poderíamos estudar a variação de um êrro standard observado, em relação ao seu valor ideal ou médio, dividindo a diferença pelo êrro :

$$\frac{\sigma - \sigma_0}{\sigma_0 : \sqrt{2n}} = \frac{\sigma - \sigma_0}{\sigma_0} \cdot \sqrt{2n} = D$$

A variação dêste têrmo deve seguir a distribuição de Gauss, pois n aproxima-se ao infinito.

Agora, o têrmo mencionado pode facilmente ser transformado :

$$D = \left(\frac{\sigma}{\sigma_0} - 1\right) \cdot \sqrt{2n}$$

$$\frac{\sigma}{\sigma_0} = 1 \pm \frac{D}{\sqrt{2n}} \qquad \text{2—3}$$

Esta fórmula muito importante podemos também exprimir em palavras da seguinte maneira : **O quociente de um êrro standard e seu valor ideal deve ser igual a 1, mais ou menos as abcissas D de uma distribuição de Gauss, divididas pela raiz quadrada de 2n.** Mas nunca devemos esquecer que as fórmulas (2-2) e (2-3) podem ser aplicadas apenas quando o valor n é tão grande que êle aproxima-se ao infinito, pois no caso contrário não pode ser admitido sem prova que o índica β_2 seja igual a 3, nem que nós possamos usar a distribuição de Gauss.

E' interessante notar que aparece na fórmula (2-3) o quociente, em vez da diferença entre o valor observado e o esperado. Êste quociente foi usado pela primeira vez por LEXIS (1872). Porém, como se pode ver pela exosição no tratado de YULE AND KENDALL (1940 pp. 368-369), LEXIS estudou a sua "razão de dispersão" apenas em casos especiais e também não considerou qual a distribuição de acaso que deve ser aplicada.

Parece que a fórmula matemática de variação de um êrro standard, respectivamente de uma variança, foi estudada por HELLMERT (1876, de acôrdo com KENNY, Mathematical Satistics, 1939, pg. 128). O problema foi resolvido apenas mais tarde por KARL PEARSON (1900). Explicamos a solução seguindo uma marcha diferente daquela que nos conduziu á equação (2-3), e que também difere do caso muito especial discutido e solucionado por PEARSON.

Como explicámos acima, a variação de uma variável pode ser medida pelo seu desvio relativo $D = (v-\bar{v}) : \sigma$ e podemos assim analisar vários desvios isoladamente. Querendo porém resolver, se um conjunto dêles pode ser causado pelo acaso, nós reunimos os seus quadrados, seguindo depois dois caminhos que são apenas algebricamente diferentes. Dividindo a soma pelo quadrado do êrro ideal e extraindo a raiz obtemos o valor que denominamos X:

$$X = \sqrt{\frac{\Sigma(v-\bar{v})^2}{\sigma_0^2}}$$

ou nós dividimos ainda pelo grau de liberdade correspondente nf, obtendo um desvio relativo:

$$D = \sqrt{\frac{\Sigma(v-\bar{v})^2}{nf} : \sigma_0^2} = \frac{\sigma_1}{\sigma_0} \quad \text{---------- 2--4}$$

sendo assim: $D = X : \sqrt{nf}$

PEARSON (1900) determinou a fórmula da variação do têrmo X e nós chamamos **distribuições de Pearson** às distribuições correspondentes de

$$D' = \frac{X}{\sqrt{nf}}$$

Na fórmula correspondente aparece o valor **nf** do grau de liberdade do conjunto de desvios, de modo que para cada valor de **nf** temos uma distribuição. Existe apenas uma distribuição de GAUSS, porém muitas distribuições de PEARSON.

A descoberta de PEARSON (1900) permitiu depois a STUDENT (1917) achar a solução do caso da variação de um só desvio, quando o êrro standard correspondente é apenas baseado em poucas observações, sendo relativamente pequeno o grau de liberdade, correspondente ao êrro.

Finalmente, R. A. FISHER deu a solução final e bem geral do problema da variação de desvios, seja individual ou em conjunto, achando a fórmula geral de variação do acaso de um quociente de dois êrros standard, cada um com graus de liberdade n1 e n2 diferentes:

$$D = \frac{\sigma_1}{\sigma_2} \qquad \frac{n_1 = \ldots}{n_2 = \ldots}$$

3 — A DEFINIÇÃO MATEMÁTICA DOS QUATRO TIPOS DE DISTRIBUIÇÃO DE ACASO

a) As fórmulas das distribuições

As fórmulas que definem os quatro tipos de distribuição de acaso — GAUSS, PEARSON, STUDENT e FISHER — permitem calcular as ordenadas para as respectivas abcissas, sendo as ordenadas frequências e as abcissas desvios relativos. A maioria dos tratados de estatística não apresenta as fórmulas na sua forma mais geral. Os testes derivados e as tabelas correspondentes contêm os valores de abcissas, que correspondem a diferentes limites de probabilidade, em forma bastante alterada na sua expressão algébrica. Demonstraremos em seguida as relações matemáticas das distribuições, na sua forma mais geral.

Nos desvios relativos, podemos distinguir o grau de liberdade do numerador n1 e do denominador n2 e notamos neste ponto a situação seguinte:

<u>*Fisher*</u>

$1 < nf_1 < infinito$

$1 < nf_2 < infinito$

<u>*Student*</u>

$nf_1 = 1$

$1 < nf_2 < infinito$

<u>*Pearson*</u>

$1 < nf_1 < infinito$

$nf_2 = infinito$

<u>*Gauss*</u>

$nf_1 = 1$

$nf_2 = infinito$

Limitamos a definição das distribuições de FISHER aos graus de liberdade maiores do que 1 e menores do que infinito. Quando o grau de liberdade n1 do numerador é igual a 1 e aquêle do denominador qualquer valor maior do que 1 e menor do que infinito, encontramos uma das distribuições de STUDENT. De outro lado, quando os limites do grau de liberdade n1 são os mesmos como na distribuição de FISHER, e os graus de liberdade n2 são infinitamente grandes, então trata-se de uma das distribuições de PEARSON. Com outras palavras, se na distribuição de FISHER alterarmos apenas o grau de liberdade do numerador para 1, ficando sem alteração o valor de n2, passamos para uma distribuição de STUDENT. E se não mudarmos os limites de n1, ficando n2 infinitamente grande, obtem-se uma das distribuições de PEARSON. Finalmente, quando fizermos concomitantemente ambas as alterações (n1 = 1/ n2 = infinito) obtemos a distribuição normal ou de GAUSS.

Nas tabelas que são frequentemente usadas para o Z-teste (FISHER), F-teste (SNEDECOR) e ϑ -teste (BRIEGER), a definição das distribuições é mais geral, incluindo os extremos e dando os valores críticos dos testes não sòmente das distribuições de FISHER, mas para tôdas as quatro distribuições, FISHER, STUDENT, PEARSON e GAUSS simultaneamente.

Nas definições matemáticas dos quatro tipos de distribuição podemos distinguir duas componentes: uma que depende dos valores dos graus de liberdade apenas e por isso é constante para cada distribuição, e outra que contém o valor do desvio relativo D do qual depende o valor da frequência y.

Distribuições de Fisher

Grão de liberdade

$$y = k_1 \cdot D^{n_1 - 1} \cdot \left(1 + \frac{n_1}{n_2} \cdot D^2\right)^{-\frac{n_1 + n_2}{2}} \qquad 3 - 1$$

$1 < n_1 < inf.$

$1 < n_2 < inf.$

$$k_1 = \frac{\frac{n_1 + n_2 - 2}{2}!}{\frac{n_1 - 2}{2}! \cdot \frac{n_2 - 2}{2}!} \cdot 2 \cdot \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \qquad 3 - 1$$

Distribuições de Pearson

Grão de liberdade

$1 < n_1 < inf$

$n_2 = inf.$

$$y = k_2 \cdot D^{n_1 - 1} \cdot e^{-\frac{n_1}{2} D^2}$$

$$k_2 = \frac{n_1^{\frac{n_1}{2}}}{\frac{n_1 - 2}{2}! \, 2^{\frac{n_1 - 2}{2}}} \qquad 3—2$$

Distribuições de Student

Grão de liberdade

$n_1 = 1$

$1 < n_2 < inf.$

$$y = k_3 \cdot \left(1 + \frac{1}{n_2} D^2\right)^{-\frac{n_2 + 1}{2}}$$

$$k_3 = \frac{\frac{n_2 - 1}{2}!}{\frac{n_2 - 2}{2}! \sqrt{n_2 \pi}} \qquad 3—3$$

Distribuições de Gauss

Grão de liberdade

$n_1 = 1$

$n_2 = inf.$

$$y = k_4 \cdot e^{-\frac{1}{2} D^2}$$

$$k_4 = \frac{1}{\sqrt{2\pi}} \qquad 3—4$$

b) Cálculo das ordenadas e áreas

As fórmulas dadas acima permitem o cálculo de ordenadas de curvas com área unidade, e tendo como unidade da abcissa o êrro standard.

Precisamos explicar o que êstes últimos têrmos significam e também como calculamos a área debaixo de uma curva, isto é, entre a curva e o eixo da abcissa. Para acharmos a área dividimos primeiro a curva de tal modo que as diferenças entre abcissas seguidas seja igual

$$x_1 - x_2 = x_2 - x_3 = x_3 - x_4$$

O tamanho dêstes intervalos será igual a uma fração do êrro standard que podemos calcular, dividindo o êrro standard por **m**, sendo **m** um número livremente escolhido. A área agora será determinada para cada abcissa aproximadamente, supondo-se que podemos considerar a área total da curva acima de cada intervalo como retângulo. Assim, temos a seguinte relação:

Abcissa	Área	
x_1	$y_1 \cdot \frac{\sigma}{m}$	Área total $\Sigma\left(y \cdot \frac{\sigma}{m}\right)$
x_2	$y_2 \cdot \frac{\sigma}{m}$	

Para uma distribuição com êrro standard e área unidade, esta fórmula se transforma na seguinte:

$$\text{Area unidade} = \Sigma \frac{y}{m} = 1 \quad \text{sendo} \quad \sigma = 1$$

Para construir os gráficos dêste trabalho, foram calculadas as ordenadas em intervalos de m = 1:10, do êrro standard.

c) Forma das curvas

Na figura 2 temos quatro distribuições, tôdas representadas em escala igual: a distribuição de GAUSS, a distribuição de STUDENT com n1 = 1 / n2 = 10, a distribuição de PEAR-

SON com n1 = 10 / n2 = infinito e a distribuição de FISHER para n1 = 10 / n2 = 10. No centro do gráfico temos a abcissa zero, separando os valores positivos e negativos para a direita e para esquerda. As ordenadas sempre são positivas.

O gráfico deixa aparecer algumas propriedades especiais das curvas.

Tôdas as distribuições, como representadas, são simétricas em relação ao ponto de origem ou valor **zero** da abcissa. Esta conclusão não precisa de nenhum comentário para as distribuições de STUDENT e de GAUSS. Porém, a referida simetria é apenas de ordem secundária para as outras distribuições que são descontínuas para o valor zero da abcissa, de modo que atualmente temos duas curvas completas e separadas, uma no lado positivo e outra no lado negativo do eixo horizontal. Cada uma destas curvas isoladas é assimétrica e o seu módulo ou máximo não fica longe da abcissa D = 1. É muito importante notar que a área de cada curva isolada, de PEARSON e de FISHER, seja no lado direito ou no esquerdo de zero, corresponde apenas à metade da área das distribuições de STUDENT e GAUSS.

d) Comparação matemática das quatro distribuições

Uma análise das fórmulas e dos gráficos (Fig. 2) resulta em estabelecer quatro principais diferenças:

I) Em primeiro lugar vamos determinar as ordenadas para a abcissa D igual a zero :

$$y\left[Gauss\right] = k_4 \cdot e^{-\frac{1}{2}D^2} = k_4$$

$$y\left[Student\right] = k_3 \cdot \left(1+\frac{1}{n_2}D^2\right)^{-\frac{n_2+1}{2}} = k_3$$

$$y\left[Pearson\right] = k_2 \cdot D^{n_1-1} \cdot e^{-\frac{n_1}{2}D^2} = 0 \ (zero)$$

$$y\left[Fisher\right] = k_1 \cdot D^{n_1-1} \cdot \left(1+\frac{n_1}{n_2}D^2\right)^{-\frac{n_1+n_2}{2}} = 0 \ (zero)$$

Nos primeiros dois casos, a ordenada y para a abcissa zero torna-se igual à constante da equação, k4 e k3. Mas nos outros dois casos a ordenada é zero para a abcissa zero. No gráfico isso de faz sentir na forma já mencionada. As curvas de GAUSS e de STUDENT são contínuas desde mais infinito (canto extremo direito do gráfico) até menos infinito (canto esquerdo). As outras duas curvas, de PEARSON e de FISHER, são descontínuas. Nos gráficos temos duas curvas, separadas; uma à direita e outra à esquerda do centro.

II) O valor do desvio relativo nas distribuções de GAUSS e STUDENT de um lado e naquelas de PEARSON e FISHER representam coisas bem diferentes.

No primeiro caso dividimos um só desvio pelo seu êrro, de modo que o quociente, que serve como abcissa no gráfico, tem um sinal definido, positivo ou negativo, de acôrdo com o sinal do desvio. O valor da ordenada y de outro lado, é independente do sinal de D, pois nas equações 3-3 e 3-4 apenas aparece o quadrado de D que é sempre positivo. Assim, as curvas nos gráficos são simétricas, com uma metade à direita do valor zero da abcissa para os valores positivos do desvio relativo e outra metade à esquerda para os valores negativos.

Nos casos das distribuições de FISHER e PEARSON, o valor do desvio relativo é um quociente de dois êrros, respectivamente de um êrro por seu valor ideal. A raiz no numerador, e consequentemente também o quociente D poderá ser tanto positivo como negativo. No cálculo como na ilustração gráfica devemos atribuir ao desvio relativo, ou sòmente o sinal positivo, ou então o sinal negativo. Obtemos assim, como na figura 1, ou uma curva completa à direita do valor zero da abcissa ou uma curva completa à esquerda.

As curvas de PEARSON e FISHER se encontram sòmente num ou noutro lado do ponto D igual a zero.

III) Em terceiro lugar vamos resolver a questão da posição do máximo das curvas.

Nas distribuições de GAUSS e STUDENT temos que dividir a constante k4 ou k3 por um têrmo exponencial que atinge o seu valor mínimo para D igual a zero, e que cresce para os valores maiores ou menores do que zero. A ordenada y atingirá assim o máximo quando o denominador do quociente é um mínimo, isto é, para a abcissa D igual a zero. O valor da ordenada y para esta abcissa já determinámos acima (sob I) de modo que podemos resumir, dizendo que as curvas de GAUSS e de STUDENT atingem o seu máximo com $y = k3$ e $y = k4$, quando D igual a zero.

Nas distribuições de PEARSON e de FISHER a situação é mais complicada, pois as constantes k2 e k1 respectivamente devem ser multiplicadas por um têrmo exponencial de D e divididas por outro têrmo exponencial de uma função de D. Quando D é pequeno, o primeiro têrmo cresce mais depressa do que o outro e assim y também cresce, mas depois e na medida que D aproxima-se ao valor 1, o segundo têrmo cresce mais depressa, e o valor de y decresce consequentemente. Explicaremos isso com um exemplo, a distribuição de FISHER com n1 = n2 = 10.

D	9 log. D	10 log (1+D²)	Dif. ou log (D⁹:(1+D²)¹⁰)	
0,5	0,28703 — 3	0,96910	0,31163 — 4	
0,6	0,00335 — 2	1,33540	0,66795 4	
0,7	0,60590 — 2	1,73190	0,87400 4	
0,8	0,12772 — 1	2,14840	0,97932 4	
0,9	0,58816 — 1	2,57680	0,01136 3	
1,0	1,00000	3,01300	0,98700 4	Máximo
1,1	0,37251	3,44390	0,92861 4	
1,2	0,71262	3,87390	0,83872 4	

$$\log. k_1 = \log.\left[\frac{9!}{4!\,4!}\cdot 2\right] = 3{,}10037$$

É evidente neste caso que a diferença entre os logarítmos decimais dos dois têrmos exponenciais e com isso de y, cresce até D = 0,9 caindo depois de novo. Aproximando melhor, achamos o máximo da diferença para D = 0,904, sendo o valor do logarítmo desta diferença igual á 0,01151-3. Assim, temos enfim : y = 0,1294.

Para a distribuição de FISHER com nf1 = nf2 = 10, temos a ordenada máxima de 0,1294 na abcissa D = 0,904.

Para a distribuição de PEARSON e FISHER, a ordenada máxima é diferente para cada combinação de graus de liberdade, sendo localizada em abcissas menores do que 1 e aproximando-se ao valor D = 1, quando os valores do grau de liberdade se aproximam ao infinito. Uma inspeção dos gráficos na Fig. 2 e nas Figs. 6 a 9 demonstra claramente esta conclusão.

IV) Finalmente, consideremos a questão dos chamados "limites de probabilidade". Em todos os testes da análise estatística aplicamos êste conceito, querendo saber se um desvio relativo está dentro ou fora dos respectivos limites.

A significação dêste têrmo "limite de probabilidade" podemos fàcilmente compreender se lembrarmos da definição da área debaixo de uma curva. Esta área é a soma de tôdas as ordenadas que por sua vez são as frequências com as quais são esperados os valores correspondentes da abcissa.

Se erguermos uma ordenada num ponto qualquer da extremidade positiva da curva, dividindo assim a área em duas partes, obteremos a situação seguinte: a soma das frequências de tôdas as abcissas menores daquela que serve como limite e na qual erguemos a ordenada, seja **a** e a frequência de tôdas as abcissas maiores do que o valor do limite **b**. Supomos que a abcissa foi escolhida de modo que **a** = 95% e **b** = 5% da área total, então a soma das frequências de abcissas menores foi 0,95 e a soma das frequências de abcissas maiores apenas 0,05, sendo a área total: 0,95 + 0,05 = 1,00. O valor correspondente das abcissas chamamos o 0,05 ou 5% **limite unilateral à direita.** ou superior (**Figs. 3 e 4**).

Se tivéssemos erguido a ordenada na extremidade esquerda da curva, então se trataria do **limite unilateral inferior**

Podemos erguer ordenadas em ambas as extremidades ao mesmo tempo, de modo que a área entre as ordenadas seria igual a: **a** e aquela fora de cada uma igual a: $b = \frac{1 - a}{2}$, ou no exemplo dado acima: $a = 0{,}95$ e $b = \frac{1 - a}{2} = 0{,}025$. Assim temos marcados os **limites bilaterais**, um à esquerda e outro à direita, e a probabilidade total de se obter abcissas maiores do que o limite à direita ou menores do que o limite à esquerda é igual a 0,05 ou 5%.

Uma representação gráfica dêstes três casos: limite unilateral à direita ou superior, limite unilateral à esquerda ou inferior e limites bilaterais para a distribuição de GAUSS e para a distribuição de FISHER com nf1 = nf2 = 10 encontramos nas figs. 3 e 4. Notamos nestes gráficos duas particularidades:

As abcissas que marcam os limites unilaterais são sempre menores do que aquelas dos limites bilaterais.

Na distribuição de GAUSS (Fig. 3), e o mesmo encontramos nas distribuições de STUDENT, tanto os dois limites unilaterais como também os dois limites bilaterais são de tamanho igual, mas com sinal oposto. Os limites à esquerda ou inferio-

res têm um sinal negativo e os limites à direita ou superiores têm o sinal positivo.

A situação é bastante diferente nas distribuições de PEARSON e de FISHER. (Fig. 4). Aqui não encontramos diferenças de sinais, pois os limites à direita são abcissas maiores do que 1 e os limites à esquerda são frações menores do que 1. Tratando-se de distribuições assimétricas, não existe uma relação matemática simples entre os valores de ambos os limites, exceto em alguns casos especiais que nós iremos mencionar apenas sem entrar numa discussão matemática.

a) Chamamos de distribuições recíprocas aquelas nas quais os graus de liberdade são trocados, como n1 = a / n2 = b e n1 = b / n2 = a,. Nestas distribuições recíprocas, os limites superiores de uma distribuição são os valores recíprocos dos limites inferiores das distribuições correspondentes.

b) Nas distribuições de FISHER com n1 = n2 os valores dos limites inferiores e superiores também são valores recíprocos.

c) Quando os valores dos graus de liberdade são grandes, as distribuições correspondentes tornam-se pràticamente simétricas, e os limites superiores e inferiores se tornam equidistantes do valor 1.

Analisando as distribuições de GAUSS ou de STUDENT usamos em geral limites que separam as extremidades em **ambos os lados da curva**, quando nas distribuições de PEARSON e FISHER consideramos apenas **uma extremidade**, isto é, o limite unilateral superior. Isto é de um modo geral bem justificável.

Quando estudamos um desvio simples D igual a um desvio dividido pelo seu êrro standard, admitimos que tanto valores positivos e negativos **devem** acontecer. Uma variação excessiva deve-se fazer sentir nos dois extremos da curva, tanto no lado positivo como no negativo.

Na distribuição de PEARSON, de outro lado, com n1 entre

um e infinito e n2 = infinito, comparamos os dois erros σ_1 e σ_0. O último êrro, derivado, seja de uma fórmula teórica ou de um número ilimitado de observações, representa então o êrro inevitável do experimento em aprêço.

Assim, um valor do quociente dos êrros maior do que 1, indica que a variação na amostra, que deu o valor do primeiro êrro, foi excessivo, em consequência de fontes de variação adicionais.

Um valor menor do que 1 não podemos esperar, nem precisamos tomar em consideração, pois se de fato o valor σ_0 é a medida da variação proveniente do acaso σ qual não pode ser excluida do experimento, um valor σ_1 , significantemente menor do que σ_0 não podia ser encontrado.

Seria apenas possível que a variação fôsse aumentada pela ação de fatores especiais, e assim justifica-se plenamente considerarmos sòmente uma extremidade da curva: aquela que corresponde a abcissas maiores do que 1. As tábuas até hoje publicadas se referem apenas aos limites de probabilidade unilaterais e superiores que indicam a improbabilidade de se obter valores da abcissa maiores do que aquêles da tábua.

Na distribuição de FISHER, a situação é muitas vezes a mesma, quando a natureza dos dados nos permite indicar qual o êrro que deve ser considerado como experimental e inevitável ou de acaso. Quando o outro êrro for estatisticamente menor do que o êrro experimental, isto indicaria uma de duas coisas: ou o êrro experimental não foi determinado com exatidão suficiente e a sua estimativa era errada e grande demais, ou na parte do experimento ao qual se refere o êrro menor,condições especiais causaram uma limitação sensível da variação.

Na genética, comparando a variação em famílias resultantes de um cruzamento, nas gerações F1 e F2, etc., temos uma variação que depende da variação fenotípica e outra causada pela segregação genotípica ou mendeliana. Esta última deve ter um máximo em F2 e um mínimo em F1, atingindo um grau diferente nas famílias de F3, F4, etc.. Assim, teremos uma variação dos erros standard, por família, em dois sentidos, como indicação da maior ou menor segregação mendeliana.

Em geral, quando comparamos a variação num número grande de amostras com um conjunto de muitos erros, devemos tomar em consideração ambos os limites: o limite superior, que nos indica quais os valores máximos de abcissas, maiores do que 1, como também os limites inferiores para as abcissas menores do que 1.

Os limites em tôdas as tábuas para as distribuições de FISHER até hoje publicados se referem sempre a uma extremidade apenas. Se precisamos tomar em consideração ambas as extremidades simultaneamente, não devemos esquecer que então os limites indicados nestas tábuas para o valor de 5%, 1% ou 0,1% de probabilidade (limites unilaterais), correspondem a 10%, 2% e 0,2% de probabilidade (limites bilaterais).

Considerando que em muitos casos é necessário empregar os limites unilaterais das distribuições de STUDENT e os limites bi-laterais das distribuições de PEARSON e FISHER, preparei novas táboas que já estão em impressão. Nessa publicação será explicado com mais detalhe o problema geral dos limites. (BRIEGER, 1945).

4 — A TRANSFORMAÇÃO MATEMATICA DAS 4 DISTRIBUIÇÕES

a) Noções matemáticas

Nestas transformações somos obrigados a recorrer a algumas noções gerais do cálculo matemático, as quais podemos resumir da maneira seguinte:

Nas fórmulas aparecem sempre têrmos fatoriais da forma geral (n-2) / 2! e (n-1) / 2!. Sendo os têrmos fatoriais uma forma de expressão da função gama, êles têm os seguintes valores em dependência do valor n:

Quando $(n-2):2$ ou $(n-1):2$ são numeros positivos inteiros:

$$\frac{n-2}{2}! = 1.2.3.... \frac{n-2}{2} \qquad \frac{n-1}{2}! = 1.2.3..... \frac{n-1}{2}$$

Quando $(n-2):2$ ou $(n-1):2$ tem como valores: 0,5 ou 1,5 ou 2,5...:

$$\frac{n-2}{2}! = 0{,}5 \cdot 1{,}5 ..... \frac{n-2}{2} \cdot \sqrt{\pi} \qquad \frac{n-1}{2}! = 0{,}5 \cdot 1{,}5 ..... \frac{n-1}{2} \cdot \sqrt{\pi}$$

Quando $(n-2):2$ ou $(n-1):2$ são iguais a 1 ou 0:(zero)

$$0! = 1 \qquad 1! = 1$$

Quando $(n-2):2$ é igual a $-0{,}5$:

$$(-0{,}5)! = \sqrt{\pi}$$

Finalmente devemos considerar a possibilidade de **n** se aproximar ao valor infinito, e neste caso podemos aplicar a fórmula de STIRLING:

$$\lim_{n = inf.} n! = \sqrt{2\pi} \cdot n^{n+\frac{1}{2}} \cdot e^{-n}$$

Outra fórmula de bastante importância, derivada da definição da base e dos logarítmos naturais, é:

$$\lim_{m = inf.} \left(1+\frac{a}{m}\right) = e^{\frac{a}{m}}$$

b) Transformação geral da parte exponencial

Como uma primeira informação daremos a transformação da parte exponencial das equações 3-1 a 3-4.

Quando o grau de liberdade n2 cresce até infinito temos a seguinte transformação:

$$\lim_{n_1 = inf} y\left[Fisher\right] = k_1 \cdot D^{n_1-1} \cdot \left(1+\frac{n_1}{n_2} D^2\right)^{-\frac{n_1+n_2}{2}}$$

$$= k \cdot D^{n_1-1} \cdot e^{-\frac{n_1}{n_2} D^2 \cdot \frac{n_1+n_2}{2}}$$

igualando n2 a (n2 + n1) quando n2 infinitamente grande, de modo que (n2 + n1) : n2 = 1, temos:

$$\lim_{n_2 = inf} y\left[Fisher\right] = k_2 \cdot D^{n_1-1} \cdot e^{-\frac{n_1}{2}D^2} = y\left[Pearson\right] \quad \text{---- 4—1}$$

Do outro lado, partindo-se da distribuição de PEARSON (3-2) e se nós atribuirmos a n1 o valor de unidade, teremos:

$$\lim_{n_1 = 1} y\left[Pearson\right] = k_2 \cdot D^{n_1-1} \cdot e^{-\frac{n_1}{2}D^2} = k_4 \cdot e^{-\frac{1}{2}D^2} = y\left[Gauss\right]$$

Atribuindo na equação da distribuição de FISHER (3-1) ao grau de liberdade n1 o valor de **um**, a equação se transforma na equação da distribuição de STUDENT (3-3):

$$\lim_{n_1 = 1} y\left[Fisher\right] = k_1 \cdot D^{n_1 - 1} \cdot \left(1 + \frac{n_1}{n_2} D^2\right)^{-\frac{n_1 + n_2}{2}} = k_3 \left(1 + \frac{1}{n_2} D^2\right)^{-\frac{n_2 + 1}{2}} = y\left[Student\right] \qquad \text{---- 4—3}$$

Finalmente, se na distribuição de STUDENT o valor de n2 se aproxima ao infinito, obteremos:

$$\lim_{n_2 = inf.} y\left[Student\right] = k_3 \cdot \left(1 + \frac{1}{n_2} D^2\right)^{-\frac{n_2 + 1}{2}} = k_4 \cdot e^{-\frac{1}{n_2} D^2 \cdot \frac{n_2 + 1}{2}} = k_4 \cdot e^{-\frac{1}{2} D^2} \qquad \text{---- 4—4}$$

$$= y\left[Gauss\right]$$

pois podemos igualar n_2 *e* $(n_2 + 1)$

Neste capítulo explicamos apenas a transformação matemática da parte das equações que contém as funções do desvio relativo **D**. Para completar porém o nosso trabalho será ainda necessário demonstrar que os coeficientes k1, k2, k3 e k4 podem fàcilmente ser transformados, o que nós faremos nos capítulos seguintes.

c) Transformação da equação de FISHER para a de PEARSON

Precisamos demonstrar que a equação (3-1) passa para a equação (3-2) quando n2 se aproxima ao valor infinito.

Separemos em primeiro lugar dois grupos de têrmos na constante k1: um primeiro contendo apenas n1 e que não sofrerá alterações, sendo independente de n2 e D; um segundo que se altera com a aproximação do valor de n2 ao infinito, sendo também independente de D:

$$k_1 = \frac{n_1^{\frac{n_1}{2}}}{\frac{n_1-2}{2}!} \cdot \frac{\frac{n_1+n_2-2}{2}!\,2}{\frac{n_2-2}{2}!\;n_2^{\frac{n_1}{2}}}$$

Na transformação do segundo têrmo aplicamos os princípios da matemática acima mencionados. Escrevamos ainda para simplificar:

$$a = n_2 - 2$$

$$k' = \frac{\frac{n_1+a}{2}!\,2}{\frac{a}{2}!\;n_2^{\frac{n_1}{2}}}$$

Aplicando agora a fórmula de STIRLING, temos quando n2, e também a = n2 — 2 se aproximam ao infinito:

$$\lim_{a=inf.} \frac{n_1+a}{2}! = \sqrt{2\pi} \cdot \left(\frac{n_1+a}{2}\right)^{\frac{n_1+a+1}{2}} \cdot e^{-\frac{n_1+a}{2}}$$

$$\lim_{a=inf} \frac{a}{2}! = \sqrt{2\pi} \cdot \left(\frac{a}{2}\right)^{\frac{a+1}{2}} \cdot e^{-\frac{a}{2}}$$

$$\lim_{a=inf.} k' = \frac{\sqrt{2\pi} \quad (n_1+a)^{\frac{n_1+a+1}{2}} \cdot 2^{\frac{a+1}{2}} \cdot e^{\frac{a}{2}} \cdot 2^{\frac{n_1}{2}}}{\sqrt{2\pi} \quad 2^{\frac{n_1+a+1}{2}} \cdot a^{\frac{a+1}{2}} \cdot e^{\frac{n_1+a}{2}} \cdot n_2^{\frac{n_1}{2}}}$$

$$= \frac{(n_1+a)^{\frac{n_1}{2}} \cdot (n_1+a)^{\frac{a+1}{2}} \cdot 2^{\frac{a+1}{2}} \cdot 2 \cdot e^{\frac{a}{2}}}{n_2^{\frac{n_1}{2}} \cdot a^{\frac{a+1}{2}} \cdot 2^{\frac{n_1+a+1}{2}} \cdot e^{\frac{n_1+a}{2}}}$$

$$= \left(\frac{n_1+a}{n_2}\right)^{\frac{n_1}{2}} \cdot \left(\frac{n_1+a}{a}\right)^{\frac{a+1}{2}} \cdot 2^{-\frac{n_1-2}{2}} \cdot e^{-\frac{n_1}{2}}$$

$$= \left(\frac{n_1}{n_2}+\frac{a}{n_2}\right)^{\frac{n_1}{2}} \left(\frac{n_1}{a}+1\right)^{\frac{a+1}{2}} \cdot 2^{-\frac{n_1-2}{2}} \cdot e^{-\frac{n_1}{2}}$$

Igualemos n2—1, n2—2= a e n2, o que é justificado quando êste último valor se aproxima ao infinito, e igualemos também (n2+ n1) a n2 que é permissível quando n2 é um número muito grande e n1 um número razoàvelmente pequeno, e obteremos:

$$\lim_{n_2 = inf} (k') = \left(1+\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \cdot \left(1+\frac{n_1}{n_2}\right)^{\frac{n_2}{2}} \cdot 2^{-\frac{n_1-2}{2}} \cdot e^{-\frac{n_1}{2}}$$

$$= \left(1+\frac{n_1}{n_2}\right)^{\frac{n_1+n_2}{2}} \cdot e^{-\frac{n_1}{2}} \cdot 2^{-\frac{n_1-2}{2}}$$

$$= e^{\frac{n_1}{n_2} \cdot \frac{n_1+n_2}{2}} \cdot e^{-\frac{n_1}{2}} \cdot 2^{-\frac{n_1-2}{2}}$$

$$= e^{\frac{n_1}{2}} \cdot e^{-\frac{n_1}{2}} \cdot 2^{-\frac{n_1-2}{2}} = 2^{-\frac{n_1-2}{2}}$$

Voltando agora para a equação inicial de k1, teremos:

$$\lim_{n_2 = inf} (k_1) = \frac{n_1^{\frac{n_1}{2}}}{\frac{n_1-2}{2}!} \cdot \frac{1}{2^{\frac{n_1-2}{2}}} = k_2 \left[\text{Pearson}\right] \quad \text{---} \quad 4\text{—}5$$

Conseguimos assim demonstrar que a constante de FISHER passa para aquela de PEARSON quando n2 se aproxima a um valor infinitamente grande. Já mostrámos a transformação do resto da equação da distribuição de FISHER (4-1) e podemos escrever finalmente, combinando as equações 4-1 e 4-5:

$$\lim_{n_2 = \infty} y\left[Fisher\right] = \frac{\frac{n_1+n_2-2}{2}!}{\frac{n_1-2}{2}!\,\frac{n_2-2}{2}!} \cdot 2 \cdot \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \cdot D^{n_1-1} \cdot \left(1+\frac{n_1}{n_2}D^2\right)^{-\frac{n_1+n_2}{2}}$$

$$= \frac{n_1^{\frac{n_1}{2}}}{\frac{n_1-2}{2}!\,2^{\frac{n_1-2}{2}}} \cdot D^{n_1-1} \cdot e^{-\frac{n_1}{2}D^2}$$

$$\lim_{n_2 = \infty} y\left[Fischer\right] = y\left[Pearson\right] \qquad \text{---------- 4-6}$$

Conseguimos então demonstrar que a equação da distribuição de FISHER (equação 3-1) passa para a equação da distribuição de PEARSON (equação 3-2) quando n2 se aproxima a um valor infinitamente grande.

d) Transformação da distribuição de PEARSON para a de GAUSS

Temos agora a tarefa de demonstrar que a fórmula 3-2 passa para a fórmula 3-4 quando n1 = 1.

A constante k2 da fórmula 3-2 tomará a seguinte forma quando substituirmos n1 pelo valor **um**:

$$k_2 = \frac{n_2}{\frac{n_1-2}{2}! \; 2^{\frac{n_1-2}{2}}} = \frac{1}{\left(-\frac{1}{2}\right)! \; 2^{-\frac{1}{2}}} = \sqrt{\frac{2}{\pi}} \qquad \text{— — — — 4—7}$$

Substituindo também n1 por um nos outros termos, (equação 4-2), temos:

$$\lim y = \sqrt{\frac{2}{\pi}} \cdot e^{-\frac{1}{2}D^2}$$

Agora devemos lembrar um ponto já discutido acima sob 4b, isto é, que uma curva de PEARSON corresponde apenas à metade de uma distribuição de GAUSS. Para evitar que a área da última ficasse o dôbro da primeira, devemos dividir as ordenadas por 2:

$$\lim_{n_1 = 1} y\left[Pearson\right] = \frac{1}{2}\sqrt{\frac{2}{\pi}} \cdot e^{-\frac{1}{2}D^2} = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}D^2} = y\left[Gauss\right] \text{--}$$

$$\text{— — — — — — — 4—8}$$

e) **Transformação da equação de FISHER para a de STUDENT**

Mostraremos que basta substituir na equação da distribuição de FISHER n1 por 1, para se obter a fórmula de STUDENT:

$$\lim_{n_1 = 1} y\left[Fisher\right] = \frac{\frac{n_1+n_2-2}{2}!}{\frac{n_1-2}{2}! \; \frac{n_2-2}{2}!} \cdot 2 \cdot \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \cdot \left(1+\frac{n_1}{n_2}D^2\right)^{-\frac{n_1+n_2}{2}}$$

$$= \frac{\frac{n_2-1}{2}!}{\left(-\frac{1}{2}\right)! \; \frac{n_2-2}{2}!} \cdot 2 \cdot \left(\frac{1}{n_2}\right)^{\frac{1}{2}} \cdot \left(1+\frac{1}{n_2}D^2\right)^{-\frac{n_2+1}{2}}$$

$$= \frac{\frac{n_2-1}{2}!}{\frac{n_2-2}{2}! \; \sqrt{n_2\pi}} \cdot 2 \cdot \left(1+\frac{1}{n_2}D^2\right)^{-\frac{n_2+1}{2}}$$

Considerando que a área da curva de FISHER corresponde apenas à metade da área das curvas de STUDENT como explicado acima (4b) devemos, para igualar as áreas, dividir as ordenadas por dois:

$$\lim_{n_1 = 1} y\left[Fisher\right] = \frac{\frac{n_2-1}{2}!}{\frac{n_2-2}{2}!\sqrt{n_2\pi}} \cdot \left(1+\frac{1}{n_2}\right)^{-\frac{n_2+1}{2}} = y\left[Student\right]$$

— — — — — — 4—9

f) Transformação da equação de STUDENT para a de GAUSS quando n2 aproxima-se ao infinitamente grande

Fazendo a transformação da constante k3:

$$\lim_{n_2 = inf} \left(\frac{n_2-1}{2}\right)! = \sqrt{2\pi} \cdot \left(\frac{n_2-1}{2}\right)^{\frac{n_2}{2}} e^{-\frac{n_2-1}{2}}$$

$$\lim_{n_1 = inf} \left(\frac{n_2-2}{2}\right)! = \sqrt{2\pi} \left(\frac{n_2-2}{2}\right)^{\frac{n_2-1}{2}} e^{-\frac{n_2-2}{2}}$$

$$\lim_{n_2 = inf} k_3 = \frac{\sqrt{2\pi} \cdot (n_2-1)^{\frac{n_2}{2}} \cdot 2^{-\frac{n_2-2}{2}} \cdot e^{\frac{n_2-2}{2}}}{\sqrt{2\pi} \cdot 2^{\frac{n_2}{2}} \cdot (n_2-2)^{\frac{n_2-1}{2}} e^{\frac{n_2-1}{2}}} \cdot \frac{1}{\sqrt{n_2\pi}}$$

$$= \frac{(n_2-1)^{\frac{n_2}{2}} \cdot 1 \cdot 1}{(n_2-2)^{\frac{n_2-1}{2}} \cdot 2^{\frac{1}{2}} \cdot e^{\frac{1}{2}} \quad n_2^{\frac{1}{2}}} \cdot \frac{1}{\sqrt{\pi}}$$

$$= \frac{\left(1-\frac{1}{n_2}\right)^{\frac{n_2}{2}} \cdot n_2^{\frac{n_2}{2}}}{\left(1-\frac{2}{n_2}\right)^{\frac{n_2-1}{2}} \cdot n_2^{\frac{n_2-1}{2}} \cdot n_2^{\frac{1}{2}} \cdot e^{\frac{1}{2}}} \quad \frac{1}{\sqrt{2\pi}}$$

$$= \frac{e^{\frac{1}{2}}}{e^{\frac{n_2-1}{n_2}} \cdot e^{\frac{1}{2}}} \cdot \frac{1}{\sqrt{2\pi}}$$

Sendo, porem, 1 muito pequeno em relação a n_2, *podemos igualar* n_2 *e* (n_2-1) *e obtermos finalmente:*

$$\lim_{n_2=inf.} k_3 = \frac{1}{\sqrt{2\pi}} = k_4 \text{ ---------- } 4\text{—}10$$

O segundo têrmo da equação de STUDENT se transforma simultaneamente, como já demonstrámos acima (4-3), e assim chegamos ao fim da tarnsformação:

$$\lim_{n_2=inf.} y\,[Student] = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}D^2} = y\,[Gauss] \text{ ---} \quad 4\text{—}11$$

5 — A FORMA DAS DISTRIBUIÇÕES EM DEPENDÊNCIA DO GRAU DE LIBERDADE

a) **As distribuições de STUDENT** — A Fig. 5 contém as curvas correspondentes aos valores de n2 = 5, 10, 20 e 30, sendo n1 sempre igual a 1. Nota-se que as curvas se tornam cada vez mais altas e mais curtas em ambas as extremidades.

Quando n2 vai além de 30, podemos usar os limites da distribuição de GAUSS como uma aproximação suficiente. A razão para esta recomendação é bem evidente. A diferença entre as curvas de n2 = 20 e n2 = 30 são da mesma ordem como aquela entre n2 = 30 e n2 = infinito, isto é, a curva de GAUSS.

b) **As distribuições de PEARSON** — A Fig. 6 representa as curvas que correspondem a n1 = 2, 5, 10, 20, sendo nf2 sempre infinito. As curvas são altamente assimétricas quando n1 é baixo. Quando o grau de liberdade cresce, a curva torna-se mais alta, mais estreita e menos assimétrica.

As curvas devem se aproximar, como explicado no capítulo 2 (Fórmula 2-3) acima, à curva modificada de GAUSS com média D = 1, e um êrro standard que em vez de unidade, será dividido pela raiz quadrada de 2 n1.

Para explicar melhor daremos a fórmula de distribuição de GAUSS com uma pequena modificação: para acentuar que se trata de uma distribuição com êrro standard unidade, es-

creveremos, em vez de D, D:1, logo, substituindo, teremos a fórmula de uma distribuição modificada de GAUSS:

$$\lim_{n_1 > 30} y\left[Pearson\right] = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{D^2}{1:\sqrt{2n_1}}\right)^2} = \frac{1}{\sqrt{2\pi}} \cdot e^{-n_1 D^2} \quad \text{—5—1}$$

A Fig. 5, em baixo, ilustra a diferença da curva de PEARSON e da curva modificada de GAUSS, para n1 = 20 e n2 = 30. Especialmente as últimas duas curvas tornam-se já muito semelhantes, de modo que parece justificado, para graus de liberdade superiores a 30, usar em vez da distribuição de PEARSON, aquela de GAUSS.

Podemos determinar os limites de probabilidade pela equação:

$$\left.\begin{array}{l}\textit{limites de probabilidades das}\\ \textit{distribuições de Pearson para}\\ \textit{valores de } n_1 \textit{ maiores que } 30\end{array}\right\} = 1 \pm \textit{limites (Gauss)} : \sqrt{2n_1} \quad \text{—5—2}$$

Fisher dá à esta relação uma forma matemática diferente. Ele escreve:

$$\sqrt{2X^2} - \sqrt{2n_1 - 1} = \textit{limites (Gauss)}$$

Despresando 1 em relação a $2n_1$ *e dividindo por* $\sqrt{2n_1}$ *teremos:*

$$\sqrt{\frac{X^2}{n_1}} - 1 = \textit{limite (Gauss)} : \sqrt{2n_1}$$

$$\sqrt{\frac{X^2}{n_1}} = D = 1 \pm \textit{limite (Gauss)} \cdot \sqrt{2n_1}$$

o que é matemàticamente o mesmo do que a equação dada acima (5-2).

Os limites de GAUSS que nós devemos usar são limites uni-

laterais, pois referem-se apenas a uma extremidade da distribuição, de acôrdo com o que explicámos acima, e são para 5% — 1,64; 1% — 2,33; 1 °/oo — 3,09.

c) **Distribuição de FISHER** — Para melhor esclarecer a situação, trataremos três casos separadamente.

I — Na Fig. 7 encontramos uma série de curvas para as seguintes combinações: n1 = n2 = 2; 5; 10; 20; 30; 50.

As curvas que correspondem a graus de liberdade baixos são muito assimétricas, mas quando êste valor sobe a 30 ou além, a assimetria já parece muito pouco acentuada (Fig. 8), e evidentemente podemos esperar uma aproximação a uma distribuição modificada de GAUSS. Simultaneamente altera-se a curtosis, e as distribuições tornam-se cada vez menos chatas com o crescimento do grau de liberdade. Explicamos acima (pag. 328) que o índice de curtosis ou excessividade β_2 atinge neste casos o seu valor 3, típico para a distribuição de GAUSS.

Para definir matemàticamente a natureza da distribuição modificada de GAUSS, devemos fazer as seguintes considerações:

Supomos que dois erros σ_1 e σ_2, que nós queremos comparar, fôssem desvios de acaso de um valor σ_0 . Assim, e para valores de n1 = n2 bastante elevados, podemos determinar o êrro correnpondente à variação dos erros pela fórmula :

$$\sigma\left[\sigma\right] = \frac{\sigma_0}{\sqrt{2n}}$$

O êrro correspondente ao quociente dos dois erros pode ser determinado aplicando a fórmula geral do êrro de um quociente de duas variáveis a e b, independentes, com médias $\bar{a}$ e $\bar{b}$ e com erros σa e σb, do seguinte modo :

$$\sigma\left[\frac{a}{b}\right] = \pm \frac{\bar{a}}{\bar{b}} \sqrt{\left(\frac{\sigma_a}{\bar{a}}\right)^2 + \left(\frac{\sigma_b}{\bar{b}}\right)^2} \qquad \text{—5—3}$$

$$\sigma\left[\frac{\sigma_1}{\sigma_2}\right] = \pm \frac{\sigma_e}{\sigma_0} \sqrt{\left(\frac{\sigma_0}{\sigma_0\sqrt{2n}}\right)^2 + \left(\frac{\sigma_0}{\sigma_0\sqrt{2n}}\right)^2} = \pm \frac{1}{\sqrt{n}} \qquad \text{—5—4}$$

Assim podemos definir as ordenadas da distribuição modificada de GAUSS do quociente dos dois erros quando n1 = n2 = n como sendo igual a:

$$\lim y\left[Fisher\right] = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{D^2}{\sqrt{n}}\right)^2} = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{D^2}{2n}} \qquad -5-5$$

Os limites da probabilidade podem assim ser calculados quando n1 = n2 = n, e sendo n um valor grande pela fórmula simples:

$$\frac{\sigma_1}{\sigma_2} = 1 \pm \left(limites\ de\ Gauss\right) : \sqrt{n} \qquad -5-6$$

Os limites convencionais da distribuição de GAUSS são:

	5%	1%	1º/oo
para uma extremidade	1,64	2,33	3,09
para ambas as extremidades	1,96	2,53	3,29

Na Fig. 8 estão reproduzidas as duas curvas de FISHER para n1 = n2 = 30 e n1 = n2 = 50, em comparação com as respectivas curvas modificadas de GAUSS. Evidentemente a aproximação não é suficiente para n1 = n2 = 30, e apenas moderada para n1 = n2 = 50. Como melhor aproximação podemos recomendar aplicar a fórmula dada acima (5-6) quando n1 = n2 é igual ou superior a 100.

II — Em segundo lugar estudaremos casos onde os valores de n1 e n2 são diferentes.

A parte superior da figura (Fig. 9) contém as curvas que têm tôdas o mesmo valor n1 = 10, nas seguintes combinações:

n1	10	10	10	10	10	10	10	10
n2	1	2	5	10	20	50	100	infinito

A primeira destas distribuições é evidentemente uma curva recíproca de STUDENT para n1 = 10/n2 = 1 em vez de n1 = 1 / n2 = 10. A fórmula desta nova distribuição podemos fàcilmente obter por uma transformação semelhante àquela do capítulo 5e.

Transformemos em primeiro lugar o último têrmo da equação 3-1 para n2 = 1.

$$\frac{D^{n_1-1}}{(1+n_1D^2)^{\frac{n_1+1}{2}}} = \frac{D^{n_1-1}\cdot(n_1D^2)^{-\frac{n_1+1}{2}}}{\left(1+\frac{1}{n_1D^2}\right)^{\frac{n_1+1}{2}}} = \frac{1}{n_1^{\frac{n_1+1}{2}}}\cdot D^{-2}\cdot\left(1+\frac{1}{n_1D^2}\right)^{-\frac{n_1+1}{2}}$$

Transformando agora também a constante k1, e reagrupando os têrmos da área da curva além do valor D igual a zero, obteremos para a distribuição recíproca de STUDENT, com:

$$\lim_{n_2=1} y\left[Fisher\right] = \frac{\frac{n_1-1}{2}!}{\frac{n_1-2}{2}!\sqrt{\pi}}\cdot\frac{n_1^{\frac{n_1}{2}}}{n_1^{\frac{n_1+1}{2}}}\cdot D^{-2}\cdot\left(1+\frac{1}{n_1D^2}\right)^{-\frac{n_1+1}{2}}$$

$$= \frac{\frac{n_1-1}{2}!}{\frac{n_1-2}{2}!\sqrt{n_1\pi}}\cdot D^{-2}\cdot\left(1+\frac{1}{n_1D^2}\right)^{-\frac{n_1+1}{2}} \quad \text{---} \quad \text{— 5—7}$$

De outro lado, a última curva da Fig. 9 (parte superior), com n1 = 10 / n2 = infinito, é uma distribuição típica de PEARSON.

III — Na Fig. 9, parte inferior, temos representadas as curvas das distribuições recíprocas, sendo agora n2 = 10 :

n1	1	2	5	10	20	50	100	infinito
n2	10	10	10	10	10	10	10	10

A primeira curva é evidentemente uma curva de STUDENT, mas para manter as curvas comparáveis, será necessário usar apenas a metade da curva de STUDENT, multiplicando as ordenadas por 2 para ter-se uma área de unidade.

O outro extremo podemos considerar como a curva "recíproca" da distribuição de PEARSON sendo n1 = infinito, em vez de n2. As ordenadas de uma tal curva podem ser determinadas fàcilmente, depois de uma transformação conveniente da fórmula de FISHER:

$$y\left[Fisher\right] = \frac{1}{\frac{n_2-2}{2}!}\cdot\frac{\frac{n_1+n_2-2}{2}!\cdot 2}{\frac{n_1-2}{2}!}\cdot\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}\cdot D^{n_1-1}\cdot\left(1+\frac{n_1}{n_2}D^2\right)^{-\frac{n_1+n_2}{2}}$$

O primeiro têrmo é independente de n1. O segundo têrmo, que podemos chamar K", podemos transformar em analogia ao têrmo k' da equação 4-5.

$$\lim_{n_1 = inf.} \left[\frac{\frac{n_1+n_2-2}{2} \,!\, 2}{\frac{n_1-2}{2} \,!\, n_1^{\frac{n_2}{2}}} \right] = \lim\, k''.\left(n_1\right)^{\frac{n_2}{2}} = \frac{n_1^{\frac{n_2}{2}}}{2^{\frac{n_2-2}{2}}}$$

Temos ainda que tomar em consideração os têrmos restantes:

$$\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \cdot \frac{D^{n-1}}{\left(1+\frac{n_2}{n_1}D^2\right)^{\frac{n_1+n_2}{2}}} = \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{D^{n-1}}{\left(\frac{n_2}{n_1}D^2+1\right)^{\frac{n_1+n_2}{2}}} \left(\frac{n_2}{n_1}D^2\right)^{\frac{n_1+n_2}{2}}$$

$$\equiv \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \left(\frac{n_2}{n_1}\right)^{\frac{n_1+n_2}{2}} D^{(n-1)} \; D^{-(n_1+n_2)} \left(1+\frac{n_2}{n_1D^2}\right)^{-\frac{n_1+n_2}{2}}$$

$$\equiv \left(\frac{n_2}{n_1}\right)^{\frac{n_2}{2}} D^{-(n_2+1)} \left(1+\frac{n_2}{n_1D^2}\right)^{-\frac{n_1+n_2}{2}}$$

O último têrmo transforma-se ainda em:

$$\lim_{m=inf.}\left(1+\frac{n_2}{n_1D^2}\right)^{-\frac{n_1+n_2}{2}} \equiv e^{\frac{n_2}{n_1D^2}\cdot -\frac{n_1+n_2}{2}} = e^{-\frac{n_2}{2D^2}\cdot\frac{n_1+n_2}{n_1}} \equiv e^{-\frac{n_2}{2}}$$

pois sendo n2 pequeno em relação a n1, que é igual a infinito, podemos escrever (n1 + n2) : n1 igual a 1.

Assim, chegámos finalmente à seguinte transformação do conjunto, para a equação da distribuição recíproca de PEARSON:

$$\lim_{n_1 = inf.} y\left[Fisher\right] = \frac{1}{\frac{n_2-2}{2}!} \cdot \frac{n_1^{\frac{n_2}{2}}}{2^{\frac{n_2-2}{2}}} \cdot \left(\frac{n_2}{n_1}\right)^{\frac{n_2}{2}} \cdot D^{-(n_2+1)} \cdot e^{-\frac{n_2}{2}D^{-2}}$$

$$= \frac{n_2^{\frac{n_2}{2}}}{\frac{n_2-2}{2}! \cdot 2^{\frac{n_2-2}{2}}} \cdot D^{-(n_2+1)} \cdot e^{-\frac{n_2}{2}D^{-2}} = y\left[Pearson\ recipr.\right] \quad 5—8$$

Mencionamos acima sob 3a que nós podemos considerar as distribuições de STUDENT e de PEARSON como os casos extremos da distribuição de FISHER, e tambem explicamos sob 3e que a área das curvas de FISHER corresponde à metade da área das de STUDENT. Agora comprovamos êstes pontos de forma matemática mais concisa.

Uma inspeção das curvas na Fig. 9 mostra que a diferença entre as curvas para a combinação 10/50 e 10/100 de um lado, e de 10/100 e 10/infinito são da mesma ordem. Daí podemos concluir que quando um dos graus de liberdade é maior do que 100 nós podemos substituir a distribuição de FISHER pela distribuição correspondente de PEARSON para todos os fins da análise estatística.

Explicámos acima que as distribuições de PEARSON por sua vez podem ser substituidas por uma distribuição modificada de GAUSS quando o grau de liberdade n1 ultrapassa 30. Agora devemos determinar qual a distribuição modificada de GAUSS que corresponde ao limite de distribuição de FISHER quando n1 difere de n2 e quando um dêles é igual ou maior do que 50 e o outro do que 100. A média destas distribuições é D=1, e a unidade da abcissa será um êrro standard que podemos derivar, como no caso anterior de n1 = n2, da fórmula do quociente (5—5), pela transformação :

$$\sigma\left(\frac{\sigma_1}{\sigma_2}\right) = \pm \frac{\sigma_0}{\sigma_0} \sqrt{\left(\frac{\sigma_0}{\sigma_0\sqrt{2n_1}}\right)^2 + \left(\frac{\sigma_0}{\sigma_0\sqrt{2n_2}}\right)^2}$$

$$= \pm \sqrt{\frac{n_1 + n_2}{2 \cdot n_1 \cdot n_2}} \qquad \text{— — — — — 5—9}$$

Assim teremos:

$$lim\ y\ [fisher] = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}D^2 \cdot \frac{n_1+n_2}{2 \cdot n_1 \cdot n_2}} \quad \text{---------- --- 5—10}$$

para $n_1 = 50/n_2 = 100$ *ou* $n_1 = 100/n_2 = 50$

Os limites de probabilidade podemos assim determinar pela equação:

$$\left.\begin{array}{l} \textit{Limites de probabilidade das distribuições de Fisher} \\ \textit{para } n_1 > 50 \ ; \ n_2 > 100 \\ \textit{ou } n_1 > 100 \ ; \ n_2 > 50 \end{array}\right\} = 1 \pm \text{limites}\,[Gauss] : \sqrt{\frac{n_1+n_2}{2n_1 n_2}}$$

— — — — — — 5—11

6 — RELAÇÃO ENTRE AS DISTRIBUIÇÕES E OS TESTES DERIVADOS

Discutimos acima as distribuições de acaso empregando para isso fórmulas estritamente comparáveis, com a mesma

área e usando sempre o valor do desvio relativo D como a abcissa. Vamos mostrar numa publicação futura e já em preparação, que os testes necessários na análise estatística, podem ser derivados todos destas mesmas fórmulas. Porém, os testes como usados da literatura, se referem geralmente a distribuições com outras unidades como abcissas.

Assim, as distribuições de FISHER são em geral usadas hoje com três abcissas diferentes: o têrmo original z de FISHER (1923), o têrmo F de SNEDECOR (1937) e o têrmo ϑ de BRIEGER (1937), sendo êstes ligados pela equação:

$$D = \frac{\sigma_1}{\sigma_2} = z\,[\text{Fisher}] = \sqrt{F}\ [\text{Snedecor}] = \vartheta\,[\text{Brieger}] \quad \text{---} \qquad \text{— — — — — — 6—1}$$

Igualmente, a distribuição de PEARSON é aplicada no X^2 teste com a abcissa X^2, sendo:

$$D = \frac{\sigma_1}{\sigma_2} = \sqrt{\frac{X^2}{n_1}} \quad ou \quad X^2 = \frac{\sigma_1^2 \cdot n_1}{\sigma_2^2} = D^2 \cdot n_1 \quad \text{-----} \qquad \text{— — — — — — 6—2}$$

A aplicação dos testes nestas formas derivadas parece-me ter bastante conveniências.

Em primeiro lugar o técnico, quando não especializado, é em geral incapaz de compreender que todos os testes, que empregam os limites das distribuições de acaso, são intimamente ligados de modo que êle aplica os testes sem compreender do que se trata de fato — uma situação bastante perigosa, especialmente na análise de casos complicados.

Em segundo lugar, o técnico acostuma-se a exigir um grau diferente de exatidão nos testes, pois no t-teste, derivado das distribuições de STUDENT, êle analisa os valores simples dos desvios relativos sendo $D = t$, e no F-teste, de SNEDECOR de acôrdo com as distribuições de FISHER, êle usa os quadrados dos desvios $D^2 = F$. Algèbricamente, a segunda casa decimal no t-teste corresponde à quarta casa decimal no F-teste, sendo por exemplo o quadrado de 0,01 igual a 0,0001.

Na aplicação do X^2 teste a situação é pior ainda e por exemplo aqui temos o seguinte, para n1 = 10:

D = 9,0 X^2 = 81,0	D = 1,0 X^2 = 10
D = 0,9 X^2 = 8,1	D = 0,1 X^2 = 0,1
D = 0,09 X^2 = 0,081	D = 0,01 X^2 = 0,001

Unidades de D dão assim dezenas ou centenas em X^2, decimais de D dão unidades ou décimos de X^2, centésimos de D dão centésimos ou milésimos de X^2.

Para o caso da distribuição de PEARSON vamos mostrar como se deve transformar as fórmulas para manter a área unidade, apesar da alteração da unidade da abcissa. A equação para estas distribuições com abcissa D era a equação (3-2):

$$y\left[Pearson\right] = \frac{n_1^{\frac{n_1}{2}}}{\frac{n_1-2}{2}\,!\;2^{\frac{n_1-2}{2}}} \cdot D^{n_1-1} \cdot e^{-\frac{n_1}{2}D^2} \qquad \text{---} \quad 6\text{—}3$$

Substituimos agora D *por* $X\ n_1^{\frac{1}{2}}$

$$y\left[Pearson\right] = \frac{n_1^{\frac{n_1}{2}}}{\frac{n_1-2}{2}\,!\;2^{\frac{n_1-2}{2}}} \cdot X^{n_1-1} \cdot n_1^{-\frac{n_1-1}{2}} \quad e^{-\frac{n_1}{2}\cdot X^2 \cdot \frac{1}{n_1}}$$

$$= \frac{\sqrt{n_1}}{\frac{n_1-2}{2}\,!\;2^{\frac{n_1-2}{2}}} \cdot X^{n_1-1} \cdot e^{-\frac{1}{2}X^2}$$

Para poder agora passar da abcissa D para X sem alterar a área da curva, temos que dividir tambem y por $\sqrt{n_1}$:

$$y\left[X\right] = \frac{1}{\frac{n_1-2}{2}\,!\;2^{\frac{n_1-2}{2}}} \quad X^{n_1-1} \cdot e^{-\frac{1}{2}X^2} \qquad \text{------} \quad \text{---} \quad 6\text{—}4$$

Se agora quisermos ainda passar para a distribuição de X^2 em vez de X, teremos ainda que dividir as ordenadas por X:

$$y\left[X^2\right] = \frac{y[X]}{X} = \frac{y\left[p_{pearson}\right]}{X : \sqrt{n_1}} = \frac{1}{\frac{n_1-2}{2}\,!\;2^{\frac{n_1-2}{2}}} \cdot X^{n_1-2} \cdot e^{-\frac{1}{2}X^2}$$

$$= \frac{1}{\frac{n_1-2}{2}\,!} \cdot \left(\frac{X^2}{2}\right)^{\frac{n_1-2}{2}} e^{-\frac{X^2}{2}} \qquad \text{---------- 6—5}$$

As três equações 6-3, 6-4 e 6-5 podem ser encontradas na literatura, porém em geral os autores não explicam claramente que se trata da mesma distribuição básica do ponto de vista matemático, porém com diferentes sistemas de ordenadas e abcissas ,sendo a unidade das abcissas o valor $D = \sigma_1 : \sigma_0$, na fórmula 6-3, X na fórmula 6-4 e finalmente X^2 na fórmula 6-5.

7 — CONSIDERAÇÕES FINAIS

Os quatro tipos de distribuições de acaso, discutidos nesta publicação, servem como base da análise estatística. Grande foi o mérito da escola inglêsa de KARL PEARSON, STUDENT e R. A. FISHER ter mostrado que a chamada distribuição "normal" é apenas um caso especial, com um nome bastante enganador, pois um caso especial nunca podia ser considerado como norma. Com as descobertas dêstes três autores muitos casos que antes não podiam ser submetidos a uma análise efi-

ciente, puderam finalmente ser estudados e analisados com exatidão.

Não devemos, porém, esperar que agora todos os problemas encontrarão uma solução fácil para a sua análise. Frequentemente temos que submeter antes os dados a transformações algébricas, para poder submetê-los à análise, e estas transformações não raramente envolvem princípios da teoria estatística. Em outros casos a situação torna-se ainda mais complicada, pois trata-se de combinações de variáveis, cada uma das quais segue um outro tipo de distribuição de acaso. Explicaremos estas complicações com algum detalhe.

a) **Distribuição de FISHER** — Quando analisarmos um desvio relativo

$$D = \sqrt{\frac{\Sigma(v-\bar{v})^2}{n_1}} : \sigma_e$$

podemos aplicar o princípio da decomposição ou da composição. Supomos que temos uma série de **m** amostras, com médias $\bar{v}_a, \bar{v}_b \ldots$ e graus de liberdade $n_a, n_b \ldots$ Então podemos decompor o conjunto $\Sigma (v-\bar{v})^2$ em partes, devendo ser porém os respectivos graus de liberdade todos iguais:

$$n_a = n_b = \ldots\ldots$$

$$n_a + n_b \ldots\ldots = m \cdot n_a = m \cdot n_b$$

$$\frac{\sigma_a^2+\sigma_b^2+\ldots}{m} = \frac{1}{m}\left(\frac{\Sigma(v-\bar{v}_a)^2}{n_a} + \frac{\Sigma(v-\bar{v}_b)^2}{n_b}+\ldots\right)$$

$$= \frac{\Sigma(v-\bar{v}_a)^2 + \Sigma(v-\bar{v}_b)^2\ldots}{n_a + n_b \ldots\ldots} = \frac{\Sigma(v-\bar{v})^2}{m.np} = \sigma_b^2$$

— — — — — — 7—1

Esta derivação pode também ser lida no sentido inverso, e teremos então em vez de uma composição de um êrro total, a sua decomposição.

Todos os erros parciais σa, σb, se baseiam no mesmo grau de liberdade, na = nb, de modo que êles foram determinados com a mesma exatidão e podem ser comparados em conjunto, pelo teste de FISHER.

Porém, se os graus de liberdade individuais forem diferentes, o conjunto não mais pode ser considerado como homogêneo. Para a análise de cada êrro parcial teremos outro grau de liberdade e assim outros limites calculados de diferentes distribuições de FISHER. Daí a exigência na decomposição do êrro ("analysis of variance") que as partes que são reunidas no êrro residual simples deveriam ser baseadas no mesmo grau de liberdade.

Uma outra complicação pode ser causada às vezes quando os diferentes erros parciais σa, σb não podem ser considerados comparáveis. Por exemplo, em experiências de adubação, pode acontecer que a variação dos canteiros com adubo e sem adubo seja diferente. No primeiro entram os enganos na distribuição, mistura e pesagem dos adubos, e nos outros o efeito de elementos mínimos no solo que terão um efeito mais forte por falta de qualquer compensação biológica. Aqui então um teste estatístico de homogeneidade será necessário para provar se as componentes podiam ser combinadas estatisticamente.

b) Depois discutiremos o **X^2 teste.** Temos às vezes que comparar desvios de dimensão diferente, isto é, diferenças $(v - \bar{v})$ nas quais os valores de $\bar{v}$ são de dimensões bastante diferentes. Assim, PEARSON, quando introduziu o X^2 teste para comparar frequências observadas com frequências esperadas,se viu na dificuldade de reduzí-las a uma base comparável e que permitisse a sua combinação. Por isso devemos no X^2 teste dividir primeiro cada quadrado da diferença entre uma frequência observada e esperada pela última, para depois somar todos êstes quocientes.

$$X^2 = \Sigma \frac{(f_{ob.} - f_{esp.})^2}{f_{esp}} \qquad \text{7—2}$$

Se num caso especial todos os valores de frequência esperada fôssem os mesmos, poderíamos escrever:

$$X^2 = \frac{\Sigma\,(f\,ob - f\,esp)^2}{f\,esp} = \frac{\Sigma\,(v - v_0)^2}{v_0} \qquad \text{7—3}$$

PEARSON (1900) demonstrou que as raizes dêstes quocientes são distribuidas de acôrdo com a fórmula 6-5 com área unidade e usando X2 com uma unidade de abcissa. Se extrairmos a raiz quadrada e dividirmos os valores ainda por $\sqrt{n_1}$ êles ficam distribuidos segundo uma distribuição típica de PEARSON, de acôrdo com a nossa fórmula 3-2, com área unidade e usando o êrro standard como unidade da abcissa.

c) Uma situação semelhante encontramos com referência à distribuição de FISHER, quando queremos comparar dois erros σ_1 e σ_2 que se referem a amostras com médias bastante diferentes. Êste é um problema que aparece frequentemente na genética quando temos que analisar a segregação mendeliana de caracteres quantitativos. Por exemplo, em F2 e em F3 de um cruzamento, teremos nas diferentes famílias médias bastante diferentes em consequência da segregação na geração anterior e devemos investigar se há ainda em cada família indicaçõs da continuação da segregação mendeliana. Devemos assim comparar os respectivos erros entre si ou com aquêles de famílias de gerações anteriores num teste que usa as distribuições de FISHER. Em analogia ao caso de X2 do capítulo anterior (b) dividimos os desvios quadrados pelas suas médias, ou os erros pela raiz quadrada da média, de modo que obtemos para o desvio relativo:

$$D = \sqrt{\frac{\sigma_1^2 : \bar{v}_1}{\sigma_2^2 : \bar{v}_2}} = \frac{\sigma_1\sqrt{\bar{v}_2}}{\sigma_2\sqrt{\bar{v}_1}} \qquad \text{---------7-4}$$

Tratamos as médias como valores constantes, apenas como indicações de dimensão algébrica desprezando o fato que elas também variam entre certos limites.

Assim parece justificado supor que os quocientes do índice da variação $(\sqrt{\sigma^2 : \bar{v}})$ variam de acôrdo com uma distribuição de FISHER (BRIEGER 1943).

d) Na análise de desvios simples devemos também resolver se as médias podem ser consideradas como constantes ou não. Nós não podemos considerar o desvio $(v-\bar{v})$ como uma só variavel, pois o êrro se refere apenas à variação dos valores v e não a diferença $(v-\bar{v})$. De fato, deveríamos aplicar uma fórmula mais complicada:

$$\frac{v-\bar{v}}{\sigma_{dif}} = \frac{v-\bar{v}}{\sqrt{\sigma^2+\sigma_{\bar{v}}^2}} \qquad \text{---------7-5}$$

Existem duas fórmulas para determinar o valor do êrro da média. Se pudermos repetir o experimento em idênticas condições m vezes, cada vez obtendo um outro valor para a média $\overline{v}$, além da média geral de tôdas as repetições $\overline{\overline{v}}$, o êrro das médias parciais poderá ser calculado pela fórmula:

$$\sigma_{\bar{v}} = \pm \sqrt{\frac{\sum(\bar{v} - \bar{\bar{v}})^2}{m-1}} \quad \text{------} \quad \text{---} \quad 7\text{—}6$$

Podemos neste caso esperar que os valores das diferentes médias parciais variem em redor da média geral $\overline{\overline{v}}$ e com êrro standard de acôrdo com uma distribuição de STUDENT com (m-1) graus de liberdade.

De outro lado, quando há uma só série de variáveis, nós não conhecendo o valor de $\overline{\overline{v}}$, temos que aplicar uma fórmula teórica bastante conhecida:

$$\sigma_{\bar{v}} = \pm \frac{\sigma}{\sqrt{N}} \quad \text{---------} \quad \text{---} \quad 7\text{—}7$$

A derivação da última fórmula inclui a hipótese que a média pode variar em volta de um valor ideal desconhecido, seguindo a distribuição normal ou de GAUSS. Aparece na fórmula 7-10 não o grau de liberdade, mas o número total (N) de variáveis ao qual se refere a média.

Devemos agora substituir os valores das fórmulas 7-6 e 7-7 respectivamente na fórmula 7-5. No último caso obteremos a equação:

$$\frac{v - \bar{v}}{\sigma\, dif.} = \frac{v - \bar{v}}{\sqrt{\sigma^2 + \frac{\sigma^2}{N}}} = \frac{v - \bar{v}}{\sqrt{1 + \frac{1}{N}}} \quad \text{---} \quad 7\text{—}8$$

Agora temos que resolver dois problemas, um que é relativamente fácil e representa uma simples questão de álgebra, enquanto o outro envolve problemas da teoria estatística.

Quanto ao primeiro podemos dizer que em geral a inclusão do valor do êrro da média muito pouco altera o valor final

Êste é especialmente claro quando comparamos os dois valores:

$$\left(\frac{v-\bar{v}}{\sigma}\right) \text{ e } \left(\frac{v-\bar{v}}{\sigma} \cdot \frac{1}{\sqrt{1+\frac{1}{N}}}\right)$$

N	$\sqrt{1+\frac{1}{N}}$	$1:\sqrt{1+\frac{1}{N}}$
5	1,10	0,91
10	1,05	0,95
20	1,02	0,98
50	1,01	0,99

O valor do desvio, calculado do modo comum já para N=5 é apenas 1,10 vezes maior do que quando nós incluimos o êrro da média.

Para o caso do emprêgo da fórmula 7-6 a situação não pode ser resolvida de uma maneira geral. Devemos em cada caso comparar os valores algébricos de σ e $\sigma_{\bar{v}}$ podendo desprezar o último quando fôr pequeno em relação ao primeiro.

O problema estatístico se resolve diferentemente nos dois casos.

Empregando as fórmulas 7-5 e 7-6 em combinação temos um quociente no qual o numerador tem 1 grau de liberdade e o denominador $(N_2-1) + (m-1)$ graus de liberdade, devendo o quociente seguir assim uma distribuição de STUDENT. Êste no mínimo é o processo comumente empregado para calcular o grau de liberdade do êrro de uma diferença, que porém é apenas justificado quando os dois erros, no nosso caso σ e $\sigma_{\bar{v}}$, são mais ou menos do mesmo tamanho. Quando desprezamos o valor $\sigma_{\bar{v}}$ não devemos tomar em consideração o seu grau de liberdade.

Quando usamos as fórmulas 7-8 e 7-10 em combinação, a situação é mais complicada, pois agora nós combinamos na fórmula do êrro da diferença, um êrro standard que segue uma distribuição de STUDENT (σ e outro ($\sigma_{\bar{v}}$)) que segue uma distribuição de GAUSS. A combinação dos graus de liberdade pelo modo comum terá o seguinte valor: (N-1) + infinito; assim, o grau de liberdade do êrro mais importante σ pràtica-

mente fica sem efeito por ser sempre pequeno em relação ao valor infinito.

Sendo $\sigma_{\overline{v}}$ peqeuno em relação a σ, usamos apenas o valor σ, mas sendo o grau de liberdade n pequeno em relação ao grau de liberdade de $\sigma_{\overline{v}}$, desprezamos o primeiro. Chegamos assim a um absurdo: combinamos o valor de um êrro com o grau de liberdade do outro!

As dificuldades desaparecem quando nós desprezamos por completo tanto o valor como o grau de liberdade do êrro das médias. Se julgarmos que isso não é aconselhável, deve-se aplicar a transformação de BEHRENS, que permite combinar uma distribuição de STUDENT com uma de GAUSS (FISHER and YATES 1943).

e) Uma situação semelhante encontramos quando queremos analisar a diferença entre dois valores a e b, que podem ser simples variáveis ou médias, com êrro standard σ a e σ b, baseados em n_a e n_b graus de liberdade. Devemos calcular como sempre o desvio relativo:

$$\frac{a-b}{\sigma_{dif.}} \qquad \text{— — — — — — — — — 7—9}$$

Limitando-nos a diferenças entre valores a e b, não coordenados e não correlacionados, temos duas fórmulas para calcular o êrro da diferença, uma que nós podemos chamar a fórmula clássica ,e outra a fórmula de FISHER:

Fórmula clássica:

$$\sigma_{dif} = \pm \sqrt{\sigma_a^2 + \sigma_b^2} \qquad \text{— — — — — — — — — — — — — 7—10}$$

Fórmula de Fisher

$$\sigma_{dif} = \pm \sqrt{\sigma_a^2 + \sigma_b^2} = \pm \sigma_o \sqrt{2} \qquad \text{— — — — — 7—11}$$

$$\sigma_o = \pm \sqrt{\frac{\sum (a - \bar{a})^2 + \sum (b - \bar{b})^2}{n_a + n_b}}$$

$$= \pm \sqrt{\frac{n_a \sigma_a^2 + n_b \sigma_b^2}{n_a + n_b}}$$

A derivação da primeira fórmula é fácil e não envolve nenhum princípio estatístico complicado. Devemos resolver qual distribuição, que a diferença (a-b) medida por êste êrro standard, deve seguir. De acôrdo com os números n_a e n_b os respectivos erros σ a e σ b correspondem a diferentes distribuições de STUDENT. Se queremos aplicar a teoria da estatística com todo rigor, deveriamos recorrer a uma modificação do teste de BEHRENDS mencionado no capítulo anterior. Como uma aproximação satisfatória podemos usar uma distribuição de STUDENT com ($n_a + n_b$) graus de liberdade, exceto quando um dos graus de liberdade fôr muito maior do que o outro. Neste último caso encontramos uma situação idêntica àquela discutida no capítulo anterior (7d).

Na fórmula de FISHER (7-11)para o êrro de uma diferença encontramos uma situação diferente. Como se pode ver, trata-se de uma média balançada entre as duas variânças. Esta fórmula pode ser aplicada apenas quando estamos certos que os dois erros não são estatisticamente diferentes, porque sòmente neste caso é justificado reunir as variânças para obter uma única estimativa, por sua vez baseada em ($n_a + n_b$) graus de liberdade.

A questão da distribuição não oferece agora nenhuma dificuldade. A diferença (a-b) dividida pela estimativa única do seu êrro (7-10) deve seguir uma distribuição de STUDENT com ($n_a + n_b$) graus de liberdade.

Quando, em vez de variáveis simples a e b, se tratar de médias, as duas fórmulas sofrem apenas uma ligeira alteração, quando podemos calcular erros da média pela fórmula: $\sigma : \sqrt{N}$

Substituindo teremos:

Fórmula clássica:

$$\sigma\left[dif\ \bar{v}\right] = \pm \sqrt{\frac{\sigma_a^2}{N_a} + \frac{\sigma_b^2}{N_b}} \quad \text{---- --- --- 7—12}$$

Fórmula de Fisher:

$$\sigma\left[dif.\ \bar{v}\right] = \pm \sigma_o \sqrt{\frac{1}{N_a} + \frac{1}{N_b}} \quad \text{--- --- --- 7—13}$$

f) A substituição de uma estimativa do êrro standard por outra é geralmente feita na análice estatística chamada "analysis of variance", sem que em geral os autores estejam ao par desta substituição. Usamos como exemplo um caso simples de um teste "entre-dentro" ou "betwen-within".

Supomos que temos m amostras, cada uma com (np + 1) observações ou medidas, e queremos saber se as médias

$\bar{v}_1$ e $\bar{v}_2$ são diferentes da média geral $\bar{\bar{v}}$.

Poderíamos calcular o êrro para cada amostra:

$$\sigma_1 = \pm \sqrt{\frac{\Sigma(v - \bar{v}_a)^2}{n_a}}$$

$$\sigma_b = \pm \sqrt{\frac{\Sigma(v - \bar{v}_b)^2}{n_b}} \quad \text{----}$$

$$\sigma_c \text{ ---------}$$

Mas em vez dêstes erros usa-se em geral uma estimativa geral, que é o êrro dentro, calculado pela fórmula: (veja fórmula 7-1):

$$\sigma_D = \pm \sqrt{\frac{\sum(v-\bar{v})^2}{m.n_p}} = \pm \sqrt{\frac{\sum(v-\bar{v}_a)^2 + \sum(v-\bar{v}_b)^2}{m \cdot n_p}}$$

$$= \pm \sqrt{\frac{np\sigma_a^2 + np\sigma_b^2 \ldots}{m \cdot n_p}} = \pm \sqrt{\frac{\sigma_a^2 + \sigma_b^2 \ldots}{m}}$$

Analisando as diferenças entre médias parciais e a média geral, temos:

$$D = \frac{\bar{v} - \bar{\bar{v}}}{\sigma \text{ dif.}} = \frac{\bar{v} - \bar{\bar{v}}}{\sigma : \sqrt{N_p}} = \frac{\bar{v} - \bar{\bar{v}}}{\sigma} \cdot \sqrt{N_p} \quad --$$

Que valor de σ deveremos usar? Poderemos empregar o êrro de cada média, por exemplo:

$$\left.\begin{aligned} D &= \frac{\bar{v}_a - \bar{\bar{v}}}{\sigma_a} \cdot \sqrt{N_p} \\ D &= \frac{\bar{v}_b - \bar{\bar{v}}}{\sigma_b} \cdot \sqrt{N_p} \end{aligned}\right\}$$

e comparar êste valor com os limites da respectiva distribuição de STUDENT com np graus de liberdade.

Porém, em vez disso, emprega-se geralmente o êrro dentro, que por sua vez é calculado de m.np graus de liberdade:

$$D = \frac{\bar{v} - \bar{\bar{v}}}{\sigma_p} \cdot \sqrt{N_p} \quad ----$$

Êste têrmo segue uma distribuição de STUDENT, com m.np graus de liberdade.

Substituindo σ_1 por σD, não sòmente usamos um outro valor do êrro, mas atribuimos ao desvio relativo um grau de liberdade muito mais elevado, isto é, (m.).np em vez de np. O efeito desta substituição relativo ao limite de probabilidade é muito grande, pois em geral np é bem pequeno, raramente mais do que 5, e o valor de (m.).np chega muitas vezes perto ou além de 30. Dêste modo e conforme explicado acima, substituimos os limites de uma distribuição de STUDENT com n1 igual ou menor do que 5 pelos limites da distribuição de GAUSS.

Antes de substituir o valor individual σ_1, por uma média balançada σD, seria imprescindível provar que a substituição é permissível. Podemos comparar os dois êrros, determinando o quociente

$$\frac{\sigma_1}{\sigma_2} \text{ com } \frac{n_1 = np}{n_2 = m.np}$$

e comparando-o com os limites da distribuição de FISHER correspondente.

Êste teste em geral é de pouca confiança quando n1 é muito pequeno. Quando σ_1 fôr estatisticamente diferente de σD o que não é raro, a substituição não pode ser feita. Mas quando a comparação não indica diferenças significantes, podemos ainda ficar na dúvida por causa do número pequeno de graus de liberdade. Nestes casos recomendo fazer a substituição e usar em vez de σ_1 o valor σD que provàvelmente é uma melhor estimativa do êrro standard, porém continuar a usar o grau de liberdade que se refere a σ_1, isto é np.

Não é nossa intenção entrar aqui em mais detalhes, uma vez que pretendemos tratar dos testes estatísticos em outra publicação. Os exemplos servem para mostrar que os técnicos devem tomar cuidado na aplicação dos limites de probabilidade na análise de casos complicados, para evitar que êles empreguem conceitos e limites errados.

As nossas considerações finais indicam também em que direção novos trabalhos são necessários. A preparação de uma tabela com os limites bilaterais para as distribuições de FISHER é uma tarefa laboriosa, porém não envolve mais um estudo de problemas fundamentais. O principal problema de ordem teórica é a combinação de diferentes distribuições para

os casos nos quais devemos, com todo rigor, comparar variáveis que seguem cada um tipo diferente de distribuição de acaso. O teste de BEHRENDS é um início nesta direção.

RESUMO

1) Chamamos um desvio relativo simples o quociente de um desvio, isto é, de uma diferença entre uma variável e sua média ou outro valor ideal, e o seu êrro standard.

$$D = \frac{v - \bar{v}}{\sigma} \quad ou \quad D = \frac{v - v_0}{\sigma}$$

Num desvio composto nós reunimos vários desvios de acôrdo com a equação:

$$D = \pm \sqrt{\frac{\Sigma (v - \bar{v})^2}{n_1}} : \sigma_0 = \frac{\sigma_1}{\sigma_0}$$

Todo desvio relativo é caracterizado por dois graus de liberdade (número de variáveis livres) que indicam de quantas observações foi calculado o numerador (grau de liberdade nf1 ou simplesmente n2) e o denominador (grau de liberdade nf2 ou simplesmente n2).

2) Explicamos em detalhe que a chamada distribuição normal ou de GAUSS é apenas um caso especial que nós encontramos quando o êrro standard do dividendo do desvio relativo é calculado de um número bem grande de observações ou determinado por uma fórmula teórica. Para provar êste ponto foi demonstrado que a distribuição de GAUSS pode ser derivada da distribuição binomial quando o expoente desta torna-se igual a infinito (Fig.1).

3) Assim torna-se evidente que um estudo detalhado da variação do êrro standard é necessário. Mostramos ràpidamente que, depois de tentativas preliminares de LEXIS e HELMERT, a solução foi achada pelos estatísticos da escola londrina: KARL PEARSON, o autor anônimo conhecido pelo nome de STUDENT e finalmente R. A. FISHER.

4) Devemos hoje distinguir quatro tipos diferentes de dis-

tribuições de acaso dos desvios relativos, em dependência de combinação dos graus de liberdade n1 e n2.

Distribuição de:

Fisher	$1 < nf_1 <$ *infinito*	$1 < nf_2 <$ *infinito* (*formula 3-1*)
Pearson	$1 < nf_1 <$ *infinito*	$nf_2 =$ *infinito* (*formula 3-2*)
Student	$nf_1 = 1$	$1 < nf_2 <$ *infinito* (*formula 3-3*)
Gauss	$nf_1 = 1$	$nf_2 =$ *infinito* (*formula 3-4*)

As formas das curvas (Fig. 2) e as fórmulas matemáticas dos quatro tipos de distribuição são amplamente discutidas, bem como os valores das suas constantes e de ordenadas especiais.

5) As distribuições de GAUSS e de STUDENT (Figs. 2 e 5) que correspondem a variação de desvios simples são sempre simétricas e atingem o seu máximo para a abcissa D = O, sendo o valor da ordenada correspondente igual ao valor da constante da distribuição, k1 e k2 respectivamente.

6) As distribuições de PEARSON e FISHER (Fig. 2) correspondentes à variação de desvios compostos, são descontínuas para o valor D = O, existindo sempre duas curvas isoladas, uma à direita e outra à esquerda do valor zero da abcissa.

As curvas são assimétricas (Figs. 6 a 9), tornando-se mais e mais simétricas para os valores elevados dos graus de liberdade.

7) A natureza dos limites de probabilidade é discutida. Explicámos porque usam-se em geral os limites bilaterais para as distribuições de STUDENT e GAUSS e os limites unilaterais superiores para as distribuições de PEARSON e FISHER (Figs. 3 e 4).

Para o cálculo dos limites deve-se então lembrar que o desvio simples, $D = (v - \bar{v}) : \sigma$ tem o sinal positivo ou negativo, de modo que é em geral necessário determinar os limites bilaterais em ambos os lados da curva (GAUSS e STUDENT).

Os desvios relativos compostos da forma $D = \sigma_1 : \sigma_2$ não têm sinal determinado, devendo desprezar-se os sinais. Em geral consideramos apenas o caso σ_1 ser maior do que σ_2 e os limites se determinam apenas na extremidade da curva que corresponde a valores maiores do que 1. (Limites unilaterais superiores das distribuições de PEARSON e FISHER).

Quando a natureza dos dados indica a possibilidade de aparecerem tanto valores de σ_1 maiores como menores do que σ_2, devemos usar os limites bilaterais, correspondendo os limites unilaterais de 5%, 1% e 0,1% de probabilidade, correspondendo a limites bilaterais de 10%, 2% e 0,2%.

8) As relações matemáticas das fórmulas das quatro distribuições são amplamente discutidas, como também a sua transformação de uma para outra quando fazemos as necessárias alterações nos graus de liberdade. Estas transformações provam matemàticamente que tôdas as quatro distribuições de acaso formam um conjunto.

Foi demonstrado matemàticamente que a fórmula das distribuições de FISHER representa o caso geral de variação de acaso de um desvio relativo, se nós extendermos a sua definição desde nf1 = 1 até infinito e desde nf2 = 1 até infinito.

9) Existe apenas uma distribuição de GAUSS; podemos calcular uma curva para cada combinação imaginável de graus de liberdade para as outras três distribuições. Porém, é matemàticamente evidente que nos aproximamos à distribuições limitantes quando os valores dos graus de liberdade se aproximam ao valor infinito.

Partindo de fórmulas com área unidade e usando o êrro standard como unidade da abcissa, chegamos às seguintes transformações:

a) A distribuição de STUDENT (Fig. 5) passa a distribuição de GAUSS quando o grau de liberdade n2 se aproxima ao valor infinito. Como aproximação ao infinito, suficiente na prática, podemos aceitar valores maiores do que n2 = 30.

b) A distribuição de PEARSON (Fig. 6) passa para uma de GAUSS com média zero e êrro standard unidade quando n1 é igual a 1.

Quando de outro lado, n1 torna-se muito grande, a distribuição de PEARSON podia ser substituida por uma distribuição modificada de GAUSS, com média igual a 1 e unidade da abcissa igual a $1 : \sqrt{2\,n\,1}$. Para fins práticos, valores de n1 maiores do que 30 são em geral uma aproximação suficiente ao infinito.

c) Os limites da distribuição de FISHER são um pouco mais difíceis para definir.

I) Em primeiro lugar foram estudadas as distribuições com n1 = n2 = n e verificamos (Figs. 7 e 8) que aproximamo-nos a uma distribuição transformada de GAUSS com média 1 e êrro standard $1 : \sqrt{n}$, quando o valor cresce até o infinito. Como aproximação satisfatória podemos considerar n1 = n2 = 100, ou já n1 = n2 = 50 (Fig. 8)

II) Quando n1 e n2 diferem (Fig. 9) podemos distinguir dois casos:

Se n1 é pequeno e n2 maior do que 100 podemos substituir a distribuição de FISHER pela distribuição correspondente de PEARSON. (Fig. 9, parte superior).

Se porém n1 é maior do que 50 e n2 maior do que 100, ou vice-versa, atingimos uma distribuição modificada de GAUSS com média 1 e êrro standard

$$1 : \sqrt{\frac{2n_1\ n_2}{n_1 + n_2}}$$

10) As definições matemáticas e os limites de probabilidade para as diferentes distribuições de acaso são dadas em geral na literatura em formas bem diversas, usando-se diferentes sistemas de abcissas.

Com referência às distribuições de FISHER, foi usado por êste autor, inicialmente, o logarítmo natural do desvio relativo, como abcissa. SNEDECOR (1937) emprega o quadrado dos desvios relativos e BRIEGER (1937) o desvio relativo proprio.

As distribuições de PEARSON são empregadas para o X^2 teste de PEARSON e FISHER, usando como abcissa os valores de

$$X^2 = D^2.\ n1$$

Foi exposto o meu ponto de vista, que estas desigualdades trazem desvantagens na aplicação dos testes, pois atribui-se um pêso diferente aos números analisados em cada teste, que são somas de desvios quadrados no X^2 teste, somas des desvios quadrados divididos pelo grau de liberdade ou variançaş no F-teste de SNEDECOR, desvios simples no t-teste de STUDENT, etc..

Uma tábua dos limites de probabilidade de desvios relativos foi publicada por mim (BRIEGER 1937) e uma tábua mais extensa será publicada em breve, contendo os limites unilaterais e bilaterais, tanto para as distribuições de STUDENT como de FISHER.

11) Num capítulo final são discutidas várias complicações que podem surgir na análise. Entre elas quero apenas citar alguns problemas.

a) Quando comparamos o desvio de um valor e sua média, deveríamos corretamente empregar também os erros de ambos êstes valores:

$$D = \frac{v - \bar{v}}{\sqrt{\sigma^2 + \sigma_{\bar{v}}^2}}$$

Mas não podemos aqui imediatamente aplicar os limites de qualquer das distribuições do acaso discutidas acima. Em geral a variação de v, medida por σ, segue uma distribuição de STUDENT e a variação da média $\bar{v}$ segue uma distribuição de GAUSS. O problema a ser solucionado é, como reunir os limites destas distribuições num só teste. A solução prática do caso é de considerar a média como uma constante, e aplicar diretamente os limites de probabilidade das dstribuições de STUDENT com o grau de liberdade do êrro σ. Mas êste é apenas uma solução prática. O problema mesmo é, em parte, solucionado pelo teste de BEHRENDS.

b) Um outro problema se apresenta no curso dos métodos chamados "analysis of variance" ou decomposição do êrro.

Supomos que nós queremos comparar uma média parcial $\bar{v}a$ com a média geral $\bar{\bar{v}}$. Mas podemos calcular o êrro desta média parcial, por dois processos, ou partindo do êrro individual σa ou do êrro "dentro" σD que é, como explicado acima, uma média balançada de todos os m erros individuais. O emprêgo dêste último garante um teste mais satisfatório e severo, pois êle é baseado sempre num grau de liberdade bastante elevado.

Teremos que aplicar dois testes em seguida:

Em primeiro lugar devemos decidir se o êrro σa difere do êrro dentro:

$$D = \frac{\sigma_a}{\sigma_p} \qquad \frac{n_1 = n_p}{n_2 = m \cdot n_p}$$

Se êste teste fôr significante, uma substituição de σa pelo σD não será admissível. Mas mesmo quando o resultado fôr insignificante, ainda não temos certeza sôbre a identidade dos dois erros, pois pode ser que a diferença entre êles é pequena e os graus de liberdade não são suficientes para permitir o reconhecimento desta diferença como significante.

Podemos então substituirmos σa por σD de modo que n2 = m : np:

$$D = \frac{\bar{v}_a - \bar{\bar{v}}}{\sigma_a} \sqrt{N_p} \qquad \frac{n_1 = 1}{n_2 = n_p}$$

passa para

$$D = \frac{\bar{v}_a - \bar{\bar{v}}}{\sigma_p} \sqrt{N_p} \qquad \frac{n_1 = 1}{n_2 = m \cdot n_p}$$

Mas como podemos incluir neste último teste uma apreciação das nossas dúvidas sôbre o teste anterior σa; σD ?

A melhor solução prática me parece fazer uso da determinação de σD, que é provàvelmente mais exata do que σa, mas usar os graus de liberdade do teste simples: np = 1 / n2 = np para deixar margem para as nossas dúvidas sôbre a igualdade de σa a σD.

Êstes dois exemplos devem ser suficientes para demonstrar que apesar dos grandes progressos que nós podíamos registrar na teoria da variação do acaso, ainda existem problemas importantes a serem solucionados.

ABSTRACT

1) The present paper deals with the mathematical basis and the relations of the different chance distributions. It is shown that the concepts of classical statistics may only be applied correctly when dealing with illimited populations where the number of variables is so large that it may be considered as infinite. After the attempts of LEXIS and HELMERT, a partial solution was found by KARL PEARSON and by STUDENT, until finally R. A. FISHER gave the general solution, solving the problem of statistical analysis in a general form and determining the chance distribution in small samples.

2) As a basis for the formulas, I am using always the relative deviate, which may be determined in two ways :

the simple relative deviate :

$$D = \frac{v - \bar{v}}{\sigma}$$

the compound relative deviate :

$$D = \pm \sqrt{\frac{\Sigma(v-\bar{v})^2}{n_1}} : \sigma_0 \quad \text{ou} \quad D = \frac{\sigma_1}{\sigma_2}$$

3) The deviates are always defined by two degrees of freedom, n1 for the dividend and n2 for the divisor. According to the values combined in any given case, we may distinguish four basic chance distributions which we shall call according to the respective authors: the distributions of GAUSS, STUDENT, PEARSON and FISHER.

The mathematical definition and the corresponding degrees of freedom are given both in formulae 3-1 to 3-4 on pg. and in the lower part of Fig. 2. The upper part of Fig. 2. represents grafically these four distributions. The equations and the forma of the corresponding curves are discussed in detail.

4) The main differences between the simple and the compound relative deviate are discussed:

a) Simple deviates have always a definite signe and are either positive or negative, according to the signe of the numerator. Correspondingly the distributions of GAUSS and STUDENT are symetrical with regards to the abscissa zero and extend on both sides of it untill plus and minus infinite. Compound deviates on the other side, have no definite sign, since the numerator is a square root. The distributions of PEARSON and FISHER, accordingly, are discontinuous for the value zero and we obtain two identical and independent curves which go from zero to plus infinite, resp. from zero to minus infinite.

b) Secondly when studying simple deviates we admitt that both positive and negative large deviates may occur in consequence of an increase in variability. Consequently we ha-

ve to use, in the corresponding tests, bilateral limits of probabillity (Fog. 3).

When analysing compound deviates, we are comparing one standard error with another, which may either be an ideal value or at least a better estimate. Admitting that only an increase of variability may occurr, we apply in tests, based on PEARSON's or FISHER's distribuitions, only the upper (superior) unilateral limit of probability (Fig. 4).

The tables thus far published, for these distributions contain the unilaeeral limits only. A more complete table, including bilateral limits, has been computed by myself and is already in press.

5) Discussing the relations of the four distribuitions, it is shown that mathematically their formulas can be easily transformed from one to the other by changing the respective values of degrees of freedom. The application of a few principles of mathematics is sufficient, besides remembering that the distributions of PEARSON and FISHER correspond only to half a distribution of STUDENT and GAUSS.

Thus it is shown:

a) that for n1 bigger than 30, the distribution of STUDENT is so near to that of GAUSS (or normal), to permitt its substitution.

b) that for n1 bigger than 30, the distribution of PEARSON becomes almost symetrical and may be substituted by a modified distribution of GAUSS (or normal) with mean equal to one and error standard

$$1 : \sqrt{2n_1}$$

c) That the distribution of FISHER with n1 = n2 becomes more or less symetrical when both reach the limit of 50 or bester still 100, and than may be substituted by a modified distribution of GAUSS with mean one and error standard.

$$1 : \sqrt{n}$$

d) That the distributions of FISHER, when n1 differs from n2, may be substituted either by the correspondent distribution of PEARSAN, if n1 is small and n2 bigger than 100, or by a modified distribution of GAUSS with mean unity and error standard equal to

$$1 \cdot \sqrt{\frac{2n_1n_2}{n_1+n_2}}$$

when n1 goes beyond 50 and n2 beyond 100 ou vice versa.

6) The formulas, generally given in the litterature to characterize the different distribuitions are far from being uniform and use differents measures for the abcissa.

Thus in the tests for FISHER's distribution, the natural logarithm for the deviate were used initially (FISER's z-test). Later on SNEDECOR (1937) recommended the square of the deviate (F-test) and BRIEGER (1937) the deviate itself (ϑ— test).

In the X^2 test, based on PEARSON's distribution, one generally uses the square of the compound deviate, multiplied by the degree of freedom n1.

The t-test, based on STUDENT's distribution, finally makes use of the simple deviate itself.

The inevitabal algebric consequences of this variation of unitis of emasure is, that the severity and thus the statistical efficiency of the tests is not comparable. Decimals in the t-test and ϑ-test correspond to hundreds in the F-test and to almost anything, depending upon the values of n1, in the X^2-test.

7) In the last chapter a few rather complicated problems are discussed, which can be solved with approximation in praetical tests, but wich are still unsolved from the theoretical point of view. We shall mention here only two of the questions raised:

a) Analyzing the difference between a variable and its mean (or of a partial mean and a general mean), only the standard error of the first term is used generally, considering the other as a constant:

$$D = \frac{v-\bar{v}}{\sigma_v}$$

Howewer with more justification both terms may be considered as variable und thus one should apply the formula :

$$D \cdot \frac{v-\bar{v}}{\sqrt{\sigma^2-\sigma_v^2}}$$

The first mentioned simple value of D chould follow a distribution of STUDENT and its analysis thus does not present any difficulties. But in the second term we combine the term, v, with standard error σ which should follow STUDENT's distribution and the mean, $\overline{v}$, with standard error σv which generally will follow the distribution of GAUSS. How shall we combine the requeriments of those two distribuitions simultaneously? BEHREND's test seems to give a solution, which however is not very easy to apply and which is not suficient when the second term follows also a distribution of STUDENT, but with different degree of freedom.

b) The second problem arises in connection with the anaysis of variance in its most simple form, i. e. the test "within-between". If we want to compare by a t-test the partial mean of one sample $\overline{v}a$ with the general mean $\overline{\overline{v}}$ of the whole experiment, we must decide wether we should use standard error of this sample σa, based on np degrees of freedom or the error "within" σD which is a balanced mean value of all the m individual sample errors. At the same time we have an alternative choice with regards to the degree of freedom:

$$D = \frac{\overline{v}_a - \overline{\overline{v}}}{\sigma_a} \sqrt{N_p} \qquad \frac{n_1 = 1}{n_2 = n_p} \qquad ou$$

$$D = \frac{\overline{v}_a - \overline{\overline{v}}}{\sigma_D} \sqrt{N_p} \qquad \frac{n_1 = 1}{n_2 = m \cdot n_p}$$

Thus it is evident that the use of the value σD not only alters the value of the relative deviate D, but also the limits of probability to be applied which depend upon the degree of greedom. Howewer we must justify the substitution of the partial error σa by the error "within" σD and this should be done by determining wether the value $\sigma a : \sigma D$ is due to chance only, i. e. that there is really no difference between the two errors from a statistical point of view. The necessary test howewer:

$$D = \frac{\sigma_a}{\sigma_D} \qquad \frac{n_1 = n_p}{n_2 = m \cdot n_p}$$

generally does not allow a very decisive answer since the degree of freedom np is in most cases small.

Whenever there is some reason to doubt wether the substitution is really justified, it seems to me reasonable to use the probably better estimate σD, instead of the individual sample error σa, while at the same time make allowances for doubts by not substituting the degrees of freedom:

$$D = \frac{\bar{v}_a - \bar{\bar{v}}}{\sigma_D} \sqrt{N_p} \qquad \frac{n_1 = 1}{n_2 = n_p}$$

A more complete formula naturally would be the following:

$$D = \frac{\bar{v}_a - \bar{\bar{v}}}{\sigma_a \left(\frac{\sigma_p}{\sigma_a}\right)} \sqrt{N_p} \qquad \frac{n_1 = 1}{n_2 = \ldots}$$

c) These too examples should be sufficient to show that there are still important theoretical problems to be solved, in spite of the really very considerable progress achieved with regards to theory and methods of analysis of simple and compound relative deviates from uniform small or large, but always limited samples.

REFERÊNCIAS

1 — BRIEGER, F. G. — 1937 — Tábuas e Fórmulas para Estatistica, 46 pgs. — Cia. Melhoramentos de S. Paulo — S. Paulo.

2 — FISHER, R. A. — 1941 — Statistical Methods for Research Workers — 8th Ed. Oliver and Boyd — London — 334 pgs.

3 — FISHER, R. A. and FRANK YATES — 1943 — Statistical Tables, for Bioloical, Agricultural and Medical Research — Second Edition — Oliver and Boyd Ltd. — London — 98 pgs.

4 — KENNEY, J. F. — 1939 — Mathematics of Statistics — Van Nostrand Company — New York — Vol. 1, 248 pgs — Vol. 2, 202 pgs.

5 — SNEDECOR G. W. — 1937 — Statistical Methods — Collegiate Press — Ames, Iowa.

6 — YULE, G. U. and M. G. KENDALL — 1940 — An Introduction to the Theory of Statistics — Charles Griffin 12th — 570 pgs.

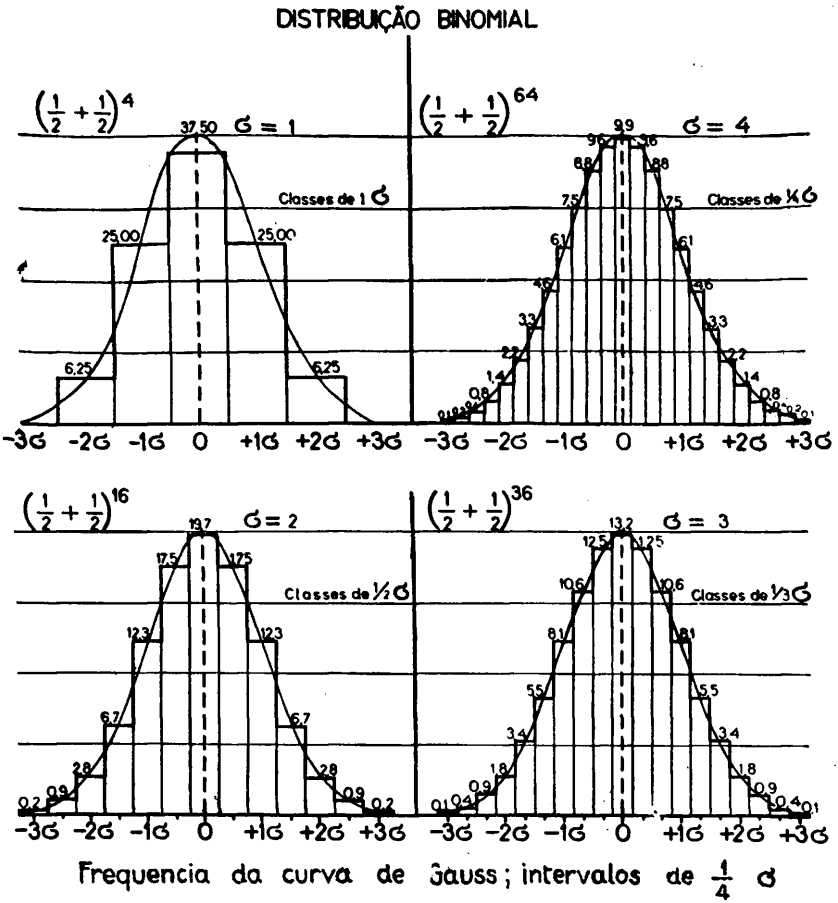
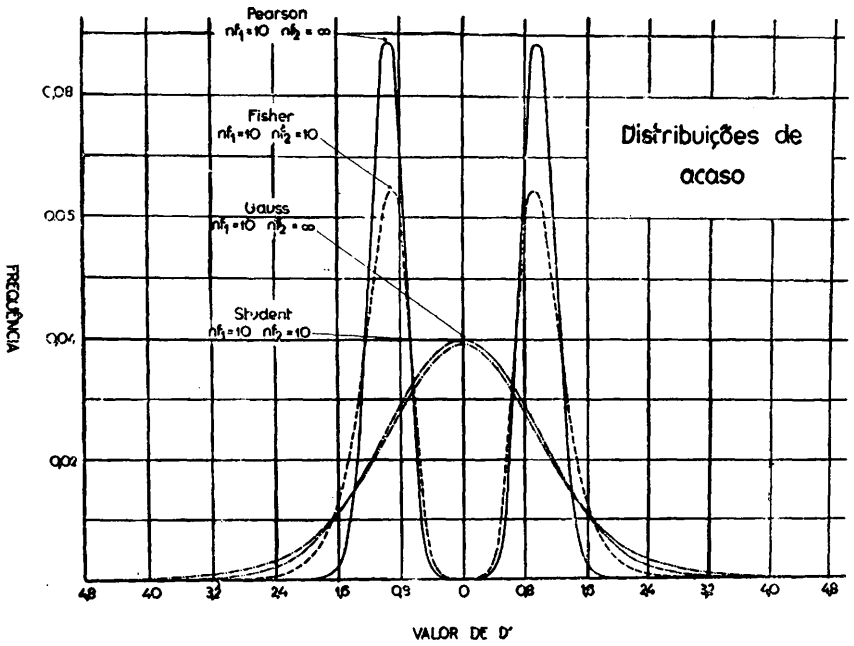

DISTRIBUIÇÃO BINOMIAL
$\left(\frac{1}{2}+\frac{1}{2}\right)^4$ σ = 1
Classes de 1 σ
37,50
25,00
25,00
6,25
6,25
$\left(\frac{1}{2}+\frac{1}{2}\right)^{64}$ σ = 4
Classes de ¼ σ
$\left(\frac{1}{2}+\frac{1}{2}\right)^{16}$ σ = 2
Classes de ½ σ
$\left(\frac{1}{2}+\frac{1}{2}\right)^{36}$ σ = 3
Classes de ⅓ σ
-3σ -2σ -1σ 0 +1σ +2σ +3σ
Frequencia da curva de Gauss; intervalos de ¼ σ
|0,1|0,2|0,3|0,6|1,0|1,7|2,7|3,9|5,3|6,8|8,2|9,3|9,9|9,9|9,3|8,2|
|0,1|0,2|0,4|0,8|1,4|2,2|3,3|4,5|6,1|7,5|8,8|9,6|9,9|9,6|8,8|

Fig. 1



Graus de liberdade

Fisher:

$$y = k_1 \cdot D^{n_1-1} : \left(1+\frac{n_1}{n_2}D^2\right)^{-\frac{n_1+n_2}{2}}$$

$1 < n_1 < \infty$, $1 < n_2 < \infty$

Pearson:

$$y = k_2 \cdot D^{n_1-1} : e^{-\frac{n_1}{2}D^2}$$

$1 < n_1 < \infty$, $n_2 = \infty$

$D = \frac{\sigma_1}{\sigma_2}$

Student:

$$y = k_3 \cdot \left(1+\frac{1}{n_2}D^2\right)^{-\frac{n_2+1}{2}}$$

$n_1 = 1$, $1 < n_2 < \infty$

Gauss:

$$y = k_4 \cdot e^{-\frac{1}{2}D^2}$$

$n_1 = 1$, $n_2 = \infty$

$D = \frac{v-\heartsuit}{\sigma}$

Fig. 2

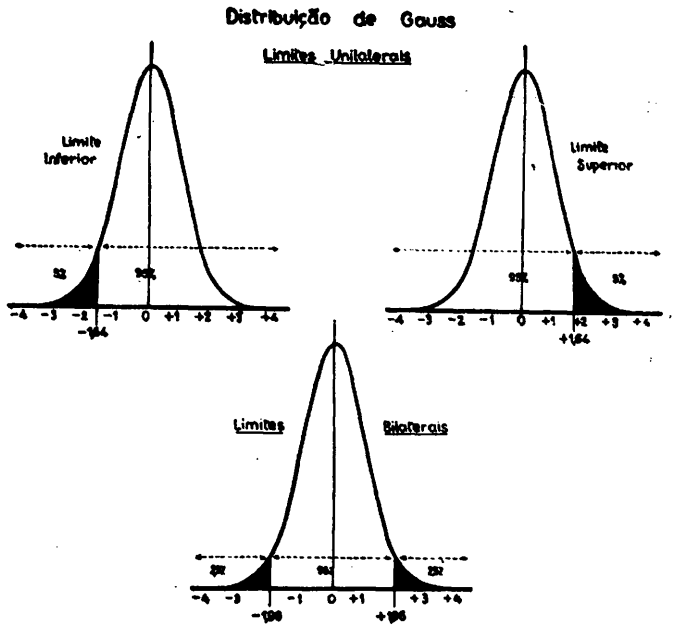


Fig. 3

Distribuição de Fisher: $nf_1 = nf_2 = 10$

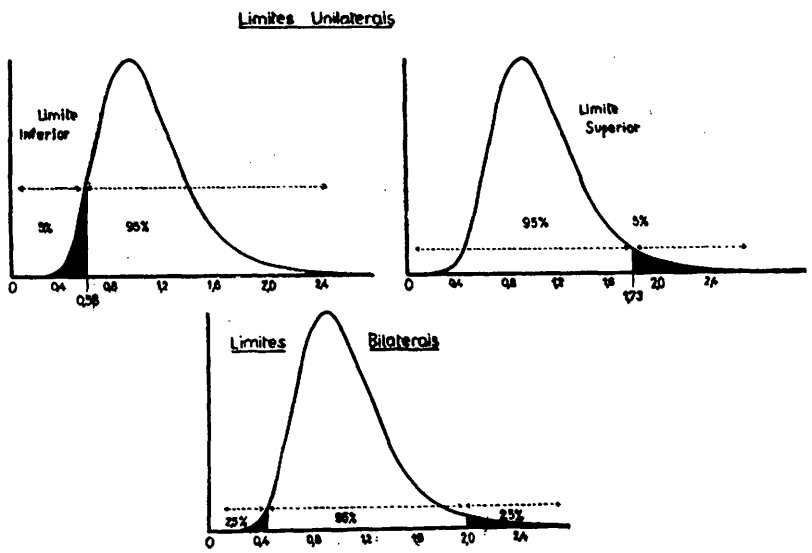


Fig. 4

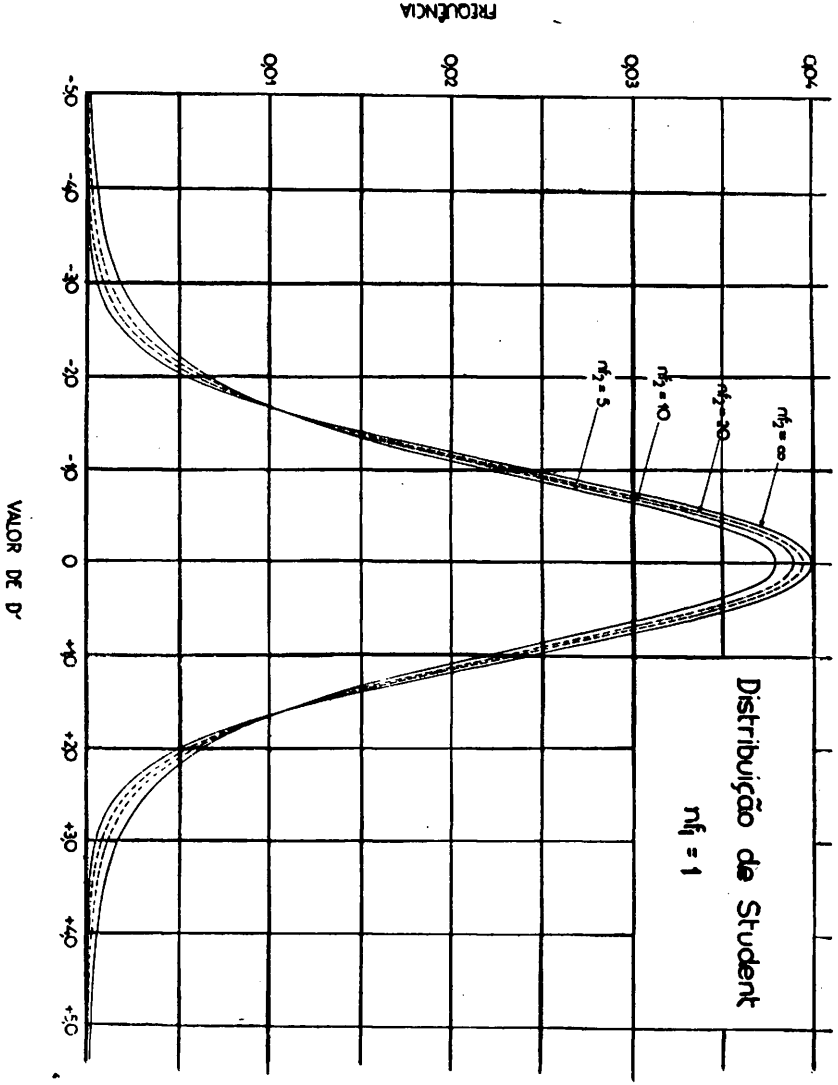
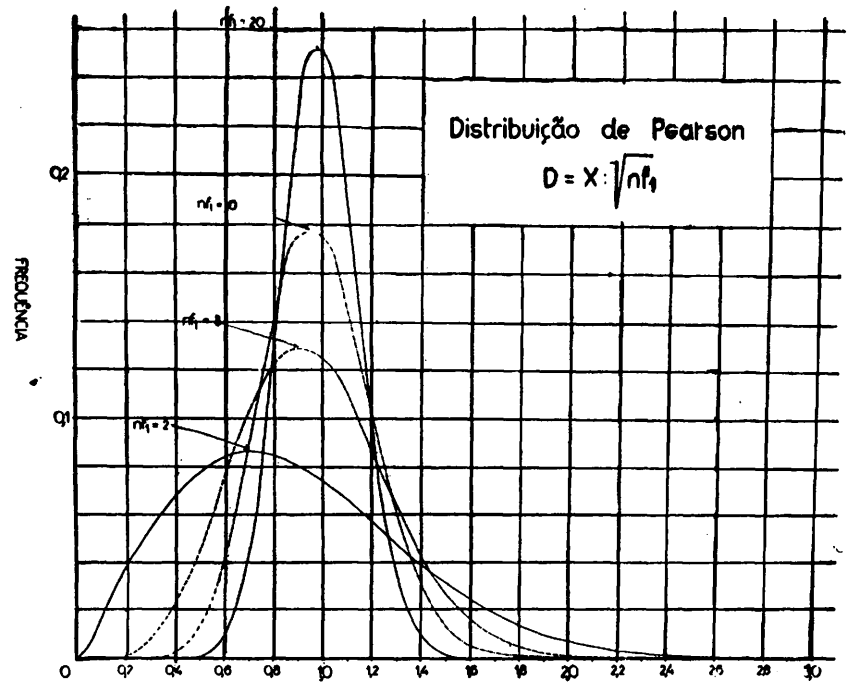

FREQUÊNCIA
0,01
0,02
0,03
0,04
-50
-40
-30
-20
-10
0
+10
+20
+30
+40
+50
VALOR DE D'
$n'_2 = 5$
$n'_2 = 10$
$n'_2 = 20$
$n'_2 = \infty$
Distribuição de Student
$n'_1 = 1$

Fig. 5


Distribuição de Pearson
D = X : √nf₁
nf₁ = 20
nf₁ = 10
nf₁ = 5
nf₁ = 2
FREQUÊNCIA
VALOR DE D

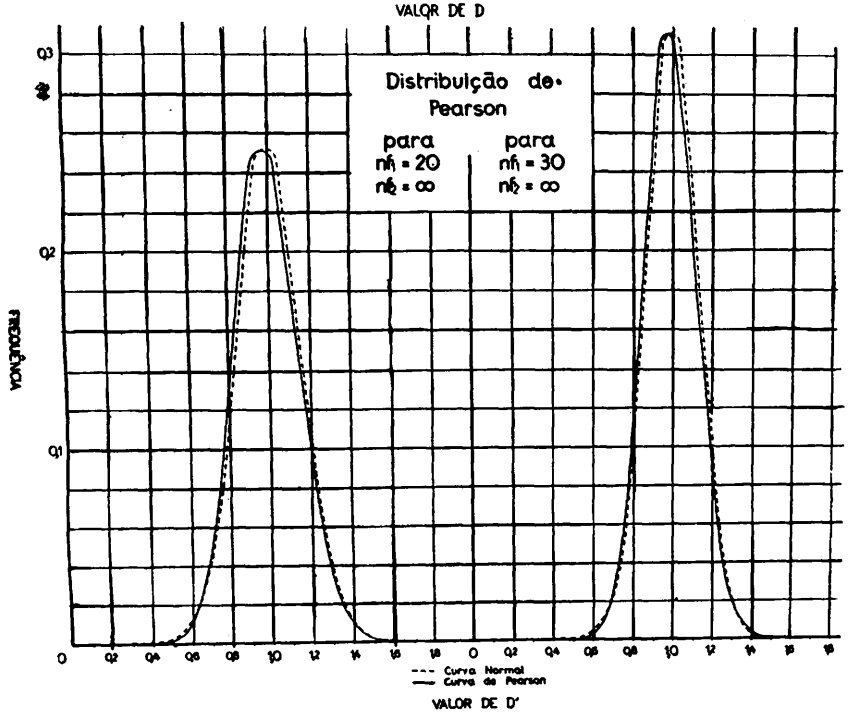

Distribuição de Pearson
para nf₁ = 20 nf₂ = ∞
para nf₁ = 30 nf₂ = ∞
FREQUÊNCIA
Curva Normal
Curva de Pearson
VALOR DE D'

Fig. 6

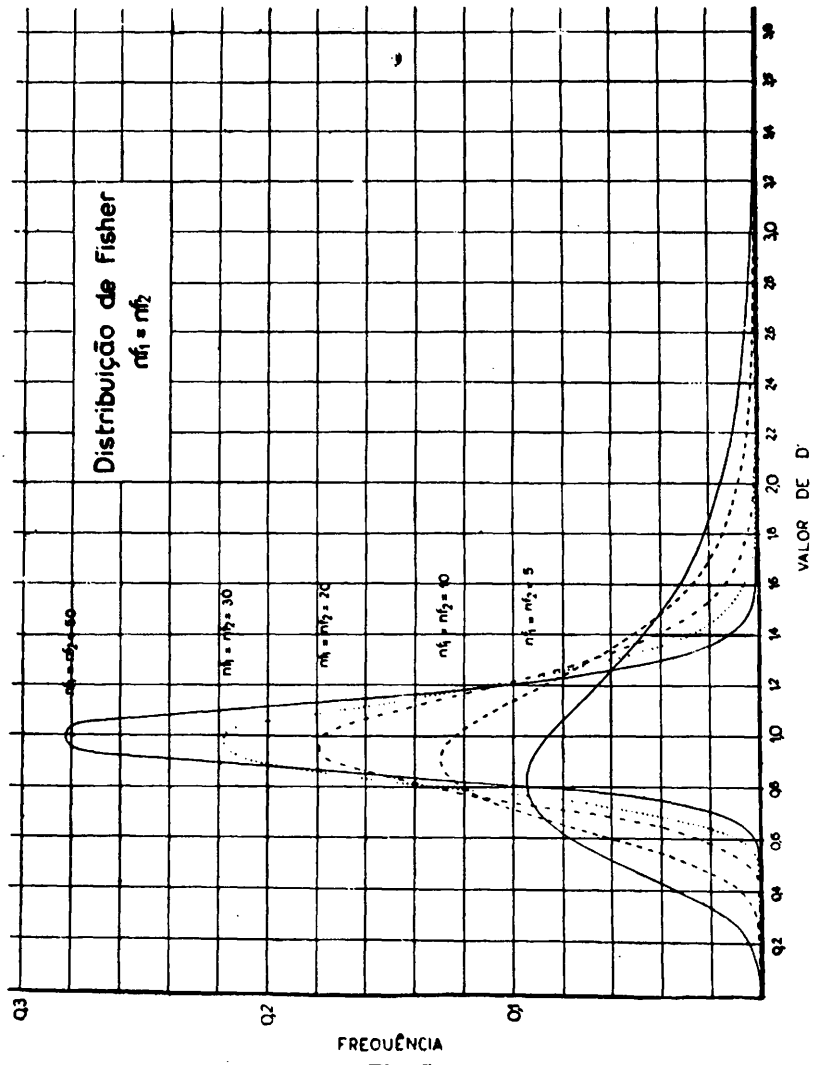

Distribuição de Fisher
$n_1 = n_2$
$n_1 = n_2 = 30$
$n_1 = n_2 = 20$
$n_1 = n_2 = 10$
$n_1 = n_2 = 5$
Q3
Q2
Q1
FREQUÊNCIA
VALOR DE D

Fig. 7

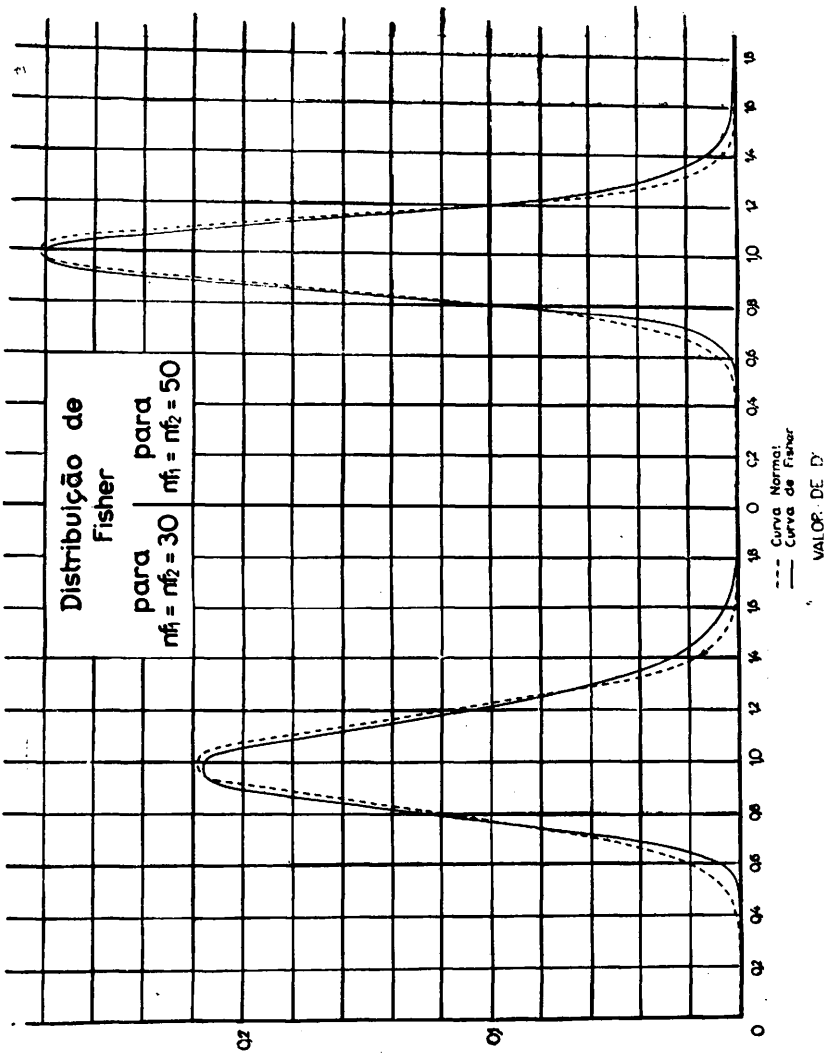
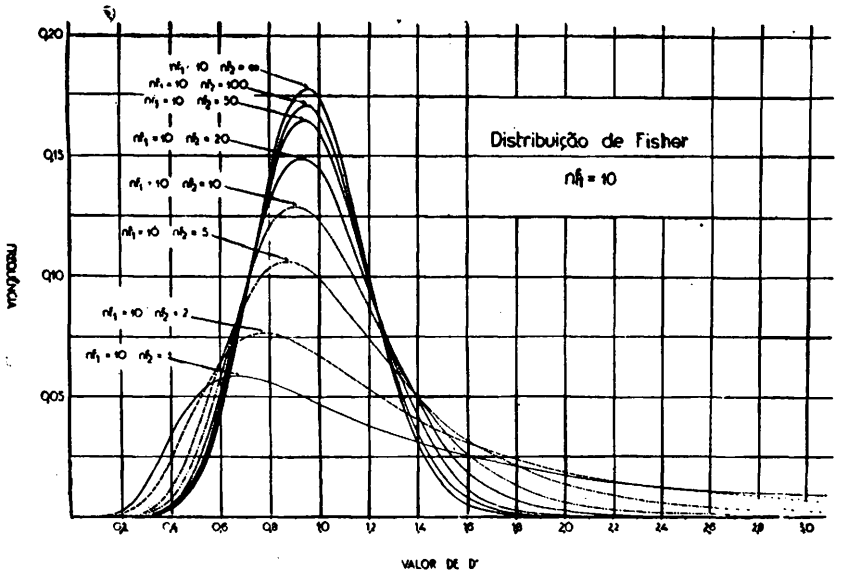

Distribuição de Fisher
para $nf_1 = nf_2 = 30$
para $nf_1 = nf_2 = 50$
Curva Normal
Curva de Fisher
VALOR DE D'
FREQUÊNCIA

Fig. 8


Distribuição de Fisher
$n_1 = 10$
FREQUÊNCIA
VALOR DE D'

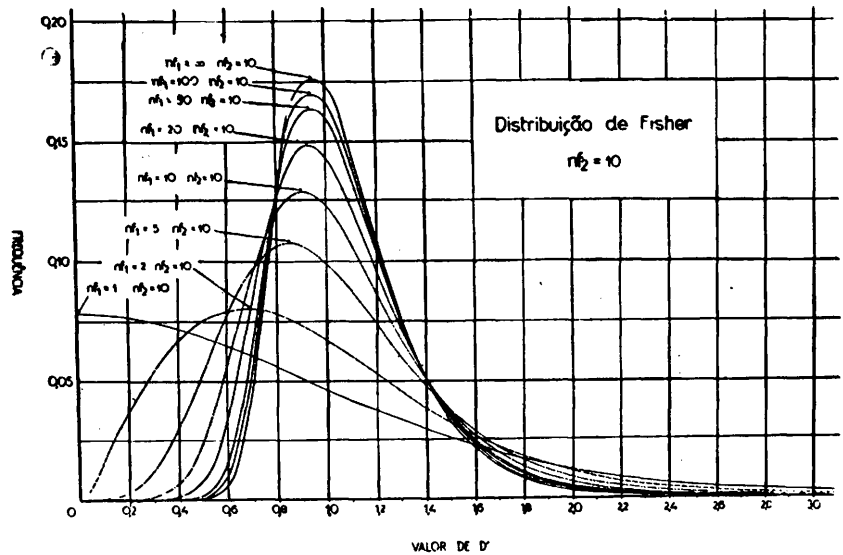

Distribuição de Fisher
$n_2 = 10$
FREQUÊNCIA
VALOR DE D'

Fig. 9

NOTA

O manuscrito do presente trabalho serviu como base à várias conferências, realizadas no Ministério de Agricultura no Rio de Janeiro e na Escola Superior Agrícola de Viçosa, em Setembro de 1944. Por razões técnicas atrazou-se a impressão.

CORREÇÕES

Infelizmente escaparam ao autor alguns enganos nas fórmulas :

Pg. 343 — Em baixo do têrmo: **lim y [Fisher]** deve constar: **n2 = inf.** em vez de **n1 = inf.**.

Pg. 348 — Em baixo do têrmo: **lim y [Fisher]** deve constar: **n2 = inf.** em vez de **n1 = 1.**

Pg. 349 — Na fórmula do têrmo k2, no início da página, deve-se ler no dividendo **n1, elevado n1 : 2**, em vez de **n2.**

Pg. 360 — Deve-se ler na segunda linha das fórmulas: "Substituimos agora D por **X : n, ½**, em vez de **X n, ½.**

Pg. 365 — No último têrmo da fórmula 7-8 falta o fator **sigma** antes da raiz quadrada no divisor.

Pg. 371 — Na fórmula, no centro da página, deve-se ler **sigma 1 : sigma D**, em vez de **sigma 1 : sigma 2.**

Pg. 380 — Na primeira fórmula da página deve-se substituir o **sinal de multiplicação** pelo de **divisão.**

Na última fórmula deve-se substituir o **sinal de multiplicação** pelo de **igualdade.**