ELZA SALVATORI BERQUÓ

SÔBRE UMA DISTRIBUIÇÃO DISCRETA DE PROBABILIDADES

Tese apresentada no concurso para provimento efetivo da CADEIRA VI - ESTATÍSTICA I - da Faculdade de Filosofia, Ciências e Letras da Universidade de São Paulo.

SÃO PAULO
1952

SÔBRE UMA DISTRIBUIÇÃO DISCRETA DE PROBABILIDADES

Tese apresentada ao concurso para provimento efetivo da CADEIRA VI - ESTATÍSTICA I - da Faculdade de Filosofia, Ciências e Letras da Universidade de São Paulo

SÃO PAULO

1952

I N T R O D U Ç Ã O

Sempre tivemos para nós que a inclusão da prova de tese em um concurso visa precìpuamente permitir à Comissão Julgadora aquilatar a eventual originalidade científica do candidato.

Êste ponto de vista constitui, no âmbito estatístico, para aquêles que têm o ensino como missão primordial, um sério óbice. Efetivamente, pela própria natureza e condições do seu trabalho, são êles conduzidos, quasi forçosamente, a perquirir o algo de novo exigido, no domínio puramente teórico. Ora, em um ambiente como o nosso - que é forçoso reconhecer ainda incipiente em questões estatísticas - o conseguir oportunidade para realizar qualquer cousa de inédito em matéria doutrinária é tarefa que demanda não só assiduidade e perseverança no estudo, como tempo de duração dificilmente previsível. Se isto não bastasse para arrefecer o entusiasmo de pesquisadores honestos em empreender a grande caminhada, teríamos ainda de lembrar o entrave decorrente de escassez bibliográfica que dificulta sobremodo a cada qual objetivar, de forma positiva, a convicção de originalidade que êle, em sã consciência, empresta, por vêzes, a certos resultados que consegue.

As considerações supra visam realçar a sa -

tisfação íntima que experimentamos quando, ao estudar certos problemas de probabilidade, imaginamos uma questão de amostragem de postos que nos conduziu a uma distribuição discreta que nos pareceu, desde logo, digna de um estudo acurado.

Começamos, então, a viver sinceramente o assunto, dando de nós mesmos os melhores e mais ingentes esforços. Como, a medida que prosseguiam as nossas indagações, fossem aparecendo certos resultados interessantes, seja de natureza matemática (por exemplo, o encontro de uma expressão geral da soma das potências do grau ℓ dos n primeiros números naturais) seja de natureza pròpriamente estatística (v.g. a generalização da distribuição para o caso multidimensional - e o que é ainda mais importante, a prova do teorema básico dos testes não paramétricos), persuadimo-nos que os frutos da tenacidade do nosso labor mereciam as honras de corporificar uma tese destinada a concorrer ao mais alto pôsto na Faculdade que, sem dúvida, é a cúpola da Universidade de São Paulo.

1 - O problema

Uma população finita contém n indivíduos ordenados de 1 a n segundo um certo critério. Toma-se uma amostra casual de k ($k \leq n$) indivíduos, sem reposição, cograduando-os segundo o pôsto por êles ocupado na população.

Pergunta-se : Qual é a probabilidade do indivíduo que na população ocupa o pôsto y vir a ocupar na amostra o pôsto h ($1 \leq h \leq k$) ?

Para que o indivíduo de pôsto y na população ocupe o pôsto h na amostra é necessário e suficiente que na amostra apareçam h-1 indivíduos com pôsto, na população, menor do que y , o indivíduo de pôsto y e k-h indivíduos com pôsto, na população, maior do que y .

Ora, h-1 indivíduos com pôsto menor do que y podendo ser escolhidos de $\binom{y-1}{h-1}$ diferentes maneiras e k-h com pôsto maior do que y de $\binom{n-y}{k-h}$ diferentes maneiras , então, o número de casos favoráveis é :

$$\binom{y-1}{h-1}\binom{n-y}{k-h}$$

quantidade que somente é diferente de zero para :

$$y = h,\ h+1,\ \ldots,\ (n-k+h)$$

Assim sendo, e desde que o número de amostras cograduadas é $\binom{n}{k}$, a probabilidade procurada -

- que indicaremos com f(y) - vale :

$$f(y) = \begin{cases} \dfrac{\binom{y-1}{h-1}\binom{n-y}{k-h}}{\binom{n}{k}} & \text{se } y = h, h+1, \ldots, n-k+h \\ 0 & \text{em contrário} \end{cases} \qquad (1)$$

A f(y) depende dos parâmetros n , h e k , tomando-se êste vocábulo no sentido matemático e não no de parâmetros populacionais . Apresenta ela certa analogia formal com a distribuição hipergeométrica, da qual difere, entretanto, pelo fato de que, enquanto nesta, a variável é a classe da combunação (denominador) , naquela, a variável é o número de elementos (numerador) .

A f(y) dá, naturalmente, a distribuição da variável aleatória " pôsto populacional que ocupa na amostra o pôsto h " quando se tomam, sem reposição, de forma casual, amostras de tamanho k de uma população constituida de n pôstos.

Como a grandeza pôsto pode ser equiparada a uma variável definida no campo dos números naturais, a interpretação supra se traduz, dizendo que a f(y) dá a distribuição amostral da variável aleatória " valor que ocupa o pôsto h " quando se tomam, sem reposição, amostras casuais de tamanho k de uma população constituida pelos n valores 1,2 ... n .

Por último, salientemos que a nossa questão difere do problema clássico da determinação da distribuição amostral do valor que ocupa um certo pôsto, nisto que, en-

quanto aqui a amostragem é sem reposição, no problema clássico, a amostra é simples.

À primeira vista, poderá causar extranheza a natureza da nossa variável aleatória. Esta impressão, entretanto, se desvanecerá lembrando que no problema básico das provas não paramétricas, a variável é dada por $\int_{x_r}^{x_s} f(x)\,dx$ ou seja, o total de área da distribuição populacional compreendida entre os dois valores x_r e x_s que ocupam na graduatória amostral os postos r e s $(s > r)$.

Como primeira importante conseqüência, notemos que a (1) definindo uma função de freqüência, a soma dos valores de f(y) estendida ao campo de variação de y vale 1. Assim sendo, a (1) dá uma demonstração estatística da fórmula matemática

$$\sum_{z=0}^{n-k} \binom{z+k-1}{k-1}\binom{n-z-k}{k-k} = \binom{n}{k}$$

obtida da f(y) pela transformação $y-h = z$.

2 - Determinação dos momentos

É fácil depreender, por uma análise ainda que perfunctória da (1), que a determinação dos momentos de potência da variável y apresenta sérias dificuldades. Pensamos, desde logo, em remover êste impecilho, calculando as referidas características por via dos momentos fato-

riais . Novamente, aqui, deparava-se uma dificuldade consistente em que $y^{[j]}$ não se prestava a simplificações convenientes com os fatoriais constantes da expressão de $E\left(y^{[j]}\right)$. Notando que no denominador desta última aparece $(n-y-k+h)!$, pensamos, então, começar a realização da tarefa constante desta epígrafe, procurando o valor de

$$E\left[(n-y-k+h)^{[j]}\right]$$

o que se processou como segue:

$$\alpha'_{[j]} = E\left[(n-y-k+h)^{[j]}\right]$$

$$= \frac{1}{\binom{n}{k}} \sum_{y=k}^{n-k+h} \binom{y-1}{h-1} (n-y-k+h)^{[j]} \binom{n-y}{k-h}$$

$$= \frac{1}{\binom{n}{k}} \sum_{y=h}^{n-k+h} \binom{y-1}{h-1} \frac{(n-y)!}{(n-y-k+h-j)!(k-h)!}$$

$$= \frac{(k-h+j)^{[j]}}{\binom{n}{k}} \sum_{y=h}^{n-k+h} \binom{y-1}{h-1} \frac{(n-y)!}{(n-y-k+h-j)!(k-h+j)!}$$

$$= \frac{(k-h+j)^{[j]}}{\binom{n}{k}} \sum_{y=h}^{n-k+h} \binom{y-1}{h-1} \binom{n-y}{k+j-h}$$

$$= \frac{(k-h+j)^{[j]}}{\binom{n}{k}} \sum_{y=h}^{n-(k+j)+k} \binom{y-1}{h-1}\binom{n-y}{k+j-h}$$

$$= \frac{(k-h+j)^{[j]}}{\binom{n}{k}} \binom{n}{k-j}$$

$$= \frac{(k-h+j)^{[j]}\ (n-k)^{[j]}}{(k+j)^{[j]}}$$

Utilizando, agora, a relação entre momentos de potências e momentos fatoriais, em tôrno de mesma origem, dada por

$$\alpha_t = \sum_{j=0}^{t} T_{tj}\ \alpha_{[j]}$$

podemos escrever para o momento de potência de ordem t, não centrado, da variável $n-y-k+h$:

$$\alpha'_t = \sum_{j=0}^{t} T_{tj}\ \frac{(k-h+j)^{[j]}\ (n-k)^{[j]}}{(k+j)^{[j]}} \qquad (2)$$

onde os T_{tj} são números de Stirling de segunda espécie.

Da (2) segue:

$$E(-y)^t = \sum_{j=0}^{t} T_{tj}\ \frac{(n+j)^{[j]}\ (-h)^{[j]}}{(k+j)^{[j]}}$$

e, portanto, simbolizando por α_ℓ o momento não centrado, de ordem ℓ, da variável y:

$$\alpha_\ell = (-1)^\ell \sum_{j=0}^{\ell} T_{\ell j} \frac{(n+j)^{[j]}\,(-k)^{[j]}}{(k+j)^{[j]}} \qquad (3)$$

Notando que os momentos centrados de ordem ℓ da variável y, a menos de $(-1)^\ell$ coincidem com os da variável $n-y-k+h$, então, simbolizando-os por μ_ℓ, de (2) segue também:

$$\mu_\ell = E\left[(n-y-k+h)-\alpha_1'\right]^\ell$$

$$= \sum_{s=0}^{\ell} \binom{\ell}{s} (-1)^s \left(\alpha_1'\right)^{\ell-s} \alpha_s'$$

$$= \sum_{s=0}^{\ell} \sum_{j=0}^{s} \binom{\ell}{s} (-1)^s T_{sj} \left[\frac{(k-h+1)(n-k)}{(k+1)}\right]^{\ell-s} \frac{(k-h+j)^{[j]}\,(n-k)^{[j]}}{(k+j)^{[j]}} \qquad (4)$$

Como casos particulares merecedores de registo explícito, estabeleçamos:

$$\gamma_1 = \frac{(n+1)h}{k+1} \qquad (5)$$

$$\mu_2 = \sum_{s=0}^{2} \sum_{j=0}^{s} \binom{2}{s} (-1)^s T_{sj} \left[\frac{(k-h+1)(n-k)}{k+1}\right]^{2-s} \frac{(k-h+j)^{[j]}\,(n-k)^{[j]}}{(k+j)^{[j]}}$$

$$= \left[\frac{(k-n+1)(n-k)}{k+1}\right]^2 - 2\left[\frac{(k-n+1)(n-k)}{k+1}\right]^2 +$$

$$+ \frac{(k-n+1)(n-k)}{k+1} + \frac{(k-n+2)(k-n+1)(n-k)(n-k-1)}{(k+2)(k+1)}$$

$$= \frac{n(n+1)(k-n+1)(n-k)}{(k+1)^2(k+2)} \qquad (6)$$

$$\mu_3 = \sum_{r=0}^{3} \sum_{j=0}^{3} \binom{3}{j}(-1)^j T_{3j} \left[\frac{(k-n+1)(n-k)}{k+1}\right]^{3-j} \frac{(n-n+1)^{[j]}(n-k)^{[3]}}{(k+1)^{[j]}}$$

$$= \left[\frac{(k-n+1)(n-k)}{k+1}\right]^3 - 3\left[\frac{(k-n+1)(n-k)}{k+1}\right] + 3\left[\frac{(k-n+1)(n-k)}{k+1}\right]^2 +$$

$$+ 3\,\frac{(k-n+2)(k-n+1)^2(n-k)^2(n-k-1)}{(k+2)(k+1)^2} - \frac{(k-n+1)(n-k)}{k+1} -$$

$$- 3\,\frac{(k-n+2)(k-n+1)(n-k)(n-k-1)}{(k+2)(k+1)} -$$

$$- \frac{(k-n+3)(k-n+2)(k-n+1)(n-k)(n-k-1)(n-k-2)}{(k+3)(k+2)(k+1)}$$

$$= -\frac{(k-p_1+1)(n-p_1)}{(p_1+1)^3(p_1+2)(p_1+3)}\left\{(p_2-p_1+1)^2\left[2(n-p_2)^2(p_2+2)(p_2+3)-\right.\right.$$

$$-3(n-p_2)(n-p_2-1)(p_2+1)(p_2+3)+(n-p_2-1)(n-p_2-2)(p_2+1)^2\Big]+$$

$$+3(p_2-p_1+1)(p_2+1)\Big[-(n-p_2)(p_2+2)(p_2+3)-(n-p_2)(n-p_2-1)(p_2+3)+$$

$$+(n-p_2-1)(p_2+1)(p_2+3)+(n-p_2-1)(n-p_2-2)(p_2+1)\Big]+$$

$$\left.+(p_2+1)^2\Big[(p_2+2)(p_2+3)+3(n-p_2-1)(p_2+3)+2(n-p_2-1)(n-p_2-2)\Big]\right\}$$

$$= -\frac{(p_2-p_1+1)(n-p_2)}{(p_1+1)^3(p_2+2)(p_2+3)}\left\{(p_2-p_1+1)^2\Big[2(n-p_2)(p_2+3)\Big\{(n-p_2)(p_2+2)-(n-p_2-1)(p_2+1)\Big\}+\right.$$

$$+(n-p_2-1)(p_2+1)\Big\{-(n-p_2)(p_2+3)+(n-p_2-2)(p_2+1)\Big\}\Big]+$$

$$+3(p_2-p_1+1)(p_2+1)\Big[-(n-p_2)(p_2+3)\Big\{(p_2+2)+(n-p_2-1)\Big\}+$$

$$+(n-p_2-1)(p_2+1)\Big\{(p_2+3)+(n-p_2-2)\Big\}\Big]$$

$$\left.+(p_2+1)^2\Big[(p_2+3)\Big\{(p_2+2)+(n-p_2-1)\Big\}+2(n-p_2-1)\Big\{(p_2+3)+(n-p_2-2)\Big\}\Big]\right\}$$

$$= -\frac{(p_2-p_1+1)(n-p_2)(n+1)}{(p_1+1)^3(p_2+2)(p_2+3)}\left\{(p_2-p_1+1)^2\Big[2(n-p_2)(p_2+3)-2(n-p_2-1)(p_2+1)\Big]+\right.$$

$$+3(p_2-p_1+1)(p_2+1)\Big[-(n-p_2)(p_2+3)+(n-p_2-1)(p_2+1)\Big]+$$

$$\left.+(p_2+1)^2\Big[(p_2+3)+2(n-p_2-1)\Big]\right\} =$$

$$= \frac{(k-p_1+1)(n-p_2)(n+1)(2n-p_2+1)}{(p_1+1)^3(p_2+2)(p_2+3)}\Big[2(p_2-p_1+1)^2-3(p_2-p_1+1)(p_2+1)+(p_2+1)^2\Big]$$

donde, finalmente :

$$\mu_3 = - \frac{(k-h+1)(n-k)(n+1)\,h\,(2n-k+1)(2h-k-1)}{(k+1)^3(k+2)(k+3)} \qquad (7)$$

Para o momento quarto centrado, tem-se :

$$\mu_4 = \sum_{s=0}^{4} \sum_{j=0}^{s} \binom{4}{s} (-1)^4 \, T_{sj} \left[\frac{(h-h+1)(n-h)}{(k+1)}\right]^{4-s} \frac{(k-h+j)^{[s]}\,(n-k)^{[s]}}{(k+j)^{[s]}} =$$

$$= \frac{(k-h+1)^4\,(n-k)^4}{(k+1)^4} - \frac{4(k-h+1)^4\,(n-k)^4}{(k+1)^4} +$$

$$+ 6\,\frac{(k-h+1)^2(n-k)^2}{(k+1)^2}\left[\frac{(k-h+1)(n-k)}{k+1} + \frac{(k-h+2)(k-h+1)(n-k)(n-k-1)}{(k+1)(k+2)}\right] -$$

$$- 4\,\frac{(k-h+1)(n-k)}{k+1}\left[\frac{(k-h+1)(n-k)}{k+1} + 3\,\frac{(k-h+2)(k-h+1)(n-k)(n-k-1)}{(k+1)(k+2)} + \right.$$

$$\left. + \frac{(k-h+3)(k-h+2)(k-h+1)(n-k)(n-k-1)(n-k-2)}{(k+1)(k+2)(k+3)}\right] +$$

$$+ \frac{(k-h+1)(n-k)}{k+1} + 7\,\frac{(k-h+2)(k-h+1)(n-k)(n-k-1)}{(k+1)(k+2)} +$$

$$+ 6\,\frac{(k-h+3)(k-h+2)(k-h+1)(n-k)(n-k-1)(n-k-2)}{(k+1)(k+2)(k+3)} +$$

$$+ \frac{(k-h+4)(k-h+3)(k-h+2)(k-h+1)(n-k)(n-k-1)(n-k-2)(n-k-3)}{(k+1)(k+2)(k+3)(k+4)}$$

$$\mu_4 = \frac{(k-R+1)(n-k)}{(k+1)^4(k+2)(k+3)(k+4)} \Big\{ (k-R+1)^3 \Big[-3(n-k)^3(k+2)(k+3)(k+4) +$$

$$+6(n-k)^2(n-k-1)(k+1)(k+3)(k+4) -$$

$$-4(n-k)(n-k-1)(n-k-2)(k+1)^2(k+4) +$$

$$+(n-k-1)(n-k-2)(n-k-3)(k+1)^3 \Big] +$$

$$+6(k-R+1)^2(k+1) \Big[(n-k)^2(k+2)(k+3)(k+4) +$$

$$+(n-k)^2(n-k-1)(k+3)(k+4) -$$

$$-2(n-k)(n-k-1)(k+1)(k+3)(k+4) -$$

$$-2(n-k)(n-k-1)(n-k-2)(k+1)(k+4) +$$

$$+(n-k-1)(n-k-2)(k+1)^2(k+4) +$$

$$+(n-k-1)(n-k-2)(n-k-3)(k+1)^3 \Big]$$

$$+(k-k+1)(k+1)^2 \Big[-4(n-k)(k+2)(k+3)(k+4) -$$

$$-12(n-k)(n-k-1)(k+3)(k+4) -$$

$$-8(n-k)(n-k-1)(n-k-2)(k+4) +$$

$$+7(n-k-1)(k+1)(k+3)(k+4) +$$

$$+18(n-k-1)(n-k-2)(k+1)(k+4) +$$

$$+12(n-k-1)(n-k-2)(n-k-3)(k+1) \Big] +$$

$$+(k+1)^3 \Big[(k+2)(k+3)(k+4) + 7(n-k-1)(k+3)(k+4) +$$

$$+12(n-k-1)(n-k-2)(k+4) +$$

$$+6(n-k-1)(n-k-2)(n-k-3) \Big] \Big\}$$

$$\mu_4 = \frac{(k-1)(n-k)}{(n-1)^4(n-2)(n-3)(n-4)} \Big\{ (n-k+1)^3 \Big[3(n-k)^2(k+1)(k-4) \Big\{ -(k-2)(n-k) + (k+1)(n-k-2) \Big\} +$$

$$+ 3(n-k)(n-k-1)(k-1)(k-4) \Big\{ (n-k)(k-1) - (n-k-2)(k+1) \Big\}$$

$$- (n-k-1)(n-k-2)(k+1)^2 \Big\{ (n-k)(k-4) - (n-k-2)(k-1) \Big\} \Big]$$

$$+ 6(n-k+1)^2(k+1) \Big[(n-k)^2(k-3)(k-4) \Big\{ (k-2) + (n-k-1) \Big\}$$

$$- 2(n-k)(n-k-1)(k+1)(k-4) \Big\{ (k-2) + (n-k-2) \Big\} +$$

$$+ (n-k-1)(n-k-2)(k+1)^2 \Big\{ (k-4) + (n-k-3) \Big\} \Big] +$$

$$+ (k-k+1)(k+1)^3 \Big[-4(n-k)(k-3)(k-4) \Big\{ (k+2) + (n-k-1) \Big\} -$$

$$- 8(n-k)(n-k-1)(k-4) \Big\{ (k+1) + (n-k-2) \Big\} +$$

$$+ 7(n-k-1)(k+1)(k-4) \Big\{ (k-3) + (n-k-2) \Big\} +$$

$$+ 12(n-k-1)(n-k-2)(k-1) \Big\{ (k-4) + (n-k-3) \Big\} \Big]$$

$$+ (k+1)^3 \Big[(k-3)(k-4) \Big\{ (k-2) + (n-k-1) \Big\} +$$

$$+ 6(n-k-1)(k-4) \Big\{ (k-3) + (n-k-2) \Big\} +$$

$$+ 6(n-k-1)(n-k-2) \Big\{ (k-4) + (n-k-3) \Big\} \Big] \Big\}$$

$$\mu_4 = \frac{(k-k+1)(n-k)(n+1)}{(k+1)^4(k+2)(k+3)(k+4)} \left\{ 3(k-k+1)^3 \left[(n-k)(k+4)(2n-k+1) + (n-k-1)(k+1)(3n-k+2) \right] + \right.$$

$$+ 6(k-k+1)^2(k+1)\left[(2n-k)(k+4)(2n-k-1) - (n-k-1)(k+1)(3n-k+2) \right] +$$

$$+ (k-k+1)(k+1)^2 \left[6(n-k)(k+4)(2n-k+1) + (n-k-1)(k+1)(11n-4k+6) \right] +$$

$$\left. + (k+1)^3 \left[(k+4)(2n-k+1) + 2(n-k-1)(3n-k+2) \right] \right\} =$$

$$= \frac{(k-k+1)(n-k)(n+1)\,k_2}{(k+1)^4(k+2)(k+3)(k+4)} \left\{ -3k(k-k+1)\left[(n-k)(k+4)(2n-k+1) - (n-k-1)(k+1)(3n-k+2) \right] + \right.$$

$$\left. + (k+1)^2 \left[(2n-k)(2n-k+1) + 2(n-k)(n+2) \right] \right\} \qquad (8)$$

Dêstes resultados seguem, por seu turno:

$$\gamma_1^2 = \frac{\mu_3^2}{\mu_2^3} = \frac{(2k-k-1)^2(2n-k+1)^2(k+2)}{k_2(n+1)(k-k+1)(n-k)(k+3)^2}$$

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} = \frac{k+2}{(k+3)(k+4)(n+1)k_2(k-k+1)(n-k)} \left\{ -3k(k-k+1)\left[(n-k)(k+4)(2n-k+1) - \right.\right.$$

$$\left. - (n-k-1)(k+1)(3n-k+2) \right] +$$

$$\left. + (k+1)^2 \left[(2n-k)(2n-k+1) + \right.\right.$$

$$\left.\left. + 2(n-k)(n+2) \right] \right\}$$

3 - Algumas propriedades .

a)- Determinação do máximo

Para que f(y) seja máxima no ponto y é necessário e suficiente que :

$$\begin{cases} f(y-1) \leq f(y) \\ f(y+1) \leq f(y) \end{cases}$$

ou sejam :

$$\binom{y-a}{k-1}\binom{n-y+1}{k-k} \leq \binom{y-1}{k-1}\binom{n-y}{k-k}$$

$$\binom{y}{k-1}\binom{n-y-1}{k-k} \leq \binom{y-1}{k-1}\binom{n-y}{k-k}$$

ou, ainda, após simplificações óbvias :

$$y \leq \frac{n(k-1)}{k-1} + 1$$

$$y \geq \frac{n(k-1)}{k-1}$$

ou, finalmente :

$$\frac{n(k-1)}{k-1} \leq y \leq \frac{n(k-1)}{k-1} + 1 \qquad (k>1) \qquad (9)$$

Os resultados obtidos mostram, pois, que se $n(h-1)$ fôr múltiplo de $k-1$, a distribuição admite dois máximos, em correspondência aos valores $\frac{n(h-1)}{k-1}$ e $\frac{n(h-1)}{k-1}+1$; se, ao contrário, $n(h-1)$ não fôr múltiplo de $k-1$, a distribuição admite um único máximo em correspondência ao inteiro compreendido entre os dois limites das desigualdades supra.

b)- Pondo na (1) $y=h$ e considerando o resultado como função de h, tem-se :

$$\varphi(h) = \frac{\binom{n-h}{k-h}}{\binom{n}{k}} \qquad h = 1, 2, \ldots, k$$

que dá a probabilidade de um elemento populacional de pôsto $\leq k$ vir a ocupar na amostra o mesmo pôsto que êle ocupa na população. É fácil mostrar que $\varphi(h)$ é decrescente e conseqüentemente máxima no ponto $h = 1$. Com efeito, tem-se :

$$\binom{n-h}{k-h} = \binom{n-h-1}{k-h-1} + \binom{n-h-1}{k-h} > \binom{n-[h+1]}{k-[h+1]}$$

c) - Em particular, se $k = n$, isto é, se a amostra é a própria população, y admite o único valor h e a distribuição torna-se :

$$f(y) = \begin{cases} 1 & \text{se } y = h \\ 0 & \text{em contrário} \end{cases}$$

que nada mais é do que uma distribuição δ.

O fato mais interessante a assinalar, no caso em aprêço, é que da (6) segue:

$$\mu_2 = 0$$

como deveria acontecer, posto que, como vimos, estamos diante de uma distribuição com tôda a massa concentrada em um único ponto.

d) - Para $k = 1$ e conseqüentemente $h = 1$, o nosso problema reduz-se à determinação da probabilidade de se retirar um único elemento da população. A distribuição torna-se :

$$f(y) = \begin{cases} \frac{1}{n} & \text{se } y = 1,2 \ldots n \\ 0 & 0 \text{ em contrário.} \end{cases}$$

o que mostra que a nossa distribuição contém, como caso particular, a distribuição uniforme.

Assim sendo, pareceu-nos de importância aproveitar alguns dos resultados obtidos para estabelecer outras tantas propriedades desta distribuição que tamanho papel desempenha seja em estatística seja no cálculo de probabilidades.

Neste sentido, citemos, de início, que da (3) resulta :

$$n\,\bar{x}_t = \sum_{y=1}^{n} y^t = (-1)^t \sum_{j=0}^{t} T_{tj}\, \frac{(n+j)^{[j+1]}\,(-1)^{[j]}}{(j+1)^{[j]}} \qquad (10)$$

com o que se obtém, de forma devéras elegante e elementar, uma expressão geral para a soma das t -ésimas potências dos n primeiros números naturais.

Em particular, fazendo $t = 1, 2, 3$ e 4, dá (10) seguem as fórmulas clássicas :

$$\sum_{y=1}^{n} y = -\sum_{j=0}^{1} T_{1j}\, \frac{(n+j)^{[j+1]}\,(-1)^{[j]}}{(j+1)^{[j]}} = \frac{n(n+1)}{2}$$

$$\sum_{y=1}^{n} y^2 = \sum_{j=0}^{2} T_{2j}\, \frac{(n+j)^{[j+1]}\,(-1)^{[j]}}{(j+1)^{[j]}} = \frac{n(n+1)(2n+1)}{6}$$

$$\sum_{y=1}^{n} y^3 = -\sum_{j=0}^{3} T_{3j}\, \frac{(n+j)^{[j+1]}\,(-1)^{[j]}}{(j+1)^{[j]}} = \frac{n^2(n+1)^2}{4}$$

$$\sum_{y=1}^{n} y^4 = \sum_{j=0}^{4} T_{4j}\, \frac{(n+j)^{[j+1]}\,(-1)^{[j]}}{(j+1)^{[j]}} = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}$$

Outro importante resultado diz respeito aos momentos centrados em tôrno da média $\bar{y}$ dos n primeiros números naturais. Tem-se, genèricamente pela (4) :

$$\frac{1}{n}\sum_{y=1}^{n}(y-\bar{y})^{l} = \sum_{s=0}^{l}\sum_{j=0}^{s}\binom{l}{s}(-1)^{s}\,T_{sj}\left(\frac{n-1}{2}\right)^{l-s}\frac{j^{[j]}\,(n-1)^{[j]}}{(j+1)^{[j]}}$$

e, em particular pelas (6), (7) e (8) :

$$\mu_2 = \frac{n^2-1}{12}$$

$$\mu_3 = 0$$

$$\mu_4 = \frac{(n^2-1)(3n^2-7)}{240}$$

e) - Sob a hipótese $h = 1$ e $k > 1$, tem-se para a distribuição do extremo inferior da graduatória amostral :

$$f(y) = \begin{cases} \dfrac{k(n-k)^{[y-1]}}{n^{[y]}} & \text{se } y = 1, 2, \cdots, (n-k+1) \\ 0 & \text{em contrário} \end{cases}$$

A (9) mostra que a moda é $y = 1$ e a expressão supra que o valor máximo de $f(y)$ é $\frac{k}{n}$.

Em conformidade com as (5), (6) e (7), tem-se :

$$\alpha_1 = \frac{n+1}{k+1}$$

$$\mu_2 = \frac{k(n+1)(n-k)}{(k+1)^2(k+2)}$$

$$\mu_3 = -\frac{k(n+1)(k-1)(2n-k+1)(n-k)}{(k+1)^3(k+2)(k+3)} < 0$$

Dos resultados acima segue que o gráfico representativo da função é da forma de J invertido.

f) - Se $h = k$ $(k > 1)$ tem-se para a distribuição do extremo superior da graduatória amostral :

$$f(y) = \begin{cases} \dfrac{\binom{y-1}{k-1}}{\binom{n}{k}} & \text{se } y = k, k+1, \ldots, n \\ 0 & \text{em caso contrário} \end{cases}$$

A (9) mostra que a moda é $y = n$ e a expressão supra que o valor máximo de $f(y)$ é k/n . Em conformidade com as (5), (6) e (7) tem-se :

$$\alpha_1 = \frac{k(n+1)}{k+1}$$

Dos resultados acima segue que o gráfico representativo da função é da forma de J .

É fácil verificar que a atual distribuição coincide com a estudada em e) quando se submete esta a uma translação de amplitude $n-1$ seguida de uma rotação de π em tôrno de um eixo vertical levantado perpendicularmente ao eixo das abcissas no seu ponto n .

g) - A última propriedade salientada em f) relacionando as distribuições dos valores que ocupam o primeiro pôsto na co e contragraduatórias pode ser fàcilmente generalizada. Efetivamente, se na (1) substituirmos o parâmtro h por $k-h+1$, a distribuição assumo a forma:

$$f_1(y) = \begin{cases} \dfrac{\binom{y-1}{k-h}\binom{n-y}{h-1}}{\binom{n}{k}} & \text{se } y=(k-h+1),\dots,(n-h+1) \\ 0 & \text{em contrário} \end{cases}$$

Comparando-se $f_1(y)$ com $f(y)$, é imediato verificar quo :

$$f(h+\delta) = f_1(n-h-\delta+1) \qquad (\delta = 0, 1, \dots, n-h)$$

De fato:

$$f(h+\delta) = \frac{\binom{h+\delta-1}{h-1}\binom{n-h-\delta}{h-h}}{\binom{n}{h}} = f_1(n-h-\delta+1)$$

Considerando-se, agora, que um indivíduo que ocupa o pôsto $k=h+1$ na cograduatória ocupa na contragraduatória o pôsto h, estabeleçamos, pois, genèricamente, que o gráfico representativo da distribuição dos valores que ocupam o pôsto h na contragraduatória coincide com o representativo da distribuição daqueles que ocupam o mesmo pôsto h mas no cograduatória, quando se submete êste último gráfico a uma translação de amplitude $n-2h+1$ seguida de uma rotação de π em tôrno de um eixo vertical levantado perpendicularmente ao eixo das abcissas no seu ponto $n-h+1$.

h)- Vamos supor aqui que h é o posto amostral mediano. Para evitar complicações formais, faremos $k=2p+1$ e, conseqüentemente $h = p+1$. A distribuição da mediana amostral é:

$$f(y) = \begin{cases} \dfrac{\binom{y-1}{p}\binom{n-y}{p}}{\binom{n}{2p+1}} & \text{se } y = p+1, \cdots, n-p \\ 0 & \text{em contrário} \end{cases}$$

Da (9) segue que a moda y deverá satisfazer as desigualdades:

$$\frac{n}{2} \leq y \leq \frac{n}{2} + 1$$

o que mostra que se n fôr par, haverá dois máximos e se n fôr ímpar, um único, dado por $\frac{n+1}{2}$.

Em conformidade com a (5), tem-se :

$$\alpha_1 = E(y_d) = \frac{n+1}{2}$$

oque, em palavras, pode-se traduzir, dizendo que a média dos valores dos postos populacionais que vêm a ocupar na amostra o pôsto mediano é igual ao posto mediano populacional.

Das (6), (7) e (8) deduzem-se :

$$\mu_2 = \frac{(n+1)(n-2p-1)}{4(2p+3)}$$

$$\mu_3 = 0$$

$$\mu_4 = \frac{(n+1)(n-2p-1)}{16(2p+3)(p+3)(2p+5)}\left(3n^2p - 6np^2 - 10p^2 - 12np - 27p + 6n^2 - 14\right)$$

$$\gamma_1 = 0$$

$$\gamma_2 = \frac{2p+3}{(n+1)(p+2)(2p+5)(n-2p-1)}\left(3n^2p - 6np^2 - 10p^2 - 12np - 27p + 6n^2 - 14\right)$$

Desta última expressão resulta que para $p \to \infty$ e conseqüentemente $n \to \infty$, a distribuição é me ocúrtica ; com efeito, como nas variáveis n e p , a ordem do infinito do denominador de γ_2 é 4 , tem-se:

$$\lim_{n,p\to\infty} \gamma_2 = \lim_{n,p\to\infty} \frac{(2p+3)(3n^2p - 6np^2)}{(n+1)(p+2)(2p+5)(n-2p-1)} =$$

$$= \lim_{n,p\to\infty} \frac{(2+\frac{3}{p})(3 - 6\frac{p}{n})}{(1+\frac{1}{n})(1+\frac{2}{p})(2+\frac{5}{p})(1-2\frac{p}{n}-\frac{1}{n})} = 3$$

Resumindo, estabeleçamos: a distribuição dos valores que vêm a ocupar na amostra o pôsto mediano é simétrica em tôrno do ponto $\frac{n+1}{2}$, tem variância dada por $\frac{(n+1)(n-2p-1)}{4(2p+3)}$ e é mesocúrtica para amostras suficientemente grandes.

Para finalizar, demonstremos que para $p \to \infty$, a variável reduzida :

$$t = \frac{y - \frac{n+1}{2}}{\sqrt{\frac{(n+1)(n-2p-1)}{4(2p+3)}}}$$

é normal, com parâmetros 0 e 1.

Com efeito, pondo:

$$x = t \sqrt{\frac{(n+1)(2n-p-1)}{4(2p+3)}}$$

tem-se:

$$P\left(x + \frac{n+1}{2}\right) = P_y = \frac{\binom{\frac{n+2x-1}{2}}{p}\binom{\frac{n-2x-1}{2}}{p}}{\binom{n}{2p+1}}$$

Ora, o valor da probabilidade máxima sendo dada por:

$$P\left(\frac{n+1}{2}\right) = P_0 = \frac{\binom{\frac{n-1}{2}}{p}\binom{\frac{n-1}{2}}{p}}{\binom{n}{2p+1}}$$

pode-se escrever :

$$\frac{P_y}{P_0} = \frac{\binom{\frac{n+2x-1}{2}}{p}\binom{\frac{n-2x-1}{2}}{p}}{\binom{\frac{n-1}{2}}{p}\binom{\frac{n-1}{2}}{p}} =$$

$$= \frac{\left(\frac{n+2x-1}{2}\right)!\left(\frac{n-2x-1}{2}\right)!\left(\frac{n-2p-1}{2}\right)!\left(\frac{n-2p-1}{2}\right)!}{\left(\frac{n+2x-1-2p}{2}\right)!\left(\frac{n-2x-1-2p}{2}\right)!\left(\frac{n-1}{2}\right)!\left(\frac{n-1}{2}\right)!}$$

do que segue, por utilização da fórmula de Stirling $n! \cong n^n \cdot e^{-n}\sqrt{2\pi n}$:

$$\frac{P_y}{P_0} \cong \frac{\left(\frac{n+2x-1}{2}\right)^{\frac{n+2x}{2}}\left(\frac{n-2x-1}{2}\right)^{\frac{n-2x}{2}}\left(\frac{n-2p-1}{2}\right)^{n-2p}}{\left(\frac{n+2x-2p-1}{2}\right)^{\frac{n+2x-2p}{2}}\left(\frac{n-2x-2p-1}{2}\right)^{\frac{n-2x-2p}{2}}\left(\frac{n-1}{2}\right)^{n}} =$$

$$= \frac{\left(1+\frac{2x}{n-1}\right)^{\frac{n}{2}+x}\left(1-\frac{2x}{n-1}\right)^{\frac{n}{2}-x}\left(\frac{n-1}{2}\right)^{\frac{n}{2}+x}\left(\frac{n-1}{2}\right)^{\frac{n}{2}-x}\left(\frac{n-2p-1}{2}\right)^{n-2p}}{\left(1+\frac{2x}{n-2p-1}\right)^{\frac{n}{2}+x-p}\left(\frac{n-2p-1}{2}\right)^{\frac{n}{2}+x-p}\left(1-\frac{2x}{n-2p-1}\right)^{\frac{n}{2}-x-p}\left(\frac{n-2p-1}{2}\right)^{\frac{n}{2}-x-p}\left(\frac{n-1}{2}\right)^{n}} =$$

$$= \frac{\left(1+\frac{2x}{n-1}\right)^{\frac{n}{2}+x}\left(1-\frac{2x}{n-1}\right)^{\frac{n}{2}-x}}{\left(1+\frac{2x}{n-2p-1}\right)^{\frac{n}{2}+x-p}\left(1-\frac{2x}{n-2p-1}\right)^{\frac{n}{2}-x-p}}$$

Assim sendo, tem-se :

$$\log \frac{P_y}{P_o} = \left(\frac{n}{2}+x\right)\log\left(1+\frac{2x}{n-1}\right)+\left(\frac{n}{2}-x\right)\log\left(1-\frac{2x}{n-1}\right)-$$

$$-\left(\frac{n}{2}+x-p\right)\log\left(1+\frac{2x}{n-2p-1}\right)-\left(\frac{n}{2}-x-p\right)\log\left(1-\frac{2x}{n-2p-1}\right)=$$

$$=\left(\frac{n}{2}+x\right)\left[\frac{2x}{n-1}-\frac{4x^2}{2(n-1)^2}+\cdots\right]-\left(\frac{n}{2}-x\right)\left[\frac{2x}{n-1}+\frac{4x^2}{2(n-1)^2}+\cdots\right]-$$

$$-\left(\frac{n}{2}+x-p\right)\left[\frac{2x}{n-2p-1}-\frac{4x^2}{2(n-2p-1)^2}+\cdots\right]+\left(\frac{n}{2}-x-p\right)\left[\frac{2x}{n-2p-1}+\frac{4x^2}{2(n-2p-1)^2}+\cdots\right]=$$

$$=-n\left[\frac{4x^2}{2(n-1)^2}+\frac{16x^4}{4(n-1)^4}+\cdots\right]+n\left[\frac{4x^2}{2(n-2p-1)^2}+\frac{16x^4}{4(n-2p-1)^2}+\cdots\right]+$$

$$+2x\left[\frac{2x}{n-1}+\frac{8x^3}{3(n-1)^3}+\cdots\right]-2x\left[\frac{2x}{n-2p-1}+\frac{8x^3}{3(n-2p-1)^3}+\cdots\right]-$$

$$-2p\left[\frac{4x^2}{2(n-2p-1)^2}+\frac{16x^4}{4(n-2p-1)^4}+\cdots\right]$$

Considerando que para n e p tendentes ao infinito, x é um infinito de ordem 1/2 , então :

$$\lim_{n,p\to\infty}\log\frac{P_y}{P_o}=\lim_{n,p\to\infty}\frac{4t^2(n+1)(n-2p-1)}{4(2p+3)}\left\{-\frac{n}{2(n-1)^2}+\frac{n}{2(n-2p-1)^2}+\frac{1}{n-1}-\right.$$

$$\left.-\frac{1}{n-2p-1}-\frac{p}{(n-2p-1)^2}\right\}$$

$$=t^2\left(\frac{1}{4}-\frac{1}{4}-\frac{1}{2}-\frac{1}{2}+\frac{1}{2}\right)=-\frac{t^2}{2}$$

donde, finalmente:

$$\lim_{n,p\to\infty}\frac{P_y}{P_o}=e^{-t^2/2}$$ c.s.q.d.

4 - A grande importância prática da nossa distribuição consiste em fornecer a base para inferências lí-

citas quanto às possíveis posições na população de um elemento que na amostra ocupa o pôsto h . Assim, por exemplo, podemos afirmar que a observação de posto h na amostra tem probabilidade 1 de ocupar um dos postos populacionais do intervalo

$$h \longmapsto n-k+h$$

pois êste é precisamente o campo de definição de y .

Êste intervalo, em geral, não é desejável, devido a sua amplitude, com o que somos levados à procura de um novo, mais restrito, constituido por um conjunto de valores y tendo um total de probabilidade de ocupar na amostra o pôsto h dado por P .

Por seu turno, a solução dêste último problema está na dependência da função cumulativa de y , isto é, do valor de :

$$\frac{1}{\binom{n}{k}} \sum_{y=k}^{k+m} \binom{y-1}{k-1}\binom{n-y}{n-k} \quad (m=0,1,\ldots,n-k) \qquad (11)$$

Como sempre acontece com as distribuições discretas, também aqui não é possível obter uma expressão de manuseio cômodo para a (11), razão pela qual, escolhido o tipo de intervalo - o que naturalmente depende de h * - teremos de

---

* Devido à forma da distribuição, parece-nos lógico que segundo h esteja mais próximo de 1, $k+1/2$, k , as regiões de rejeição sejam, respectivamente, a cauda superior, as caudas superior e inferior e a cauda inferior.

calcular, uma por uma, as probabilidades dos y da região a ser subtraida de h n-k+h , até que sua soma perfaça um total o mais aproximado possível de 1-P .

Com a finalidade únicas de concretizar estas considerações, vamos abordar aqui a hipótese de h ser o pôsto amostral mediano e supor novamente k = 2p+1 . Tendo em vista que no caso em aprêço, a distribuição é simétrica em tôrno do ponto médio do intervalo p+1 ⟼ n-p , isto é, de n+1/2 , parece-nos lógico que a região de rejeição seja dada pelas duas caudas da distribuição. Adotado êste ponto de vista, o nosso problema resume-se na determinaçaõ de um valor de m tal que :

$$\sum_{y=p+1}^{(p+1)+m} \frac{\binom{y-1}{p}\binom{n-y}{p}}{\binom{n}{2p+1}} = \sum_{\delta=0}^{m} \frac{\binom{\delta+p}{p}\binom{n-\delta-p-1}{p}}{\binom{n}{2p+1}} \simeq \frac{1-P}{2}$$

Para o particular caso n = 40 , p = 10 e P = 0,95, teriamos de encontrar um valor de m tal que:

$$E_m = \sum_{\delta=0}^{m} \frac{\binom{\delta+10}{10}\binom{29-\delta}{10}}{\binom{40}{21}} \simeq 0,025$$

Ora, tem-se:

$E_0 = 0,015\%$ $\quad E_1 = 0,015\% + 0,11\% = 0,125\%$

$E_2 = 0,125\% + 0,42\% = 0,545\%$ $\quad E_3 = 0,545\% + 1,16\% = 1,705\%$

$E_4 = 1,705\% + 2,5\% = 4,205\%$

Dêstes resultados seguem $m = 3$ e, portanto, o intervalo de confiança 99 % dado por 15 —— 26

Como vemos, a construção do intervalo procurado exige, mesmo neste exemplo simples, não pequeno trabalho de cálculo. Em casos ainda mais complicados, para obviar êste inconveniente, poderíamos, naturalmente, lembrar de expedientes outros, como sejam tabelas de fatoriais, emprêgo de logarítmos, fórmulas de Stirling, etc...

Voltando ao caso do pôsto mediano, além da utilização da distribuição limitante, poderíamos ainda lançar mão da propriedade empírica, segundo a qual em distribuições simétricas, a amplitude $\mu \pm 2\sigma$ abrange cêrca de 95 % das observações. Êste último expediente nos levaria ao intervalo :

$$\frac{n+1}{2} - \sqrt{\frac{(n+1)(n-2p-1)}{2p+3}} \quad \text{——} \quad \frac{n+1}{2} + \sqrt{\frac{(n+1)(n-2p-1)}{2p+3}}$$

que, no exemplo numérico apresentado torna-se $\frac{41}{2} \pm \sqrt{\frac{41 \cdot 19}{23}}$ ou seja 14,7 —— 26,3 que, neste caso, coincide com o intervalo exato.

5 - Na epígrafe 1 salientamos que devido à correspondência biunívoca entre os números naturais e a grandeza postos, a solução do problema objeto dêste trabalho podia ser interpretada como dando a distribuição da variável aleatória " valor que ocupa o pôsto h " quando se amostra, de forma casual, sem reposição, uma população cons-

tituida dos n primeiros números naturais. Esta hipótese da população ser dada pelos valores i (i = 1,2... n) pode ser substituida pela mais geral de que a população conste dos elementos a+r , a+2r , ... a+nr de uma progressão aritmética.

De fato, tem-se imediatamente para distribuição do valor a+ry que ocupa o pôsto amostral h :

$$P(a+ry) = P(y) = \frac{\binom{y-1}{h-1}\binom{n-y}{k-h}}{\binom{n}{k}} \qquad y=h,h+1 \ldots n-k$$

Dêste resultado segue que os momentos centrados da referida variável "valor que ocupa o pôsto h " são dados pelas (6),(7) e (8) multiplicadas respectivamente por r^2 , r^3 e r^4 , ao passo que se tem para a média dela :

$$E(a+ry) = a + \frac{h(n+1)}{k+1} r$$

e para sua moda o valor (ou valores) que se obtém ao substituir em a+ry , y pelo inteiro (ou inteiros) definido pelas desigualdades (9) .

Não é, entretanto, a generalização acima, a que nos parece mais importante, mas sim a que resulta da resoluçaão do problema abaixo, generalização direta do nosso primitivo problema .

" Uma população contém n indivíduos numerados de 1 a n , segundo um certo critério. Toma-se uma amostra casual de k ($k \leq n$) indivíduos, sem reposição,

o cogradua-se segundo o pôsto por êles ocupado na população. Pergunta-se : qual a probabilidade de que os indivíduos que na população ocupam os postos $y_1 < y_2 < \cdots < y_r$, $(r \leq k)$, virem a ocupar na amostra os postos respectivos 1 h_1 $\cdots$ h_r ?

Para que o evento se verifique é necessário e suficiente que na amostra apareçam $h_1 - 1$ indivíduos com postos na população menor do que y_1 , os indivíduos de postos $y_1 < y_2 < \cdots < y_r$, $(h_j - h_{j-1} - 1)$ com $j=2,3\ldots r$ indivíduos com postos, na população, compreendidos entre y_{j-1} e y_j e finalmente, $k - (h_1 - 1) - \sum_{j=2}^{r} (h_j - h_{j-1} - 1) - r = k - h_r$ indivíduos de postos, na população, maiores do que h_r .

Ora, $h_1 - 1$ indivíduos com postos, na população, menores do que y_1 podendo ser escolhidos de $\binom{y_1 - 1}{h_1 - 1}$ diferentes maneiras, $h_j - h_{j-1} - 1$ com postos, na população, compreendidos entre y_{j-1} e y_j podendo ser escolhidos de $\binom{y_j - y_{j-1} - 1}{h_j - h_{j-1} - 1}$ $j=2,3,\ldots r$ diferentes maneiras, e os $k - h_r$ de postos, na população, maiores do que y_r de $\binom{n - y_r}{k - h_r}$ diferentes maneiras, a probabilidade procurada vale, desde que existem $\binom{n}{k}$ amostras possíveis :

$$f(y_1, \ldots, y_r) = \frac{\binom{y_1 - 1}{h_1 - 1}}{\binom{n}{k}} \left\{ \prod_{j=2}^{r} \binom{y_j - y_{j-1} - 1}{h_j - h_{j-1} - 1} \right\} \binom{n - y_r}{k - h_r} \qquad (12)$$

com $y_\beta = h_\beta, h_\beta + 1, \ldots, y_{\beta+1} - (h_{\beta+1} - h_\beta)$ $\beta = 1, 2, \ldots, r-1$

$y_r = h_r, h_r + 1, \ldots, n - k + h_r$

É fácil demosstrar que :

$$I = \sum_{y_n = k_n}^{n-k+k_n} \sum_{y_{n-1} = k_{n-1}}^{y_n - (k_n - k_{n-1})} \cdots \sum_{y_1 = k_1}^{y_2 - (k_2 - k_1)} f(y_1, \ldots, y_n) = 1$$

De fato :

$$I = \frac{1}{\binom{n}{k}} \sum_{y_n = k_n}^{n-k+k_n} \binom{n - y_n}{k - k_n} \sum_{y_{n-1} = k_{n-1}}^{y_n - (k_n - k_{n-1})} \binom{y_n - y_{n-1} - 1}{k_n - k_{n-1} - 1} \cdots$$

$$\cdots \sum_{y_2 = k_2}^{y_3 - (k_3 - k_2)} \binom{y_3 - y_2 - 1}{k_3 - k_2 - 1} \sum_{y_1 = k_1}^{y_2 - (k_2 - k_1)} \binom{y_2 - y_1 - 1}{k_2 - k_1 - 1} \binom{y_1 - 1}{k_1 - 1}$$

Aplicando sucessivamente a (1), vem :

$$I = \frac{1}{\binom{n}{k}} \sum_{y_n = k_n}^{n-k+k_n} \binom{n - y_n}{k - k_n} \sum_{y_{n-1} = k_{n-1}}^{y_n - (k_n - k_{n-1})} \binom{y_n - y_{n-1} - 1}{k_n - k_{n-1} - 1} \cdots$$

$$\cdots \sum_{y_2 = k_2}^{y_3 - (k_3 - k_2)} \binom{y_3 - y_2 - 1}{k_3 - k_2 - 1} \binom{y_2 - 1}{k_2 - 1} =$$

$$= \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$= \frac{1}{\binom{n}{k}} \sum_{y_n = k_n}^{n-k+k_n} \binom{n - y_n}{k - k_n} \binom{y_n - 1}{k_n - 1} = 1 \qquad \text{c.s.q.d.}$$

Da (12) segue :

$$f_{r\dots p}(y_r, y_{r-1}, \dots, y_p) = \sum_{y_{p-1}=k_{p-1}}^{y_p-(k_p-k_{p-1})} \cdots \sum_{y_2=k_1}^{y_3-(k_2-k_1)} f(y_1, y_2, \dots, y_r) =$$

$$= \frac{\binom{n-y_r}{k-k_r} \left\{ \prod_{j=p+1}^{r} \binom{y_j - y_{j-1} - 1}{k_j - k_{j-1} - 1} \right\} \binom{y_p - 1}{k_p - 1}}{\binom{n}{k}} \qquad (13)$$

o que mostra que a marginal das $r = p + 1$ últimas variáveis y_r , y_{r-1} , $\dots$ y_p é de forma análoga a da (12).

Da (13) segue, por seu turno, para a condicional das $p - 1$ primeiras variaveis :

$$f(y_1, y_2, \dots, y_{p-1} / y_p, \dots, y_r) = \frac{f(y_1, \dots, y_{p-1}, y_p, \dots, y_r)}{f_{r\dots p}(y_r, \dots, y_p)} =$$

$$= \frac{\binom{y_1 - 1}{k_1 - 1} \prod_{j=2}^{p} \binom{y_j - y_{j-1} - 1}{k_j - k_{j-1} - 1}}{\binom{y_p - 1}{k_p - 1}} \qquad (14)$$

Como caso particular da (12), temos, pondo $r = 2$, a distribuição bidimensional :

$$f(y_2, y_1) = \frac{\binom{y_1 - 1}{k_1 - 1} \binom{y_2 - y_1 - 1}{k_2 - k_1 - 1} \binom{n - y_2}{k - k_2}}{\binom{n}{k}} \qquad (15)$$

$$y_1 = k_1, k_1 + 1, \dots, y_2 - k_2 + k_1$$

$$y_2 = k_2, k_2 + 1, \dots, n - k + k_2$$

do que segue :

$$f_2(y_2) = \frac{\binom{n - y_2}{k - k_2} \binom{y_2 - 1}{k_2 - 1}}{\binom{n}{k}}$$

com o que recaímos, como era de se esperar, na distribuição originária.

Tem-se, também :

$$f_2(y_1/y_2) = \frac{\binom{y_1-1}{h_1-1}\binom{y_2-y_1-1}{h_2-h_1-1}}{\binom{y_2-1}{h_2-1}}$$

Fazendo na (15) $h_1 = 1$ e $h_2 = k$, vem :

$$f(y_1, y_2) = \frac{\binom{y_2-y_1-1}{k-2}}{\binom{n}{k}} \qquad \begin{matrix} y_1 = 1, 2, \ldots, y_2-k+1 \\ y_2 = k, k+1, \ldots, n \end{matrix} \qquad (16)$$

que é a distribuição conjunta dos valores extremos amostrais.

Da (16) segue, por seu turno, para as marginais :

$$f_2(y_2) = \frac{\binom{y_2-1}{k-1}}{\binom{n}{k}} \qquad y_2 = k, k+1, \ldots, n$$

$$f_1(y_1) = \frac{\binom{n-y_1}{k-1}}{\binom{n}{k}} \qquad y_1 = 1, 2, \ldots, n-k+1$$

com o que recaímos nas distribuições consideradas em f) e e) da epígrafe 4 .

Ainda da (16) segue para a distribuição da amplitude t_1 de variação amostral :

$$f(t_1) = \frac{(n-t_1)\binom{t_1-1}{k-2}}{\binom{n}{k}} \qquad t_1 = k-1, k, \ldots, n-1.$$

que é a própria (1) quando se faz $h = k - 1$.

6 - Do ponto de vista teórico, dois importantes papeis podem ser atribuidos à nossa distribuição, visto como demonstraremos os dois fatos abaixo:

1º)- através dela é possível determinar a distribuição da variável aleatória A " área sob a função de freqüência de uma população finita à esquerda do valor que numa amostra, sem reposição, ocupa o pôsto h " e provar que tal distribuição independe da função de freqüência da população;

2º)- demonstrar, no caso bastante geral de uma população enumerável, o teorema básico das provas não paramétricas, ou seja, determinar a distribuição da variável aleatória "área sob a função de freqüência de uma população à esquerda de um valor que numa amostra simples ocupa o pôsto h ".

Isto é conseguido da seguinte maneira: Vimos, pela resolução do nosso problema que um valor ocupando na população postos variaveis do h até n - k + h podem ocupar na amostra o pôsto h com probabilidades dadas pela (1). Nossa única suposição ao estabelecer tais probabilidades foi a de admitir uma graduatória na população finita amostrada, nenhuma hipótese tendo sido formulada quanto à natureza desta população, isto é, quanto a sua densidade de probabilidade que simbolisaremos por $\psi(x)$. Considerando, agora, que a área sob a $\psi(x)$ à esquerda de um valor que na amostra ocupa o pôsto h depende e só depende do pôsto por êle ocupado na população, então, a distribuição de A independe, como afirmamos, de $\psi(x)$. Além disto, subsiste a seguinte correspondência:

Valores de A	Probabilidades
$\frac{k-1}{n}$	$f(k)$
$\frac{k}{n}$	$f(k+1)$
$\vdots$	
$\frac{n-k+k-1}{n}$	$f(n-k+k)$

resultado que mostra que a distribuição de A pode ser obtida da (1) quando se submete esta à transformação

$$y = nA + 1 \qquad (17)$$

Tem-se, então :

$$P(A) = \frac{\binom{nA}{k-1}\binom{n-nA-1}{k-k}}{\binom{n}{k}}$$

$$A = \frac{k-1}{n}, \ldots, \frac{n-k+k-1}{n}$$

Suponhamos, a seguir, $n \to \infty$, com o que a amostragem sem reposição transforma numa amostragem simples. No caso em aprêço, A varia entre 0 e 1 e, tem-se, notando que o jacobiano da (17) vale n :

$$\lim_{n\to\infty} p(A) = y(A) = \lim_{n\to\infty} \frac{\binom{nA}{k-1}\binom{n-nA-1}{k-k}}{\binom{n}{k}} \cdot n$$

ou seja :

$$g(A) = \frac{k!\,A^{k-1}\,(1-A)^{k-k}}{(k-1)!\,(k-k)!}$$

que é a distribuição $\beta(k, k-k+1)$, c. s. q. d.

7 - Recordando que a variável y se identifica a $nA + 1$, então, simbolizando por A_i $(i=1,2,\ldots,r)$ a área sob a $\varphi(x)$ à esquerda do valor que na amostra ocupa o pôsto h_i, estabeleçamos, como generalização :

$$p(A_1, A_2, \ldots, A_r) = \frac{1}{\binom{n}{k}} \binom{nA_1}{h_1 - 1} \left[\prod_{j=2}^{r} \binom{nA_j - nA_{j-1} - 1}{h_j - h_{j-1} - 1} \right] \binom{n - nA_r - 1}{k - h_r}$$

$$A_\beta = \frac{h_\beta - 1}{n}, \frac{h_\beta}{n}, \ldots, \frac{A_{\beta+1} - (h_{\beta+1} - h_\beta) - 1}{n} \qquad \beta = 1, 2, \ldots, r-1$$

$$A_k = \frac{h_r - 1}{n}, \ldots, \frac{n - k + h_r - 1}{n}$$

Notando que o jacobiano é n^r, tem-se, para $n \to \infty$:

$$\lim_{n\to\infty} p(A_1, A_2, \ldots, A_k) = g(A_1, A_2, \ldots, A_r) =$$

$$= \frac{n^r \binom{nA_1}{h_1 - 1} \left[\prod_{j=2}^{r} \binom{nA_j - nA_{j-1} - 1}{h_j - h_{j-1} - 1} \right] \binom{n - nA_r - 1}{k - h_r}}{\binom{n}{k}}$$

$$= \frac{k!\,A_1^{h_1-1} \prod_{j=2}^{r} (A_j - A_{j-1})^{h_j - h_{j-1} - 1} (1 - A_r)^{k - h_r}}{(h_1 - 1)! \prod_{j=2}^{r} (h_j - h_{j-1} - 1)!\,(k - h_r)!} \qquad (18)$$

$0 \le A_1 < A_2 < \ldots < A_r \le 1$

O segundo membro da (18) é a distribuição conjunta das r áreas sob a $\psi(x)$ à esquerda dos valores que ocupam na amostra os postos h_1 numa amostragem casual simples.

Em particular, se $r = k$ e conseqüentemente $h_1 = 1$, tem-se para a distribuição das k áreas sob a $\psi(x)$ à esquerda dos k valores de uma amostra casual simples:

$$g(A_1, A_2, \ldots, A_r) = k!$$

Ainda como caso particular da (18), tem-se, supondo $r = 2$:

$$g(A_1, A_2) = \frac{k!\, A_1^{h_1-1} (A_2 - A_1)^{h_2-h_1-1} (1-A_2)^{k-h_2}}{(h_1-1)!\,(h_2-h_1-1)!\,(k-h_2)!}$$

$$0 < A_1 < 1 - A_2 < 1$$

da qual resulta, para densidade de probabilidade da variável $t = A_2 - A_1$, isto é, a área sob a $\psi(x)$ compreendida entre os valores que na amostra ocupam osppostos h_2 e h_1:

$$u(t) = \frac{k!\, t^{h_2-h_1-1} (1-t)^{k-h_2-h_1}}{(h_2-h_1-1)!\,(k-h_2+h_1)!}$$

$$0 < t < 1$$

que é novamente uma função do tipo beta, com parâmetros $h_2 - h_1$ e $k - h_2 + h_1 + 1$

Para finalizar, diremos que a importância dos resultados conseguidos nesta epígrafe na teoria dos testes não paramétricos é de tal ordem, que é desnecessário encarecê-la.

SUMÁRIO E CONCLUSÕES

I - Procurando determinar a probabilidade de um indivíduo que numa graduatória populacional com n constituintes ocupa o pôsto y vir a ocupar na amostra o pôsto h , em uma amostragem, sem reposição, de tamanho k , fomos conduzidos à expressão :

$$f(y) = \frac{\binom{y-1}{h-1}\binom{n-y}{k-h}}{\binom{n}{k}} \qquad (1)$$

II - Pela messe de resultados que fornece, a (1) deve ser enquadrada entre as mais importantes distribuições discretas de probabilidades.

III - A (1) fornece uma demonstração estatística elementar da fórmula do cálculo combinatório :

$$\sum_{z=0}^{n-k} \binom{z+h-1}{h-1}\binom{n-z}{k-h} = \binom{n}{k}$$

IV - A maneira mais adequada de conseguir as expressões dos momentos da (1) é fazê-lo por via dos momentos fatoriais da variável $n - y - k + h$, obtendo-se, então as fórmulas (3) e (4) do texto.

V - A distribuição (1) admite dois ou um único máximo, segundo $n(h-1)$ fôr ou não múltiplo de $k-1$.

VI - A (1) contém, como casos particulares, as dis-

tribuição $\mathcal{E}$ e a distribuição uniforme.

VII - Foram deduzidas expressões gerais para a soma das potências e para os momentos centrados dos n primeiros números naturais.

VIII - Como outros interessantes casos particulares da (1) foram estabelecidas as distribuições dos postos populacionais que ocupam na graduatória amostral o extremo inferior, o extremo superior e o pôsto mediano.

IX - O gráfico representativo dos valores que ocupam o pôsto h na contragraduatória amostral coincide com o representativo da distribuição dos valores que ocupam o mesmo pôsto h mas na cograduatória amostral, desde que se submeta êste a uma translação de amplitude $n-2h+1$ seguida de uma rotação de π em tôrno de um eixo vertical levantado no ponto $n-h+1$ do eixo das abcissas.

X - A distribuição limitante da distribuição dos postos populacionais que na amostra ocupam o pôsto mediano é normal.

XI - A grande importância prática da nossa distribuição consiste em fornecer a base para a construção de intervalos de confiança dando, com probabilidades prefixadas, a posição na população de um indivíduo que na amostra ocupa o posto h.

XII - Como generalização da (1) foi estabelecida, para o caso r-dimensional, a fórmula (12), a distribuição marginal das $r-p+1$ últimas variaveis desta distribui-

ção constando da (13) e a condicional das p-1 primeiras variáveis aparecendo na (14) .

XIII - Como caso particular da (12) obteve-se a distribuição bidimensional (16) por via da qual deduziu-se a distribuição da amplitude de variação amostral.

XIV - Partindo da (1) foi possível demonstrar, no caso bastante geral de uma população enumerável, a proposição básica da teoria dos testes não paramétricos.

XV - Outros importantes resultados para esta última teoria foram estabelecidos na epígrafe 7 .

BIBLIOGRAFIA

Berzolari, L. , Vivanti, G. e Gigli, D. - Enciclopedia delle matematiche elementari. Vol I-Parte II. Milano, Ulrico Hoepli, 1932.

Cramér, H. - Mathematical Methods of Statistics. Princeton, Princeton University Press, 1946

Duarte, G. G. - Contribuição para o estudo dos momentos fatoriais. Tese. São Paulo, Grafico Omega, 1950.

Kendall, M. G. - The advanced theory of Statistics. Vol I . London. J. B. Lippincott Company, 1942.

Mood, A. M. - Introduction to the theory of Statistics. New York. Mc-Graw-Hill Book Company, Inc. , 1950.

Wilks, S. S. - Mathematical Statistics. Princeton. Princeton University Press, 1946.