Dedução da Equação da Curva Normal

J. M. POMPEU MEMORIA (*)

A distribuição normal é entre tôdas as distribuições a mais importante, quer sob o ponto de vista teórico, quer sob o ponto de vista aplicado. A dedução de sua expressão analítica foi feita por processos distintos por Laplace e seu contemporâneo Gauss, mas êstes métodos não são fàcilmente encontrados nos compêndios de cálculo de probabilidades e de estatística. Geralmente, a curva normal é apresentada como o limite da distribuição binomial simétrica, processo que julgamos artificial. Realmente, quando $n \rightarrow \infty$ na distribuição $(p + q)^n$, onde $p = q$, obtém-se a curva normal, mas também tende para a normalidade a distribuição de médias oriundas de qualquer população, desde que n seja suficientemente grande.

No presente artigo, tentaremos desenvolver o método de Gauss, delineado por Sokolnikoff (2). Evitaremos "saltos" na dedução, de modo que os interessados menos dotados de um conhecimento avançado de matemática possam seguí-la sem grandes dificuldades. Aliás, outra não é nossa intenção senão a de divulgar êste processo analítico, tornando-o accessível aos profissionais da agronomia que fazem uso dos métodos estatísticos.

De um modo geral, uma variável aleatória distribui-se normalmente quando suas flutuações dependem da soma de flutuações, aproximadamente da mesma grandeza, de um grande número de causas que agem independentemente, produzindo um pequeno efeito cada uma. Tal situação é encontrada na herança de alguns caracteres quantitativos, que é determinada por grande número de genes e cuja expressão depende da ação conjunta de um complexo de fatores mesológicos. A êste tipo de distribuição obedecem também os erros acidentais, que dependem de um grande número de causas desconhecidas.

Seja $X_1, X_2, \ldots X_n$ um conjunto de medições de uma grandeza, feitas independentemente, merecendo tôdas elas o mesmo grau de confiança. Se m é a melhor estimativa des-

(*) Eng. Agrônomo, M. S., Prof. do Depto. de Engenharia Rural da ESAV.

sas medidas, os erros cometidos nas medições individuais serão respectivamente:

$$x_1 = X_1 - m \ , \ x_2 = X_2 - m , \ldots \ x_n = X_n - m$$

Considerando-se que os erros positivos e negativos são igualmente prováveis, sua soma algébrica é nula:

$$\sum_{i=1}^{n} (X_i - m) = 0 \ \therefore \sum_{i=1}^{n} X_i - n\,m = 0 \text{ , e portanto}$$

$$m = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X} \text{ , e consequentemente } f(x) = f(-x) \text{, isto é,}$$

$f(x)$ é uma função par. Além disso, como os erros maiores são menos prováveis que os menores, $f(x)$ deve ser uma função decrescente para $x \geqq 0$, e desde que erros infinitamente grandes não podem ocorrer $f(\infty) = 0$. Essas considerações conduzem a uma curva que deve ter aproximadamente a forma da fig. 1, onde as ordenadas representam as probabilidades de ocurrência dos erros : $x_1, x_2, \ldots x_n$. E' claro, pois, que a área subtendida por essa curva é dada pela integral $\int_{-\infty}^{\infty} f(x)\, dx = 1$, porque $f(x)$ é uma função de probabilidade ou distribuição.

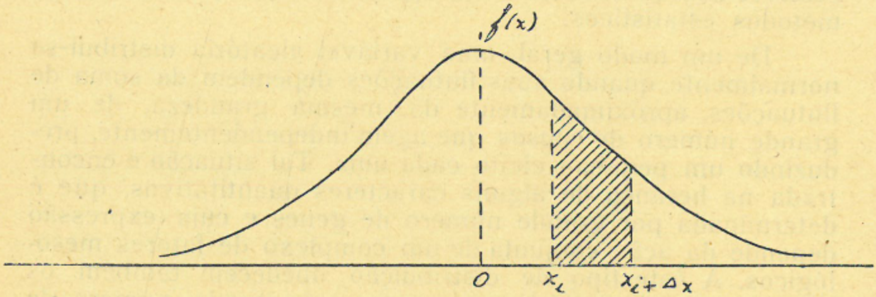

Fig. 1

A probabilidade de que um êrro ocorra entre os limites x_i e $x_i + \Delta x$ é dada pela integral $\int_{x_i}^{x_i + \Delta x} f(x)\, dx$.

Como as medidas foram feitas independentemente, a probabilidade da ocurrência simultânea dos vários erros é:

$P = f(x_1) . f(x_2) \ldots f(x_n)$ ou $P = f(X_1 - m) \ldots f(X_n - m)$ (1)

Considerando-se que a melhor estimativa das medidas é também o valor mais provável, pode-se determinar $f(x)$. Foi justamente êste o método empregado por Gauss que, assumiu axiomaticamente que a probabilidade é máxima quando $m = \bar{X}$, a média aritmética das medições. Se P é máximo, seu logarítmo também o é. Tirando-se o logarítmo neperiano de (1), derivando-se em relação a m, e anulando-se a derivada, obtem-se:

$$\frac{f'(X_1 - m)}{f(X_1 - m)} + \frac{f'(X_2 - m)}{f(X_2 - m)} + \ldots + \frac{f'(X_n - m)}{f(X_n - m)} = 0 \quad (2)$$

Fazendo-se $\frac{f'(X_i - m)}{f(X_i - m)} = g(x_i)$, pode-se escrever (2) sob a forma:

$g(x_1) + g(x_2) + g(x_3) + \ldots + g(x_n) = 0$, subordinada à condição de $\sum_{i=1}^{n} x_i = 0$. Havendo apenas duas medições tem-se: $g(x_1) + g(x_2) = 0$ e $x_1 + x_2 = 0$, donde $x_2 = - x_1$ e $g(x_1) = - g(-x_1)$. Com três medições: $g(x_1) + g(x_2) + g(x_3) = 0$ e $x_1 + x_2 + x_3 = 0$, donde $g(x_1) + g(x_2) = - g(x_3)$. Mas, $g(x_3) = - g(-x_3)$ e $- x_3 = x_1 + x_2$, portanto $- g(x_3) = g(x_1 + x_2)$, donde $g(x_1 + x_2) = g(x_1) + g(x_2)$ (3)

Obtendo-se as derivadas parciais de (3) em relação a x_1 e x_2, tem-se:
$g'(x_1) = g'(x_1 + x_2)$ e $g'(x_2) = g'(x_1 + x_2) \therefore g'(x_1) = g'(x_2)$ (4)

Como x_1 e x_2 são independentes, a equação (4) só se verifica quando $g'(x_1) = g'(x_2) = c$, portanto $g(x_1) = c\, x_1$ e $g(x_2) = c\, x_2$. Como $g(x) = \frac{f'(x)}{f(x)}$, tem-se $\frac{f'(x)}{f(x)} = c\, x$ (5), que sendo integrada fornece $f(x)$. E' conveniente, entretanto, utilizar a notação de Leibnitz antes de se fazer a integração. Seja, pois, $f(x) = y$ e $f'(x) = \frac{dy}{dx}$. A fórmula [5] reduz-se então a $\frac{dy}{dx} = c\, y\, x$. Separando-se as variáveis $\frac{dy}{y} = c\,x\,dx$, e integrando-se vem: $\ln y = \frac{cx^2}{2} + C$, onde $\underline{C}$ é a constante de integração e pode ser feita $C \equiv \ln K$, e a fórmula toma então a forma $\frac{y}{K} = e^{\frac{c\,x^2}{2}}$ ou $f(x) = K\,e^{\frac{c\,x^2}{2}}$, onde $\underline{K}$ e $\underline{c}$ são constantes arbitrárias.

Em (2) a derivada 1ª de $\underline{P}$ foi anulada, mas, esta condição tanto pode determinar um máximo como um mínimo. As duas condições que determinam o máximo de uma função, são: I) A derivada 1ª deve ser nula. II) A derivada 2ª deve ser negativa. Estas condições aplicadas ao nosso caso particular, conduzem a:

$$\text{I)}\quad \Sigma \frac{f'(x)}{f(x)} = 0 \qquad \text{II)}\quad \Sigma \frac{f(x)\,f''(x) - [\,f'(x)\,]^2}{[\,f(x)\,]^2} < 0 \quad (6)$$

$f(x) = K\,e^{\frac{c\,x^2}{2}}$, $f'(x) = c\,x\,f(x)$, $f''(x) = c\,x\,f'(x) + c\,f(x)$.

Substituindo-se êsses valores em (6), acha-se:

$$\Sigma \frac{f(x)\,[c\,x\,f'(x) + c\,f(x)] - [c\,x\,f(x)]^2}{[\,f(x)\,]^2} < 0 \text{ , donde } \Sigma\, c < 0 ,$$

e portanto $\underline{c}$ deve ser negativo. Escolhendo-se $\underline{c}$ negativo podemos, entretanto, fazer $f(x)$ decrescer ràpidamente à medida que x cresce. Fazendo-se $\frac{c}{2} \equiv -h^2$, obtem-se $f(x) = K\,e^{-h^2 x^2}$.

A constante $\underline{K}$ é determinada sob a condição de que a integral $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Mas, antes de efetuar esta integração, convém integrar-se $\int_{-\infty}^{\infty} e^{-x^2}\,dx$ que é uma função mais simples. $\int_{-\infty}^{\infty} e^{-x^2}\,dx = 2\int_{0}^{\infty} e^{-x^2}\,dx$, pois e^{-x^2} é uma função par e, portanto, a curva é simétrica em relação ao eixo das ordenadas. Fazendo-se $I = \int_{0}^{\infty} e^{-x^2}\,dx$, tem-se

$I^2 = \int_0^{\infty}\int_0^{\infty} e^{-(x^2+y^2)} dx\ dy$. Esta integral dupla é fàcilmente resolvida, transformando-se as coordenadas cartesianas em polares, o que é feito mudando-se as variáveis $\underline{x}$ e $\underline{y}$ por: $x = \varrho \cos \vartheta$, $y = \varrho \operatorname{sen} \vartheta$, $x^2 + y^2 = \varrho^2$, vide Kenney (1). O elemento de área $dA = dx\ dy$ para a função $f(x, y) = 0$ é no novo sistema de coordenadas $dA =$ $= \left[J\left(\frac{x, y}{\varrho, \vartheta}\right)\right] d\varrho\ d\vartheta$, onde $J\left(\frac{x, y}{\varrho, \vartheta}\right)$ é o símbolo que indica o valor absoluto do determinante:

$$J\left(\frac{x, y}{\varrho, \vartheta}\right) \equiv \begin{vmatrix} \frac{\delta x}{\delta \varrho} & \frac{\delta x}{\delta \vartheta} \\ \frac{\delta y}{\delta \varrho} & \frac{\delta y}{\delta \vartheta} \end{vmatrix} \text{, diferente de } 0 .$$

Êste símbolo é chamado Jacobiano ou função determinante da transformação.

Chamamos a atenção do leitor que, por falta de material tipográfico, foi empregada a letra grega δ ao invés de delta redonda para a notação de derivada parcial.

$$J\left(\frac{x, y}{\varrho, \vartheta}\right) = \frac{\delta x}{\delta \varrho} \cdot \frac{\delta y}{\delta \vartheta} - \frac{\delta x}{\delta \vartheta} \cdot \frac{\delta y}{\delta \varrho}$$

sendo $\frac{\delta x}{\delta \varrho} = \cos \vartheta$, $\frac{\delta x}{\delta \vartheta} = - \varrho \operatorname{sen} \vartheta$, $\frac{\delta y}{\delta \varrho} = \operatorname{sen} \vartheta$, e $\frac{\delta y}{\delta \vartheta} = \varrho \cos \vartheta$, donde $dA = \begin{vmatrix} \cos \vartheta & - \varrho \operatorname{sen} \vartheta \\ \operatorname{sen} \vartheta & \varrho \cos \vartheta \end{vmatrix} d\varrho\ d\vartheta =$ $= (\varrho \cos^2 \vartheta + \varrho \operatorname{sen}^2 \vartheta)\ d\varrho\ d\vartheta = \varrho d\varrho\ d\vartheta$. O elemento de integração $\underline{dx\ dy}$ torna-se $\underline{\varrho d\varrho\ d\vartheta}$ e os limites de integração são 0 e ∞ para ϱ, e 0 e $\frac{\pi}{2}$ para ϑ .

$$I^2 = \int_0^{\infty}\int_0^{\frac{\pi}{2}} e^{-\varrho^2} \varrho d\varrho\ d\vartheta = \frac{\pi}{2}\int_0^{\infty} e^{-\varrho^2} \varrho d\varrho = \frac{\pi}{4}\left[-e^{-\varrho^2}\right]_0^{\infty} = \frac{\pi}{4},$$

donde $I = \frac{\sqrt{\pi}}{2}$, e a integral $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$. Aplicando-se êste resultado em $K \int_{-\infty}^{\infty} e^{-h^2 x^2} dx = 1$, obtem-se

$$K = \frac{1}{\int_{-\infty}^{\infty} e^{-h^2 x^2} dx} = \frac{1}{\int_{-\infty}^{\infty} e^{-h^2 x^2} d(hx)} = \frac{h}{\sqrt{\pi}}$$, donde obtem-se:

$f(x) = \frac{h}{\sqrt{\pi}} e^{-h^2 x^2}$ (7). O máximo de (7) verifica-se para $x = 0$ e é $\frac{h}{\sqrt{\pi}}$. A constante h é chamada "módulo de precisão" na curva dos erros, pois, quanto maior h a curva decresce mais ràpidamente do seu valor máximo $\frac{h}{\sqrt{\pi}}$, de modo que a probabilidade de grandes erros é pequena. Isto é ilustrado gràficamente na figura 2.

Convém lembrar que $x = X - m$, sendo $m = \overline{X}$ a média da amostra. Entretanto, para se lidar com um modêlo matemático aplicável a todos os casos, o número n de dados deve ser considerado infinitamente grande, o que equivale a lidar com uma população hipotética de média

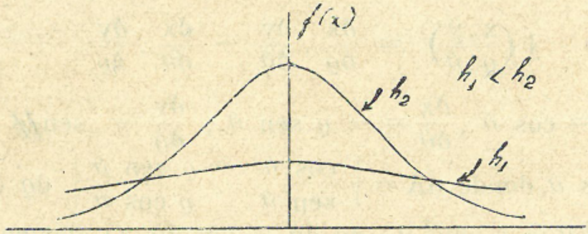

Fig. 2

aritmética μ. A melhor estimativa de μ é $\overline{X}$, média da amostra. Assim, a fórmula (7) pode ser expressa do seguinte modo: $f(x) = \frac{h}{\sqrt{\pi}} e^{-h^2 (X - \mu)^2}$ (8). A constante h, pode, por sua vez, ser expressa em função da variância σ^2, que juntamente com μ constituem os dois parâmetros da distribuição normal.

Para isso, será necessário primeiramente definir o que se entende por «expectância», vide Uspensky (3). A média aritmética de uma distribuição contínua ou a sua expectância é: $\mu = E(X) = \int_{-\infty}^{\infty} X\, f(X)\, dX$, e é às vezes chamada de 1° momento em relação à origem. A letra E é o símbolo próprio para designar a expectância, A expectância da distribuição expressa na fórmula (8) é então:

$$E(X) = \frac{h}{\sqrt{\pi}} \int_{-\infty}^{\infty} X\, e^{-h^2 (X-\mu)^2}\, dX \text{ , que desenvolvendo}$$

obtem-se:

$$E(X) = \int_{\infty}^{\infty} \frac{h}{\sqrt{\pi}} (X-\mu)\, e^{-h^2 (X-\mu)^2}\, dX + \frac{\mu h}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-h^2 (X-\mu)^2}\, dX$$

A integral $\int_{-\infty}^{\infty} (X-\mu)\, e^{-h^2 (X-\mu)^2}\, dX$ é da forma

$$\int_{-\infty}^{\infty} u\, e^{-h^2 u^2}\, du = \frac{1}{2h^2} \int_{-\infty}^{\infty} 2\, h^2 u\, e^{-h^2 u^2}\, du = \frac{1}{2h^2} \times$$

$$\times \left[-e^{-h^2 u^2} \right]_{-\infty}^{\infty} = 0 \text{ e } \int_{-\infty}^{\infty} e^{-h^2 (X-\mu)^2}\, dX = \frac{\sqrt{\pi}}{h} ,$$

como já se viu anteriormente. Portanto, $E(X) = \mu$. A variância $\sigma^2 = E(X-\mu)^2$ é também chamada 2ª momento em relacão à média, e pode ser escrita do seguinte modo:

$$\sigma^2 = \int_{-\infty}^{\infty} x^2\, f(x)\, dx = \frac{h}{\sqrt{\pi}} \int_{-\infty}^{\infty} x^2\, e^{-h^2 x^2}\, dx \text{ . Esta integração}$$

pode ser resolvida por partes, fazendo-se:

$$dv = x\, e^{-h^2 x^2}\, dx \text{ , } u = x \text{ , donde } v = \int_{-\infty}^{\infty} x\, e^{-h^2 x^2}\, dx$$

$$v = \frac{1}{2h^2} \left[-e^{-h^2 x^2} \right]_{-\infty}^{\infty} \text{, donde } \sigma^2 = \frac{x}{2h^2} \left[-e^{-h^2 x^2} \right]_{-\infty}^{\infty} -$$

$$-\frac{1}{2h^2} \int_{-\infty}^{\infty} -e^{-h^2 x^2}\, dx \text{ , como } \int_{-\infty}^{\infty} e^{-h^2 x^2}\, dx = \frac{\sqrt{\pi}}{h} \text{ , tem-se}$$

$$\text{que } \sigma^2 = \frac{1}{2h^2} \left[-x\, e^{-h^2 x^2} \right]_{-\infty}^{\infty} + \frac{1}{2h^2} \times \frac{\sqrt{\pi}}{h} \text{ ;}$$

$\left[-x\, e^{-h^2x^2} \right]_{-\infty}^{\infty}$ conduz a um valor indeterminado. Esta indeterminação pode ser levantada pela regra de L'Hospital:

$$\lim_{x \to \infty} \left[\frac{-x}{e^{h^2x^2}} \right] = \lim_{x \to \infty} \left[\frac{-1}{2h^2 x e^{h^2x^2}} \right] = 0 \text{ e o mesmo}$$

resultado é obtido quando $x \to -\infty$ Obtém-se assim

$$\sigma^2 = \frac{h}{\sqrt{\pi}} \times \frac{\sqrt{\pi}}{2h^3} = \frac{1}{2h^2} \therefore h = \frac{1}{\sigma\sqrt{2}}, \text{ donde}$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma^2}} \text{ ou } f(X) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(X-\mu)^2}{2\sigma^2}} \quad (9).$$ A fórmula (9) representa a equação da curva normal de frequência, sendo também conhecida por equação de Laplace-Gauss.

A distribuição normal é determinada por dois parâmetros: a média μ e a variância σ^2. A raiz quadrada da variância, σ, é conhecida pelo nome de desvio padrão. Pode-se obter uma infinidade de curvas normais (fig. 3), fazendo-

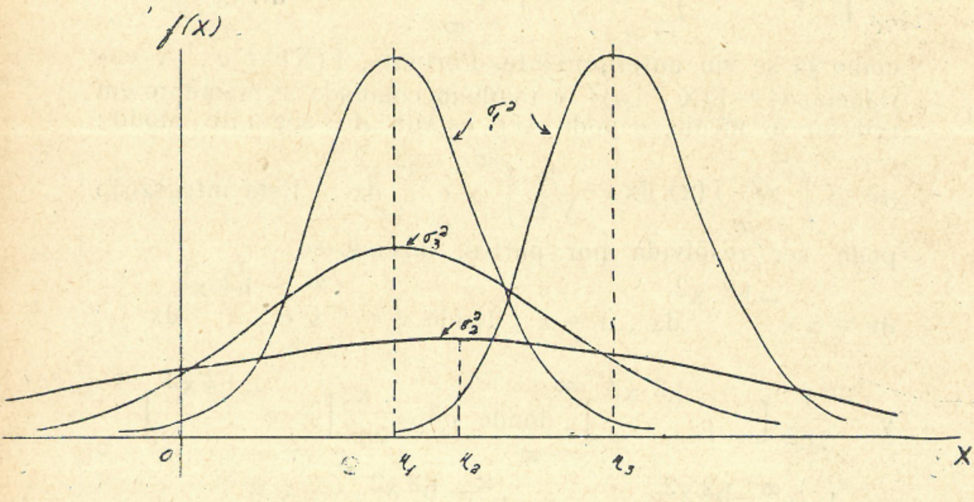

Fig. 3

se variar μ ou σ^2, ou os dois parâmetros simultâneamente. A média aritmética é uma medida da tendência central, e na curva normal ela confunde-se com a moda e a mediana.

A moda é o valor mais frequente e a mediana M é um valor tal que:

$$\int_{-\infty}^{M} f(X)dX = \frac{1}{2} = \int_{M}^{\infty} f(X)dX$$

A variação de μ desloca a curva para a direita ou para a esquerda, ao passo que σ^2 é o índice de variabilidade ou dispersão, e a sua variação determina uma curva mais ou menos extensa.

Em aplicações práticas, é conveniente reduzir-se tôdas as curvas normais a uma forma padrão, de modo que com esta curva são construídas as tabelas de ordenadas e áreas da curva normal, que são de indubitável utilidade. Isto é feito fazendo-se a variável $\frac{X-\mu}{\sigma} = y$, donde $dX = \sigma\, dy$. Isto equivale a medir as abcissas em unidades do desvio padrão (fig. 4). O cálculo da área total subtendida pela curva f (y) e o eixo das abcissas é dado pela $\int_{-\infty}^{\infty} f(y)\, dy$, e como a curva é simétrica basta efetuar-se $\int_{0}^{\infty} f(y)\, dy$. A integral $\frac{1}{\sqrt{2\pi}} \int_{0}^{y} e^{\frac{-y^2}{2}} dy$, fornece a área compreendida entre 0 e um valor qualquer y. A função f(y) é uma curva normal de $\mu = 0$ e $\sigma^2 = 1$.

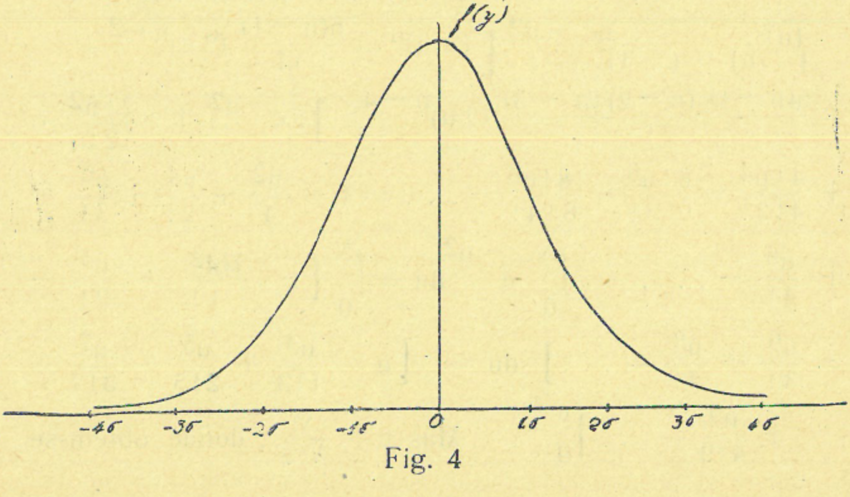

Fig. 4

Não há fórmula geral para se efetuar essa integração, mas esta dificuldade pode ser contornada desenvolvendo-se a função em uma série convergente, o que é conseguido pela série de Mac Laurin. Fazendo-se $\frac{y^2}{2} = u^2$, donde $dy = du\sqrt{2}$, a integral transforma-se em: $\frac{1}{\sqrt{\pi}}\int_0^y e^{-u^2} du$. A série de Mac Laurin nos dá:

$$f(u) = f(0) + f'(0)\frac{u}{1!} + f''(0)\frac{u^2}{2!} + f'''(0)\frac{u^3}{3!} + \ldots$$
$$\ldots + f^{(n)}(0)\frac{u^n}{n!} + \ldots$$

$$f(u) = e^{-u^2}, \quad f(0) = 1$$
$$f'(u) = -2ue^{-u^2}, \quad f'(0) = 0$$
$$f''(u) = e^{-u^2}(4u^2 - 2), \quad f''(0) = -2$$
$$f'''(u) = -e^{-u^2}(8u^3 - 12u), \quad f'''(0) = 0$$
$$f^{IV}(u) = e^{-u^2}(16u^4 - 48u^2 + 12), \quad f^{IV}(0) = 12$$
$$f^{V}(u) = -e^{-u^2}(32u^5 - 180u^3 + 120u), \quad f^{V}(0) = 0$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$f^{(n)}(u) = (-1)^n e^{-u^2}\left[(2u)^n - \frac{n(n-1)}{1!}(2u)^{n-2} + \frac{n(n-1)(n-2)(n-3)}{2!}(2u)^{n-4} - \ldots\right]; \quad e^{-u^2} = 1 - \frac{2!\,u^2}{2!} + \frac{4!\,u^4}{4!\,2!} - \frac{6!\,u^6}{6!\,3!} + \frac{8!\,u^8}{8!\,4!} - \ldots = 1 - \frac{u^2}{1} + \frac{u^4}{2!} - \frac{u^6}{3!} + \frac{u^8}{4!} - \ldots,$$

$$\int_0^y e^{-u^2} du = \int_0^y \left[1 - \frac{u^2}{1!} + \frac{u^4}{2!} - \frac{u^6}{3!} + \frac{u^8}{4!} - \ldots\right] du = \left[u - \frac{u^3}{1!\,3} + \frac{u^5}{2!\,5} - \frac{u^7}{3!\,7} + \frac{u^9}{4!\,9} - \ldots\right]_0^y$$

Mas $u = \frac{y}{\sqrt{2}}$, donde obtém-se

$$\frac{1}{\sqrt{\pi}} \int_0^y e^{\frac{-u^2}{2}} dy = \frac{1}{\sqrt{\pi}} \left[\frac{y}{\sqrt{2}} - \frac{1}{1!3} \left(\frac{y}{\sqrt{2}}\right)^3 + \frac{1}{2!5} \left(\frac{y}{\sqrt{2}}\right)^5 - \right.$$

$$\left. - \frac{1}{3!7} \left(\frac{y}{\sqrt{2}}\right)^7 + \frac{1}{4!9} \left(\frac{y}{\sqrt{2}}\right)^9 - \ldots \right]$$, que pode ser obtida com o grau de aproximação que se desejar.

ABSTRACT

This paper deals with the mathematical basis of the normal frequency distribution. The author derives the analytical expression of the normal distribution function by the method developed by Gauss to determine the error function, instead of using the binomial approximation.

The arbitrary constants that come out are determined through the conditons the function ought to fulfill. The final result is expressed in terms of the population parameters, i.e., the mean μ and the variance σ^2. The theoretical basis underlying the elaboration of the table of the area under the normal curve is explained in detail.

To make the article more accessible to persons not well acquainted with advanced calculus technique it is avoided as much as possible to skip the intermediate steps.

BIBLIOGRAFIA CITADA

(1) Kenney, J. F. — Mathematics of Statistics. D. Van Nostrand Co. Inc., New York, 1939 V. 2 Cap. II.

(2) Sokolnikoff, I. S. and E. S. — Higher Mathematics for Engineers and Physicists. Mc Graw-Hill Book Co. Inc., New York, 1941. Cap. XI.

(3) Uspensky, J. V — Introduction to Mathematical Probability. Mc Graw-Hill Book Co. Inc., New York, 1937. Cap. IX.