VOLUME XXIV NOVEMBER, 1932 PARTS III AND IV

BIOMETRIKA

FURTHER APPLICATIONS IN STATISTICS OF THE $T_m(x)$ BESSEL FUNCTION.

BY KARL PEARSON, S. A. STOUFFER AND F. N. DAVID*.

(1) THE $T_m(x)$ function was defined in a paper by Pearson, Jeffery and Elderton† to be given by

$$T_m(x) = \frac{1}{\sqrt{\pi}} \frac{1}{2^m} \frac{1}{\Gamma(m+\frac{1}{2})} x^m K_m(x) \qquad \text{(i)},$$

where $K_m(x)$ is the Bessel Function of the second order and imaginary argument. Here $T_m(x) = T_m(-x)$, while x on the right is always to be given its numerical value. Remembering this, we need not write $|x|^m K_m(|x|)$ in the equation.

If
$$y = MT_m(x) \qquad \text{(ii)}$$

be treated as a frequency curve, it will be symmetrical and run from $-\infty$ to $+\infty$ of x. The constant in (i) has been so chosen that

$$\int_{-\infty}^{+\infty} y\,dx = 2M \int_0^{\infty} T_m(x)\,dx = M.$$

An integral form of $K_m(x)$ is given by‡

$$K_m(x) = \frac{\sqrt{\pi}\, x^m}{2^m \Gamma(m+\frac{1}{2})} \int_1^{\infty} e^{-xt} (t^2-1)^{m-\frac{1}{2}}\,dt \qquad \text{(iii)}.$$

Hence we may write (ii) in the form

$$y = \frac{M}{2^{2m}} \frac{1}{\Gamma^2(m+\frac{1}{2})} x^{2m} \int_1^{\infty} e^{-xt}(t^2-1)^{m-\frac{1}{2}}\,dt \qquad \text{(iv)}.$$

(2) Consider in the next place the curve

$$y = y_0 e^{-\frac{px}{a}} \left(1 + \frac{x}{a}\right)^p \qquad \text{(v)},$$

the origin being the mode at distance a from the start of the curve.

It follows easily that

$$y_0 = \frac{M}{a} \frac{p^{p+1} e^{-p}}{\Gamma(p+1)} \qquad \text{(vi)},$$

where M is the total frequency.

* The suggestion of the problem and the selection of the illustrative examples were provided by S. A. Stouffer, the solution through the $T_m(x)$ function was given by K. Pearson, who is also responsible for the text. Florence N. David computed the table of the probability integral of the $T_m(x)$ distribution.

† *Biometrika*, Vol. XXI. p. 184.

‡ G. N. Watson: *A Treatise on the Theory of the Bessel Functions*, p. 172, Equation (4).

Thus the curve can be written

$$y = M\frac{p}{a}\frac{e^{-p\left(1+\frac{x}{a}\right)}}{\Gamma(p+1)}\left\{p\left(1+\frac{x}{a}\right)\right\}^p \quad \text{.....................(vii).}$$

Write $z = p\left(1+\frac{x}{a}\right)$ and the moments about the start of the curve can be found at once. These lead to*

$$\left.\begin{aligned} \text{Mean} = \bar{x}' &= a(p+1)/p \\ \text{Standard Deviation} = \sigma &= a\sqrt{p+1}/p \\ \beta_1 = \frac{4}{p+1}, \quad & \beta_2 = 3 + \frac{6}{p+1} \end{aligned}\right\} \quad \text{...............(viii),}$$

providing the well-known relation, $2\beta_2 - 3\beta_1 - 6 = 0$.

(3) Now suppose there are two independent variates u and v both of which have frequency distributions provided by Equation (vii). We assume the two distributions to have the same p, but to have different standard deviations σ_1 and σ_2, or, what amounts to the same thing, different modal distances a and b. We will measure our variates u and v from the start of their curves, which then take the form

$$y_1 = M\frac{p}{a}e^{-\frac{pu}{a}}\left(\frac{pu}{a}\right)^p \Big/ \Gamma(p+1),$$

and

$$y_2 = M\frac{p}{b}e^{-\frac{pv}{b}}\left(\frac{pv}{b}\right)^p \Big/ \Gamma(p+1).$$

If we take $w = M\frac{y_1}{M}\times\frac{y_2}{M}$, we obtain the combined frequency surface

$$w = M\frac{p}{a}\frac{p}{b}\frac{1}{\Gamma^2(p+1)}e^{-\left(\frac{pu}{a}+\frac{pv}{b}\right)}\left(\frac{pu}{a}\frac{pv}{b}\right)^p \quad \text{..................(ix).}$$

Now put $X = p\left(\frac{u}{a}+\frac{v}{b}\right)$ and $Y = p\left(\frac{v}{b}-\frac{u}{a}\right)$, then the element for integration of the above surface is $du\,dv$, or if we take it $d\left(\frac{pu}{a}\right)d\left(\frac{pv}{b}\right)$ we may replace it by $dX\,dY$, and we have for integration

$$\frac{M}{\Gamma^2(p+1)\,2^{2p}}e^{-X}(X^2-Y^2)^p\,dX\,dY \quad \text{.....................(x).}$$

We have to integrate this out for X to get the distribution curve of Y. In the upper octant XOB (Fig. 1, p. 295) the limit for X is clearly $X = Y$ to $X = \infty$ along the shaded area. Or, the curve of distribution of Y is

$$z = \frac{M}{\Gamma^2(p+1)\,2^{2p}}\int_Y^X e^{-X}(X^2-Y^2)^p\,dX \quad \text{..............(x } bis\text{).}$$

Put $X = Yt$ and we have

$$z = \frac{M}{\Gamma^2(p+1)\,2^{2p}}Y^{2p+1}\int_1^\infty e^{-Yt}(t^2-1)^p\,dt \quad \text{.................(xi).}$$

* *Phil. Trans.*, Vol. 185A, p. 373.

If we take the lower octant XOA, the limits of X are $-Y$ to ∞, but as Y is now negative we get precisely the same result, or we say that the whole curve of distribution of Y is (xi), Y being taken as positive, and from 0 to ∞, and mirrored in the axis of X. This result also flows from the fact that the distribution of $\frac{v}{b}-\frac{u}{a}$ must be a symmetrical curve, as the frequency curves for u/a and v/b are identical.

Now if in (iv) we write $x=Y$, $m=p+\frac{1}{2}$, we see that the z of (xi) is given by

$$z=MT_{p+\frac{1}{2}}(Y) \quad \text{.................................(xii)},$$

which leads to $\frac{1}{2}M$ for the area of our half curve. In other words our curve for Y is the $T_{p+\frac{1}{2}}$ curve mirrored on itself. The ordinates of this curve have been computed by Dr E. M. Elderton*.

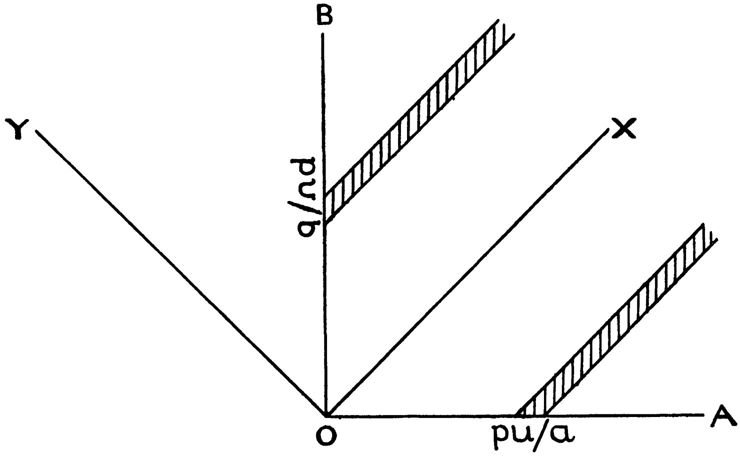

Fig. 1.

(4) Now the odd moments of the mirrored curve vanish. Let us find the even moment-coefficients. We have from (x)

$$Mu_{2s}=2\frac{M}{\Gamma^2(p+1)\,2^{2p}}\iint Y^{2s}e^{-X}(X^2-Y^2)^p\,dY\,dX,$$

where the limits of X and Y are to be chosen so as to cover the upper octant BOX. Now if we integrate first with regard to Y, the limits will be from 0 to X, and then with regard to X from 0 to ∞. Thus

$$\mu_{2s}=\frac{1}{2^{2p-1}\Gamma^2(p+1)}\int_0^\infty e^{-X}\int_0^X Y^{2s}(X^2-Y^2)^p\,dY\,dX \quad \text{.........(xiii)}.$$

Put $Y=X\lambda$ and we have

$$\mu_{2s}=\frac{1}{2^{2p-1}\Gamma^2(p+1)}\int_0^\infty e^{-X}X^{2s+2p+1}\int_0^1 \lambda^{2s}(1-\lambda^2)^p\,d\lambda\,dX,$$

* See *Biometrika*, Vol. XXI. pp. 194—201, or *Tables for Statisticians and Biometricians*, Part II. pp. lxxix—lxxxviii and 138—144.

or, if $\lambda^2 = \kappa$,

$$\mu_{2s} = \frac{1}{2^{2p}\Gamma^2(p+1)}\Gamma(2s+2p+2)\int_0^1 \kappa^{s-\frac{1}{2}}(1-\kappa)^p\,d\kappa$$

$$= \frac{1}{2^{2p}\Gamma^2(p+1)}\Gamma(2s+2p+2)\frac{\Gamma(s+\frac{1}{2})\Gamma(p+1)}{\Gamma(s+p+\frac{3}{2})}.$$

If $s=0$,
$$\mu_0 = \frac{1}{2^{2p}\Gamma(p+1)}\frac{\Gamma(2p+2)\Gamma(\frac{1}{2})}{\Gamma(p+\frac{3}{2})} = 1.$$

Hence
$$\mu_{2s} = \frac{\Gamma(2s+2p+2)}{\Gamma(2p+2)}\frac{\Gamma(p+\frac{3}{2})}{\Gamma(s+p+\frac{3}{2})}\frac{\Gamma(s+\frac{1}{2})}{\Gamma(\frac{1}{2})} \quad\text{..............(xiv)},$$

and
$$\mu_2 = \frac{(2p+3)(2p+2)}{p+\frac{3}{2}}\frac{1}{2} = 2p+2 \quad\text{......................(xv)}.$$

Generally
$$\mu_{2s} = (2s-1)(2p+2s)\,\mu_{2s-2} \quad\text{....................(xv } bis\text{)},$$

$$\beta_{2s-2} = \frac{\mu_{2s}}{(\mu_2)^s} = \frac{(2s-1)(2p+2s)}{2p+2}\frac{\mu_{2s-2}}{(\mu_2)^{s-1}},$$

or,
$$\beta_{2s-2} = (2s-1)\left(1+\frac{2(s-1)}{2(p+1)}\right)\beta_{2s-4} \quad\text{.................(xvi)}.$$

Thus finally

$$\beta_{2s-2} = (2s-1)(2s-3)\ldots 1\left(1+\frac{s-1}{p+1}\right)\left(1+\frac{s-2}{p+1}\right)\ldots\left(1+\frac{1}{p+1}\right)\ldots\text{(xvii)}.$$

It will be clear that when $p \to \infty$ we obtain

$$\beta_{2s-2} = (2s-1)(2s-3)\ldots 1,$$

the familiar β_{2s-2} formula for the normal curve, into which the $T_{p+\frac{1}{2}}$ function then passes.

Consider the Type VII curve

$$y = y_0\frac{1}{(a^2+x^2)^{\frac{1}{2}n}}.$$

Here we have

$$\beta_{2s-2} = (2s-1)(2s-3)\ldots 1\left(1+\frac{2(s-1)}{n-2s-1}\right)\left(1+\frac{2(s-2)}{n-2s+1}\right)\ldots\left(1+\frac{2}{n-5}\right)$$

and $\mu_2 = a^2/(n-3)$.

Now it is clear that we can make μ_2 and μ_4 agree in the Type VII and the $T_{p+\frac{1}{2}}$ curves*, but farther than that we cannot go, although the β_1's may not differ widely if n be considerable. The $T_{p+\frac{1}{2}}$ curve has the further advantage that no moment-coefficients tend to become infinite, while if n be an odd integer, those for the Type VII curve may become so. For values of p not too great the Type VII will fit the distribution of Y considerably better than the normal curve. For considerable values of p, both Type VII and the $T_{p+\frac{1}{2}}$ curves pass into the normal curve.

(5) A few further points may be noted. If $p=-\frac{1}{2}$ the T_0-curve asymptotes to the vertical at the origin, and this holds as long as p lies between $-\frac{1}{2}$ and 0; if

* We must take $\dfrac{1}{p+1} = \dfrac{2}{n-5}$ or $n = 2p+7$, and $a = 2\sqrt{(p+1)(p+2)}$.

$p=0$, the $T_{\frac{1}{2}}$-curve starts with a finite ordinate and makes a finite angle with the vertical, it is the exponential curve. If p be positive we see from (x *bis*) that $dz/dY=0$ for $Y=0$, or the double mirror curves have a common tangent at the axis of symmetry and will *in appearance* form a single curve. If p be a positive integer it is possible to expand z in powers of Y, but the series does not present any great advantages to the computer.

When $p=11$, Dr Elderton's Tables terminate, but it is shown in the memoir by Pearson, Jeffery and Elderton* that when $p=11$, the two curves

$$z=MT_{p+\frac{1}{2}}(Y)$$

and

$$z=\frac{M}{\sqrt{2\pi(p+1)(p+2)}}\frac{\Gamma\{\frac{1}{2}(2p+7)\}}{\Gamma(p+3)}\frac{1}{\left(1+\frac{Y^2}{4(p+1)(p+2)}\right)^{\frac{1}{2}(2p+7)}} \quad \text{......(xviii)}$$

coincide for practical statistical purposes. The areas of this latter curve up to given values of Y have been tabled† from $p=-\frac{1}{2}$ to $p=12$, but this hardly carries us beyond the T_m-tables. The completed (and now at press) *Tables of the Incomplete* B-*function* carry us up to $2p+7=101$, or $p=47$.

(6) Now let us turn to the means of samples of size n drawn from the Type III curve

$$y=y_0'\,e^{-\frac{px}{a}}\left(\frac{x}{a}\right)^p \quad \text{..............................(xix)},$$

where the origin is at the start of the curve and a is the distance to the mode from the start. Let us suppose a sample $x_1, x_2, x_3 \ldots x_n$ drawn and let its mean be $\bar{x}_n=(x_1+x_2+\ldots+x_n)/n$. Then the chance P of a sample lying between x_1 and $x_1+\delta x_1$, x_2 and $x_2+\delta x_2$, ... x_n and $x_n+\delta x_n$ is given by

$$P=\text{const.}\times e^{-\frac{p}{a}(x_1+x_2+\ldots+x_n)}\left(\frac{x_1x_2\ldots x_n}{a^n}\right)^p dx_1dx_2\ldots dx_n.$$

Now get rid of x_1 by introducing $\bar{x}_n$ as a variable and write l_2 for

$$n\bar{x}_n-x_3-x_4-\ldots-x_n.$$

We have

$$P=\text{const.}\times e^{-\frac{np\bar{x}_n}{a}}\,d\bar{x}_n\left(\frac{l_2-x_2}{a}\right)^p\left(\frac{x_2}{a}\right)^p\left(\frac{x_3\ldots x_n}{a^{n-2}}\right)^p dx_2dx_3\ldots dx_n.$$

Put $x_2=l_2x_2'$ and integrate out for $x_2=0$ to l_2 or $x_2'=0$ to 1. This will introduce a B-function into the constant, but leave us with

* Cf. *Biometrika*, Vol. XXI. pp. 171 and 173 for accordance of the curves. Their equations are given on p. 185, where we must write $\frac{1}{2}n-1=p+\frac{1}{2}$, or $n=2p+3$. The two curves have then the same first four moment-coefficients. If $\eta=Y/\{2\sqrt{(p+1)(p+2)}\}$, then the proportional area from $\eta=0$ up to any arbitrary value of η is given by $\frac{1}{2}I_\eta(\frac{1}{2}, p+1)$, where $I_\eta(\frac{1}{2}, p+1)=B_\eta(\frac{1}{2}, p+1)/B(\frac{1}{2}, p+1)$, B_η and B being the incomplete and complete Beta-functions.

† See *Biometrika*, Vol. XXII. pp. 253—283, or *Tables for Statisticians and Biometricians*, Part II, pp. cxxv—cxlii and pp. 169—177.

$$P = \text{const.} \times e^{-\frac{np\bar{x}_n}{a}} d\bar{x}_n \left(\frac{l_2}{a}\right)^{2p+1} \left(\frac{x_3 \dots x_n}{a^{n-2}}\right)^p dx_3 \dots dx_n.$$

Write $l_2 = l_3 - x_3$, and proceeding in the same way, we find

$$P = \text{const.} \times e^{-\frac{np\bar{x}_n}{a}} d\bar{x}_n \left(\frac{l_3}{a}\right)^{3p+2} \left(\frac{x_4 \dots x_n}{a^{n-3}}\right)^p dx_4 \dots dx_n,$$

where $l_3 = n\bar{x}_n - x_4 - x_5 - \dots - x_n$.

Continuing to repeat this process we ultimately get rid of all the variables but $\bar{x}_n$ and find*

$$P = \text{const.} \times e^{-\frac{np\bar{x}_n}{a}} \left(\frac{\bar{x}_n}{a}\right)^{n(p+1)-1} d\bar{x}_n \qquad \text{(xx)}.$$

We now put this into the canonical form for a Type III frequency curve, i.e.

$$y = y_0 e^{-\frac{P}{A}\bar{x}_n} \left(\frac{\bar{x}_n}{A}\right)^P \qquad \text{(xx } bis\text{)}.$$

Hence we must have $P = n(p+1) - 1$, and $P/A = np/a$, or $A = a\dfrac{n(p+1)-1}{np}$.

Accordingly:

$$\left.\begin{aligned} &\text{Mode of } \bar{x}_n = a\,\frac{n(p+1)-1}{np} \\ &\text{Mean of } \bar{x}_n = M_1' = \frac{A(P+1)}{P} = \frac{a(p+1)}{p} = \bar{x} \\ &\sigma_{\bar{x}_n}{}^2 = M_2' - \frac{A^2(P+1)}{P^2} = \frac{1}{n}\,\frac{a^2(p+1)}{p^2} = \frac{1}{n}\,\sigma_x{}^2 \end{aligned}\right\} \qquad \text{(xxi)},$$

where $\bar{x}$ and σ_x are the mean and standard deviation of the population from which the sample of n is drawn. Lastly

$$B_1 = \frac{4}{n(p+1)} \quad \text{and} \quad B_2 = 3 + \tfrac{3}{2}B_1 \qquad \text{(xxii)}.$$

Clearly, if n and p are not very small, then (xx *bis*) will approach much nearer to a normal distribution than the parent population (xix).

(7) We can now apply our results to particular cases. If we draw two individuals out of Type III curves like (xix), with the same skewness as measured by p, then if a and a' be their modal distances, and

$$Y = p\left(\frac{x_2}{a_2} - \frac{x_1}{a_1}\right) = (p+1)\left(\frac{x_2}{\bar{x}_2} - \frac{x_1}{\bar{x}_1}\right) = \sqrt{(p+1)}\left(\frac{x_2}{\sigma_{x_2}} - \frac{x_1}{\sigma_{x_1}}\right),$$

for these are all equivalent, then the distribution of Y is given by

$$z = MT_{p+\frac{1}{2}}(Y).$$

If the two individuals are taken from absolutely the same population, i.e. $a_2 = a_1 = a$, then

$$Y = p\,\frac{x_2 - x_1}{a} = (p+1)\,\frac{x_2 - x_1}{\bar{x}} = \sqrt{(p+1)}\,\frac{x_2 - x_1}{\sigma_x}.$$

* This result was published by Church: see *Biometrika*, Vol. XVIII. p. 336.

Such results, however interesting in the case of experimental sampling in the Laboratory, where we have a knowledge of the parent population, will hardly be of practical service, because we should usually lack a knowledge of p, $\bar{x}$ and σ_x.

Now turn to (xx), and suppose we have taken two samples of n and that their means are $\bar{x}_n$ and $\bar{x}_n'$, then the distribution of $Y = \frac{P}{A}(\bar{x}_n' - \bar{x}_n)$ will be

$$z = \tfrac{1}{2} MT_{P+\frac{1}{2}}(Y) = \tfrac{1}{2} MT_{n(p+1)-\frac{1}{2}}(Y) \quad \text{.................(xxiii).}$$

There are now a variety of ways in which it is possible to express Y. In the first place $P/A = \frac{np}{a}$, where p and a refer to the parent population, but mean $-$ mode $= \frac{a}{p} = \bar{x} - \tilde{x}$, say. Again $\frac{p}{a} = \frac{\bar{x}}{\sigma_x^2} = \frac{2}{\sqrt{\beta_1}\,\sigma_x}$. Thus we have

$$Y = n\frac{\bar{x}_n' - \bar{x}_n}{\bar{x} - \tilde{x}} = \frac{n\bar{x}(\bar{x}_n' - \bar{x}_n)}{\sigma_x^2} = \frac{2n}{\sqrt{\beta_1}}\frac{(\bar{x}_n' - \bar{x}_n)}{\sigma_x} \quad \text{...........(xxiv).}$$

Further, we need the value of the $p+1$ in the degree of the T_m function; we have

$$p + 1 = \frac{\bar{x}}{\bar{x} - \tilde{x}} = \frac{\sigma_x^2}{(\bar{x} - \tilde{x})^2} = \frac{4}{\beta_1} \quad \text{.......................(xxv).}$$

Here $\bar{x}$, $\tilde{x}$, σ_x and β_1 all refer like p to the parent population. Clearly some two of these quantities $\bar{x}$ and $\tilde{x}$, $\bar{x}$ and σ_x, or β_1 and σ_x must be known, or we cannot determine a and p. We shall see later that in certain other applications p is known, and then probably σ_x is the best quantity to seek for. It might be thought that $\bar{x}$ would be easy to find. It may be so, if the start of the curve can be determined, but it must be remembered that $\bar{x}$ is the mean measured from a definite point of the parent population, i.e. the start of the parent population, and this may be quite unknown, $\bar{x} - \tilde{x}$ does not involve this knowledge, but the mode is not an easily determined character. On the whole β_1 and σ_x can probably be most easily obtained from the samples. Of course this refers to cases in which the parent population is unknown, but suspected of having a skewness which may be approximated to by a Type III curve. The procedure here would be to determine to the second and third moment coefficients of the pooled samples, and thus obtain the best approximation which is available to β_1 and σ_x of the supposed parent population.

We then take $m = \frac{4n}{\beta_1} - \frac{1}{2}$, and

$$Y = \frac{2n}{\sqrt{\beta_1}}\frac{\bar{x}_n' - \bar{x}_n}{\sigma_x} \quad \text{.........................(xxvi),}$$

and test whether the probability integral of $T_m(Y)$ has a value sufficiently large to justify us in assuming that $\bar{x}_n'$ and $\bar{x}_n$ came from the same population.

Perhaps a more useful case occurs when one sample is sufficiently large to give reasonable values for the constants, and we ask whether the other could have been drawn from the same population. In this case we may determine p and a with sufficient accuracy from the large sample and measure the probability of x_n for the second sample from (xx) or (xx *bis*) by aid of the *Tables of the Incomplete* Γ*-function.*