Fragments of a New Test Formula of Normality

By Chikio HAYASHI

(Received June 3, 1948)

1. Introduction

When we want to determine the type of distribution function from the samples which were drawn from an universe, we usually presuppose a distribution function and test the goodness of fit by using χ^2- or ω^2-test. In these cases, if sample size is very large, for example 10^4 or 10^5, we are frequently obliged to reject the hypothesis. In many cases where I have encountered in social or economical phenomena or in text books and papers, χ^2- or ω^2-test tells me to reject the hypothesis whenever sample size is large. This is obviously due to the fact that sample size is large and χ^2 (or ω^2) value increases usually linearly with sample size: (this shows that sample distribution does not approach so much to mathematical formula because of complicated fluctuations in reality though sample size is large):—sample size in the case where the $\chi^2(\omega^2)$-test is applied, is moderate, for example $100 \sim 1000$. Sociometrically dealing with social phenomena which are quite different from purely experimental phenomena and fluctuate extremely on account of complicated activities of many factors, this is not appropriate in themselves. We must consider a new test, something like a risk function investigated by Wald in his paper "Statistical Inference."

On the contrary, being unable to use $\chi^2(\omega^2)$-test if sample size is very small, so we must consider another test. In some cases, we must also deal with these problems even in analysing sociometrical problems.

In this note, I will consider on a test of normality. Well known formula and tables of test of normality are β_1, β_2-test given by E. S. Pearson and ω_n-test by R. C. Geary. β_1 is a criterion to detect skewness and β_2, ω_n are criteria to detect whether the population sample is platykurtic or leptokurtic. Now I will calculate the distribution of the ratio of the median to the square root of unbiased estimate of variance which are estimated from samples and establish a criterion to detect a kind of skewness.

2. Test formulae used for tabulation

In the 1st place, the distribution function mentioned above of N samples which are independently drawn from the normal population specified by

the function $\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{x^2}{2\sigma^2}}$ will be calculated.

Let the order of these N samples is randomized, and let these variables be $X_1, \cdots\cdots, X_N$. Then we divide $X_1, \cdots\cdots, X_N$ into 2 groups, so that

the 1st group consist of $X_1, \cdots\cdots, X_{2n+1}$,

the 2nd group consist of $X_{2n+2}, \cdots\cdots, X_N$,

which we denote newly by $Y_1, \cdots\cdots, Y_{f+1}$, where

$$(2n+1)+(f+1)=N.$$

From the 1st group, median $\tilde{X}$ is estimated and unbiassed estimate S^2 of variance is obtained from the 2nd group.

$\tilde{X}$ and S are mutually independent. Accordingly, their simultaneous distribution is given in the form of density function as follows:

$$p(\tilde{x}, s)=\left(\frac{f^{\frac{f}{2}}}{2^{\frac{f-2}{2}}\Gamma\left(\frac{f}{2}\right)\sqrt{2\pi}}\cdot\frac{(2n+1)!}{(n!)^2}\right)\left(\frac{1}{\sigma^f}\cdot\frac{1}{\sigma^{2n+1}}\right)S^{f-1}e^{-\frac{fS^2}{2\sigma^2}}$$

$$\times\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{y^2}{2\sigma^2}}dy\right)^n\left(\frac{1}{\sqrt{2\pi}}\int_{\tilde{x}}^{\infty}e^{-\frac{y^2}{2\sigma^2}}dy\right)^n e^{-\frac{\tilde{x}^2}{2\sigma^2}}.$$

Changing the variables $\frac{\tilde{x}}{S}=t$, $S=S$ and integrating by S over the whole, we obtain the density function $P(t)$ of t

$$(1)\qquad p(t)=A\cdot\int_0^\infty u^f e^{-\frac{(f+t)^2}{2}u^2}\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{ut}e^{-\frac{z^2}{2}}dz\right)^n\left(\frac{1}{\sqrt{2\pi}}\int_{ut}^{\infty}e^{-\frac{z^2}{2}}dz\right)^n du$$

where

$$A=\frac{f^{\frac{f}{2}}}{2^{\frac{f-2}{2}}\Gamma\left(\frac{f}{2}\right)\sqrt{2\pi}}\cdot\frac{(2n+1)!}{(n!)^2}.$$

From this an approximate formula suitable for calculation to tabulate is induced.

(i) The case where sample size n is large.

In this case, let the density function of t be $q(t)$.

Using the relation

$$\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{ut}e^{-\frac{z^2}{2}}dz\right)^n\left(\frac{1}{\sqrt{2\pi}}\int_{ut}^{\infty}e^{-\frac{z^2}{2}}dz\right)^n=\left(\frac{1}{4}\left(\frac{1}{\sqrt{2\pi}}\int_0^{ut}e^{-\frac{z^2}{2}}dz\right)^2\right)^n$$

and Stirling's formula in (1), easy calculation gives

$$(2)\qquad q(t)=\sqrt{\frac{n}{f}}\,\frac{2}{\pi}\cdot\frac{\Gamma\left(\frac{f+1}{2}\right)}{\Gamma\left(\frac{f}{2}\right)}\left(1+\left(\frac{4n}{\pi}\cdot\frac{1}{f}\right)t^2\right)^{-\frac{f+1}{2}}\left(1+0\left(\frac{1}{n}\right)\right)$$

(Wilks' calculation of distribution of median in the case, when n is large, is used here.)

(ii) The other case.

In this case, calculation is complicated. Modifying (1) and using the precise approximate formula of $\left(\frac{1}{\sqrt{2\pi}}\int_0^{ut}e^{-\frac{z^2}{2}}dz\right)^n$ obtained by simple consideration (Williams treats this in his paper for $n=1,2$), i. e. for even $n\geqq 4$[1]

$$p_n(x)\underset{(\leqq)}{\doteq}\frac{1}{2^{\frac{n-2}{2}}\Gamma\left(\frac{n}{2}\right)}\int_0^{\left[\frac{2r}{\sqrt{\pi}}\right]\left\{\frac{n}{2}\Gamma\left(\frac{n}{2}\right)\right\}^{\frac{1}{n}}}r^{n-1}e^{-\frac{r^2}{2}}dr$$

$$=\frac{1}{2^{\frac{n-2}{2}}\Gamma\left(\frac{n}{2}\right)}\left\{(n-2)\cdots 2\right\}-e^{-\frac{2x^2}{\pi}\left\{\frac{n}{2}\cdot\Gamma\left(\frac{n}{2}\right)\right\}^{\frac{2}{n}}}$$

$$\cdot\sum_{j=0}^{\frac{n-2}{2}}{}'\left[\frac{2x}{\sqrt{\pi}}\right]^{n-2-2j}\frac{(n-2)\cdots(n-2j)}{2^{\frac{n-2}{2}}\Gamma\left(\frac{n}{2}\right)}\cdot\left\{\frac{n}{2}\Gamma\left(\frac{n}{2}\right)\right\}^{\frac{n-2-2j}{n}}$$

where

$$P_n(x)=\left[\frac{1}{\sqrt{2\pi}}\int_{-x}^{x}e^{-\frac{1}{2}t^2}dt\right]^n,$$

$$\sum_{j=0}^{\frac{n-2}{2}}{}'\left[\frac{2x}{\sqrt{\pi}}\right]^{n-2-2j}\frac{(n-2)\cdots(n-2j)}{2^{\frac{n-2}{2}}\cdot P\left(\frac{n}{2}\right)}\cdot\left\{\frac{n}{2}P\left(\frac{n}{2}\right)\right\}^{\frac{n-2-2j}{n}}\cdot$$

$$=\frac{\left[\left[\frac{2x}{\sqrt{\pi}}\right]\left\{\frac{n}{2}P\left(\frac{n}{2}\right)\right\}^{\frac{1}{n}}\right]^{n-2}}{2^{\frac{n-2}{2}}\cdot 1^{-\left(\frac{n}{2}\right)}}$$

$$+\sum_{j=1}^{\frac{n-2}{2}}\left[\frac{2x}{\sqrt{\pi}}\right]^{n-2-2j}\frac{(n-2)\cdots(n-2j)}{2^{\frac{n-2}{2}}\cdot P\left(\frac{n}{2}\right)}\left\{\frac{n}{2}P\left(\frac{n}{2}\right)\right\}\frac{n-2-2j}{n},$$

and substituting the following relation in (1)

$$\left\{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{uf} e^{-\frac{z^2}{2}}dz\right\}^n\left\{\frac{1}{\sqrt{2\pi}}\int_{uf}^{\infty} e^{-\frac{z^2}{2}}dz\right\}^n$$

$$=\left\{\frac{1}{4}-\left(\frac{1}{\sqrt{2\pi}}\int_0^{uf} e^{-\frac{z^2}{2}}dz\right)^n\right\}$$

$$=\sum_{r=0}^{n}(-1)^r\binom{n}{r}\left(\frac{1}{4}\right)^{n-r}\left\{\frac{1}{\sqrt{2\pi}}\int_0^{uf} e^{-\frac{z^2}{2}}dz\right)^2\Big\}^r$$

---

(1) Obviously we have

$$P_n(x)=\left[\frac{1}{\sqrt{2\pi}}\int_{-x}^{x} e^{-\frac{1}{2}t^2}dt\right]^n$$

$$=\left(\frac{1}{\sqrt{2\pi}}\int_{-x}^{x} e^{-\frac{1}{2}t_1^2}dt_1\right)\left(\frac{1}{\sqrt{2\pi}}\int_{-x}^{x} e^{-\frac{1}{2}t_2^2}dt_2\right)\cdots\left(\frac{1}{\sqrt{2\pi}}\int_{-x}^{x} e^{-\frac{1}{2}t_n^2}dt_n\right)$$

$$=\left(\frac{1}{\sqrt{2\pi}}\right)^n\int_{-x}^{x}\int_{-x}^{x}\cdots\int_{-x}^{x} e^{-\frac{1}{2}(t_1^2+\cdots+t_n^2)}dt_1\cdots dt_n.$$

Changing the variables $t_1\cdots t_n$ into the variables of polar, changing the integral domains (cube)—$x\leqq t,\ \leqq x,\ -x\leqq t_n\leqq x$ into the sphere which has the same volume as the cube and integrating, then the desired approximate inequality is obtained owing to the property of normality—symmetry and the higher the density, the nearer the variate appraches to the centre—.

And so the formula mentioned above is larger than

$$\frac{1}{2^{\frac{n-2}{2}}P\left(\frac{n}{2}\right)}\{(n-2)\cdots 2\}-e^{-x^2}\sum_{j=0}^{\frac{n-2}{2}}{}'\frac{(n-2)\cdots(n-2j)}{2^{\frac{n-2}{2}}P\left(\frac{n}{2}\right)}\cdot x^{n-2-2j}$$

and smaller than

$$\frac{1}{2^{\frac{n-2}{2}}P\left(\frac{n}{2}\right)}\{(n-2)\cdots 2\}-e^{-2x^2}\sum_{j=0}^{\frac{n-2}{n}}\frac{(n-2)\cdots(n-2j)}{2^{\frac{n-2}{2}}P\left(\frac{n}{2}\right)}(\sqrt{2})^{n-2-2j}x^{n-2-2j}$$

(ii) for $n=2$ $\qquad P_2(x)\underset{(\leqq)}{\fallingdotseq}[1-e^{-\frac{2x^2}{\pi}}]$

(iii)
$$P_n(x)\underset{(\leqq)}{\fallingdotseq}\frac{1}{2^{\frac{n-2}{2}}P\left(\frac{n}{2}\right)}\Bigg[\sqrt{\frac{\pi}{2}}(n-2)\cdots 3\cdot 1\left(1-e^{-\frac{8x^2}{\pi^2}\left\{\frac{n}{2}P\left(\frac{n}{2}\right)\right\}^{\frac{2}{n}}}\right)$$

$$-e^{-\frac{2x^2}{\pi}\left\{\frac{n}{2}P\left(\frac{n}{2}\right)\right\}^{\frac{2}{n}}}\cdot\sum_{j=0}^{\frac{n-3}{2}}{}'\left(\frac{2x}{\sqrt{\pi}}\right)^{n-2-2j}\cdot(n-2)\cdots(n-2j)$$

$$\cdot\left\{\frac{n}{2}P\left(\frac{n}{2}\right)\right\}^{\frac{n-2-2j}{n}}\Bigg]$$

for odd $n\gg 3$.

$$= \sum_{r=0}^{n}{}'' (-1)^r \binom{n}{r}\left(\frac{1}{4}\right)^n \left\{ \frac{1}{2^{r-1}}(2r-2)\cdots 2 - e^{-\frac{2u^2f^2}{\pi}\{r\cdot P(r)\}^{\frac{1}{r}}} \right.$$

$$\left. \times \sum_{j=0}^{r-1}{}' \left[\frac{2uf}{\sqrt{\pi}}\right]^{2r-2-2j} \frac{(2r-2)\cdots(2r-2j)}{2^{r-1}\cdot P(r)} \left\{ r - P(r) \right\}^{\frac{r-1-j}{r}} \right.$$

where

$$\sum_{r=0}^{n}{}'' = \sum_{r=2}^{n} + \left(\frac{1}{4}\right)^n - n\left(\frac{1}{4}\right)^n \cdot \left[1 - e^{-\frac{2x^2}{\pi}}\right]$$

and $\sum'$ indicates the above definition for $j=0$, then we have in the long run

$$\begin{aligned} P(t) &= A \sum_{r=0}^{n}{}'' (-1)^r \binom{n}{r}\left(\frac{1}{4}\right)^n \frac{1}{2^{r-1}P(r)} \\ &\times \left[(2r-2)\cdots 2\cdot 2^{\frac{f-1}{2}} P\left(\frac{f+1}{2}\right)(f+t^2)^{-\frac{f+1}{2}} \right. \\ &- \sum_{j-0}^{r-1}{}' \left(\frac{4}{\pi}\right)^{r-j-1} (2r-2)\cdots(2r-2j) \\ &\times \left\{ rP(r) \right\}^{\frac{r-j-1}{r}} t^{2(r-j-1)} 2^{\frac{f+2(r-j-1)-1}{2}} \\ &\left. \times P\left(\frac{f+2(r-j)-1}{2}\right)\left(f + t^2\left[1 + \frac{4}{\pi}\left\{rP(r)\right\}^{\frac{1}{r}}\right]\right)^{-\frac{f+2(r-j)-1}{2}} \right]. \end{aligned} \tag{3}$$

This formula is the combination of so-called t-distribution functions. So the mean and variance of (2), (3) are easily calculated from the formulae of t-distribution function. Thus the tables (lower and upper limits) of 1% and 5% points of $p(t)$ are obtained by combining so-called t-distribution functions tables with other common mathematical tables. These numerical values (Tables), the comparison tables with the ω_n (geary) and β_1, β_2, and the ways of applying our test mentioned above to actual problems will be published later on.

---

(2) Now for simplicity we shall consider rough approximate formula to check. This formula is as follows:

$$P(t) = A \sum_{r=0}^{n} \sum_{i=0}^{r} G_i(r,n)\, 2^{\frac{t-1}{2}} P\left(\frac{t+1}{2}\right)\left\{1 + t^2\left(\frac{1+2ci}{t}\right)\right\}^{-\frac{t+1}{2}} \cdot t^{-\frac{t+1}{2}}$$

where A is previously shown, $c = 0{,}6302$ and

$$Gi(r,n) = (-1)^{r+i}\binom{n}{r}\left(\frac{1}{4}\right)^n \binom{r}{i}.$$

3. The best way of dividing N samples

We consider the case where both n and N are large. Then because of the symmetry of $q(t)$, obviously $E(t)$—mean value of t—is zero. Accordingly, the variance is $E(t^2)$: If we divide the samples N in order to minimize the variance $E(t^2)$, we shall obtain the best result (from the point of view of testing of statistical hypothesis). Now

$$E(f^2) = \int_{-\infty}^{\infty} t^2 q(t)\, dt \fallingdotseq \frac{\pi}{4n} \frac{f}{f-2}; \quad f \geqq 3 \tag{4}$$

being obtained (we use $q(t)$ as the density function when n is large), it follows from the minimum condition of $E(t^2)$ with respect to f,

$$f = \sqrt{2(N-2)}. \tag{5}$$

From $N = 2n + 1 + (f+1)$, it is known that f must be determined as the nearest integer to $\sqrt{2(N-2)}$ such that $(N-2-f)$ is even. Thus the minimum variance is $\dfrac{\pi f_e}{(4n_e)(f_e-2)}$ where f_e, n_e satisfy the above conditions.