CERTAIN DISTRIBUTION-FREE TESTS OF REGRESSION*

By R. S. Kurup ‡

University of Kerala, Trivandrum

1. The Problem

Suppose we are given n pairs of observations (x_i, y_i), $i = 1, 2 \cdots n$ from a continuous bivariate distribution and we are required to fit a relation of the form $Y = f(x, \theta)$ where 'θ' denotes a set of parameters whose values may be found by any method of estimation. To test the significance of regression, the null hypothesis is $H_0 : \theta = 0$. Classical workers tested regression by assuming that the errors are normally and independently distributed and this forms the basis of the x^2-test. In this paper the problem is tackled without any such assumptions.

For this problem, Brown and Mood (1950),[1] (1951)[2] suggested a statistic,

$$A = \frac{8}{n}\left\{\left(r_1 - \frac{n}{4}\right)^2 + \left(r_2 - \frac{n}{4}\right)^2\right\}$$

where r_1 and r_2 are the number of positive ϵ's below and above the median of the x's, ϵ being the discrepancy between the observed 'y' and the value of 'y' under the null hypothesis. For moderately large 'n', this is distributed as a 'x^2' with 2 degrees of freedom. This statistic considers the 4 possible arrangements of signs, as shown below:

$n/2$						$n/2$					
+	+	+	+	⋯	+	−	−	−	−	⋯	−
+	+	+	+	⋯	+	+	+	+	+	⋯	+
−	−	−	−	⋯	−	−	−	−	−	⋯	−
−	−	−	−	⋯	−	+	+	+	+	⋯	+

* Paper prepared in 1955 under a Government of India Senior Scholarship.

‡ The author is at present employed in the Bereau of Economics and Statistics, Trivandrum.

Daniels in 1954[3] suggested:

$$B = \frac{2}{\sqrt{n}} \left\{ \left| r_1 - \frac{n}{4} \right| + \left| r_2 - \frac{n}{4} \right| \right\}$$

as a test criterion. This has the asymptotic distribution:

$$P(B \geqslant B_0) = 4\,\phi(B_0)\,(1 - \phi(B_0))$$

where $\phi(B_0)$ is the normal cumulative distribution. It was shown that this is more powerful than the A-test mentioned earlier. Daniels proposed another test, the m-test, based on the 2^n possible arrangements of signs.

2. The Proposed Test-Criteria and Their Distributions

Let $x_1, x_2 \cdots x_n$ denote the ordered x's in ascending order of magnitude; also let ϵ be the difference between the observed 'y' and the value of 'y' under the null hypothesis and 'R_i' denote the number of positive ϵ's up to x_i (including 'x_i'). Considering only four possible arrangements of the signs we may formulate the criterion R_n with expectation $n/2$ and variance $n/4$. It is obvious that for large n, $4/n\,(R_n - n/2)$ follows the normal distribution with mean '0' and variance unity under 'H_0'. For small 'n', the probability of any particular 'R_n' can be found easily by computation as the probability is $\frac{1}{2}$ for any ϵ to be positive or negative.

Another criterion which can be used for the test is $T = \sum_{i=1}^{n} R_i$.

The exact distribution of this criterion has been tabulated up to '$n = 10$' by considering all possible arrangements. As 'T' is found to be symmetrical about its mean $n(n+1)/4$ the upper half alone is given in Table I.

The significance of 'T' should be tested using the two tails of the distribution.

The distribution of T is symmetrical. The mean value of 'T' is $n(n+1)/4$ and variance $[n\,(n+1)\,(2n+1)]/24$

$$T' = \frac{T - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}$$

TABLE I

Ordinates P of the distribution of $T = \sum_{i=1}^{n} R_i$

n	*T*	*P*
3	6	·125
	5	·125
	4	·125
	3	·250
4	10	·0625
	9	,,
	8	,,
	7	·1250
	6	,,
	5	,,
5	15	·03125
	14	,,
	13	,,
	12	·0625
	11	,,
	10	·09375
	9	,,
	8	,,
6	21	·015625
	20	,,
	19	,,
	18	·03125
	17	,,
	16	·046875
	15	·0625
	14	,,
	13	,,
	12	·078125
	11	,,
7	28	·0078125
	27	,,
	26	,,
	25	·015625
	24	,,
	23	·0234375
	22	·03125
	21	·0390625
	20	,,
	19	·046875
	18	·0546875
	17	,,
	16	·0625
	15	,,
	14	,,
8	36	·00390625
	35	,,
	34	,,
	33	·0078125
	32	,,
	31	·01171875
	30	·015625
	29	·01953125
	28	·0234375
	27	·02734375
	26	·03125
	25	·03515625
	24	.0390625
	23	·04296875
	22	·046875
	21	·05078125
	20	,,
	19	,,
	18	·0546875
9	45	·001953125
	44	,,
	43	,,
	42	·00390625
	41	,,
	40	·005859375
	39	·0078125
	38	·009765625
	37	·01171875

TABLE I (*Contd.*)

n	T	P
	36	·015625
	35	·017578125
	34	·01953125
	33	·0234375
	32	·025390625
	31	·029296875
	30	·033203125
	29	·03515625
	28	·037109375
	27	·041015625
	26	,,
	25	·04296875
	24	·044921875
	23	,,
10	55	·0009765625
	54	,,
	53	,,
	52	·001953125
	51	,,
	50	·0029296875
	49	·00390625
	48	·0048828125
	47	·005859375
	46	·0078125
	45	·009765625
	44	·0107421875
	43	·0126953125
	42	·0146484375
	41	·0166015625
	40	·01953125
	39	·021484375
	38	·0234375
	37	·0263671875
	36	·0283203125
	35	·0302734375
	34	·0322265625
	33	·0341796875
	32	·03515625
	31	·037109375
	30	·0380859375
	29	,,
	28	·0390625

(,,) Denotes the same value as before.

will therefore tend to have a normal distribution with mean '0' and variance 1 for large 'n'.

3. ANOTHER TEST CRITERION

If

$$z_i = R_i - \frac{i}{2},$$

$$E(z_i) = 0$$

and

$$E(z_i z_j) = \frac{n_j}{4} \quad \text{for} \quad j \leqslant i$$

$$= \frac{n_i}{4} \quad \text{for } j > i.$$

Denoting by z, the matrix of values $z_1, z_2, \cdots z_n$ and by Γ the covariance matiix of the z_i's

$$T_n^2 = z\Gamma^{-1}z'$$

$$= \frac{4}{n}\left[2\sum_{i=1}^{n-1}(z_i^2 - z_i z_{i+1}) + z_n^2\right].$$

$$= \frac{4}{n}\sum_{i=1}^{n}\left(\delta_i - \frac{1}{2}\right)^2$$

where δ_i is 1 or 0 with probability $\frac{1}{2}$. Here, it may be noted that R_i (defined earlier) $= \sum_{j=1}^{i} \delta_j$.

From the structure of the criterion $T_n^2 = z\Gamma^{-1}z'$, it can be seen that for large n, it will behave as a x^2 with 'n' degrees of freedom. For small 'n' the exact distribution under H_0 can be tabulated.

4. Power of the Criteria

We may for convenience assume that we are required to test the linear relation $y = \alpha + \beta x$. Under the alternative (β, α) against (β_0, α_0), $\epsilon_i = y_i - \alpha - \beta x_i \cdots$ (1)

has probability $\frac{1}{2}$ for being positive or negative

$$y_i - \alpha_0 - \beta_0 x_i = \epsilon_i - (\alpha_0 - \alpha) - (\beta_0 - \beta)\, x_i. \quad (2)$$

Let

$$p_i = \text{prob. } \{\epsilon_i > (\alpha_0 - \alpha) + (\beta_0 - \beta)\, x_i\}. \quad (3)$$

Following Daniels[3] we may consider $p_i = p$ for all 'i'. This is the case if the alternatives are (β_0, α) and all the ϵ's have the same distribution $f(\epsilon)$. Then

$$p = \int_{\alpha_0 - \alpha}^{\alpha} f(\epsilon)\, d\epsilon \simeq \tfrac{1}{2} - (\alpha_0 - \alpha) f(0) \quad (4)$$

If as in [3]

$$p = \frac{1}{2}\left(1 - \frac{\mu}{\sqrt{n}}\right),$$

we have

$$\mu = 2\,(\alpha_0 - \alpha) f(0)\sqrt{n} \quad (5)$$

The limiting form of the distribution of R_n under the alternative hypothesis is a normal distribution with parameter,

$$\frac{np - \frac{n}{2}}{\sqrt{npq}} \simeq \mu$$

for large 'n'. Using this, the power of R_n for large 'n' has been computed and given in Table II for various values of 'μ'. For comparison the relevant results from [3] are reproduced.

TABLE II

Asymptotic power of five tests at ·05 *level for alternatives* $\alpha \neq \alpha_0$, $\beta = \beta_0$
Here $\mu = \sqrt{n}\{1 - 2 \text{ prob. } (\epsilon_i > (\alpha - \alpha_0))\}$

μ	m-Test	A	B	R_n	F-Test
·0	·05	·05	·05	·05	·05
·790	·09	·10	·10	·12	·13
1·316	·16	·20	·20	·26	·29
1·666	·25	·30	·30	·39	·45
1·958	·33	·40	·41	·50	·59
2·226	·42	·50	·51	·61	·71
2·493	·52	·60	·61	·70	·81
2·775	·62	·70	·71	·79	·89
3·104	·73	·80	·81	·87	·95
3·557	·85	·90	·91	·95	·99

The values given above show that R_n is more powerful than the other distribution-free tests namely m-test, A and B. Also it approaches the F-Test (in power) based on the assumption of normality (parameter $\frac{1}{2}\pi\mu^2$ and degrees of freedom 2).

Under the alternative hypothesis discussed above the criterion T has expectation $[n(n+1)]\,P/2$ and variance $[n(n+1)]\,pq/2$. Hence in terms of μ the parameter for T is

$$r = -\sqrt{\frac{n(n+1)\mu^2}{2(n-\mu^2)}}$$

and T is asymptotically normally distributed. For the criterion T_n^2 the corresponding parameter is r^2. The power of T and T_n^2 is not evaluated here as they behave like the 't' and x^2-tests respectively.

REFERENCES

1. Mood, A. M. .. *Introduction to the Theory of Statistics*, McGraw-Hill, New York, 1950.
2. Brown, G. W. and Mood, A. M. "On median tests of linear hypothesis," *Proceedings of the Second Berkely Symposium on Mathematical Statistics and Probability*, University of California, Berkeley, 1951, pp. 159–66.
3. Daniels, H. E. .. "A distribution-free test for regression parameters," *Annals of Math-Statistics*, 1954, **25**, 499–514.