ON LEHMANN'S TWO-SAMPLE TEST

By R. M. Sundrum

Division of Research Techniques, London School of Economics

Summary. This paper considers some properties of a two-sample test, suggested by Lehmann [2], against general alternatives. Alternative expressions are given for the test statistic; a general formula for the variance is derived and evaluated for the null case; the expectation is obtained in certain nonnull cases; and the exact distributions in the null case are tabulated for some small samples.

1. Introduction. A statistic for testing the null hypothesis that two independent random samples come from the same population against general alternatives (subject only to continuity of distribution functions) was proposed by Lehmann [2], based on the following lemma:

Lemma (4.1 of [2]). *Let X, X'; Y, Y' be independently drawn from populations with continuous cumulatives F, G respectively, and let us denote for any random variables U, U'; V, V' the event* max $(U, U') <$ min (V, V') *by $U, U' < V, V'$. Then*

$$p = P((X, X' < Y, Y') + (Y, Y' < X, X'))$$

$$= \tfrac{1}{3} + 2 \int (F - G)^2 d\left(\frac{F + G}{2}\right),$$

and hence p attains its minimum value $\tfrac{1}{3}$ if and only if $F = G$.

We can then base a test of the null hypothesis on a statistic which is a sample estimate of this probability p and test in the usual manner whether this sample estimate is significantly greater than $\tfrac{1}{3}$. For example, given a sample of X's and Y's, say of $2n$ observations each, we might choose n nonoverlapping quadruples at random each containing 2 X's and 2 Y's, and consider as our statistic the observed relative frequency of quadruples in which both X's are on the same side of both Y's. This procedure however appears to be wasting information. Lehmann has therefore suggested that it is more reasonable to consider the relative frequency of such quadruples among all the $\binom{m}{2}\binom{n}{2}$ possible quadruples that can be drawn from a sample of m X's and n Y's.

2. Alternative expressions for the test statistic. For practical purposes, Lehmann has given the following expression for the test statistic, which we denote by L.

Received 5/18/53.

$$L = \tfrac{1}{2}\binom{m}{2}^{-1}\binom{n}{2}^{-1}\Big\{(m-1)\sum_{i=1}^{m} R_i^2 - 2(m+n-2)\sum_{i=1}^{m} iR_i$$

$$- (m-2n+1)\sum_{i=1}^{m} R_i + \frac{(m+2n-3)m(m+1)(2m+1)}{6} \tag{1}$$

$$+ \tfrac{1}{2}m(m+1)(m+n^2-3n+1) - mn(n-1)\Big\}^{1}$$

([2], p. 174) where R_i is the rank of the ith ordered X-observation in the combined sequence of the $(m + n)$ members of the sample.

To see the structure of this statistic more clearly, write for the sample variance of the ranks R_i

$$S_R^2 = \frac{1}{m}\sum_{i=1}^{m}(R_i - \bar{R})^2 \quad \text{where} \quad \bar{R} = \frac{1}{m}\sum R_i$$

and for the sample "covariance" of i and R_i

$$C = \frac{1}{m}\sum_{i=1}^{m}\left(i - \frac{m+1}{2}\right)R_i.$$

Then, ignoring constant additive and multiplicative terms from (1), we have

$$L' = m(m-1)\left(\bar{R} - \frac{m+n+1}{2}\right)^2 + m(m-1)\,S_R^2 - 2m(m+n-2)C. \tag{2}$$

The test statistic has thus three components; the first term depending on the average location of the X's in the combined sequence, the second term depending on the dispersion of the R_i's and the last term depending on whether the X's are evenly spaced out as they should tend to be under the null hypothesis.

Alternatively, let (yxy) denote the event that when one X and two Y's are drawn independently from the respective populations, the X-value lies between the two Y-values; and let (xyx) denote the same event with X and Y interchanged. Then it follows quite simply that

$$p = 1 - P(yxy) - P(xyx). \tag{3}$$

Corresponding to the estimator L of p, we may consider as estimators of $P(yxy)$ and $P(xyx)$ the relative frequencies L_1 and L_2 respectively of the specified events among all possible triplets that can be drawn from the sample. In terms of ranks we have

$$L_1 = 2\sum_{i=1}^{m}(R_i - i)(n + i - R_i)/mn(n-1) \tag{4}$$

$$L_2 = 2\sum_{i=1}^{n}(S_i - i)(m + i - S_i)/mn(m-1) \tag{5}$$

where S_i is the rank of the ith ordered Y-observation in the combined sequence of the $(m + n)$ members of the sample. It can then be shown that

$$L = 1 - L_1 - L_2 \tag{6}$$

[1] The last term is omitted in Lehmann's formula.

for any sample. This gives us an expression for the test statistic L which is symmetrical in X and Y, and somewhat more convenient for practical use.

3. Expectation and variance of L. Let

$$D(i, j; k, l) = 1 \text{ if } X_i, X_j < Y_k, Y_l \text{ or } Y_k, Y_l < X_i, X_j \qquad (i \neq j; k \neq l)$$
$$= 0 \text{ otherwise.}$$

Then

$$\binom{m}{2}\binom{n}{2} L = \sum_i \sum_j \sum_k \sum_l D(i, j; k, l) \qquad (i < j; k < l) \tag{7}$$

consisting of $\binom{m}{2}\binom{n}{2}$ terms. Therefore

$$E(L) = E(D(i, j; k, l)) = p = P((X, X' < Y, Y') + (Y, Y' < X, X')). \tag{8}$$

In the null case, when $F = G$, we have $p = 1/3$ from the above lemma of Lehmann, or from the consideration that of the six possible arrangements in order of magnitude of the members of a single quadruple, all equally probable under the null hypothesis

$$x\ x\ y\ y;\ x\ y\ x\ y;\ x\ y\ y\ x;\ y\ x\ x\ y;\ y\ x\ y\ x;\ y\ y\ x\ x;$$

in two arrangements only do both X's lie on the same side of both Y's.

Further, from (7)

$$\binom{m}{2}^2\binom{n}{2}^2 L^2 = \left\{\sum_i \sum_j \sum_k \sum_l D(i, j; k, l)\right\}^2 \qquad (i < j; k < l) \tag{9}$$

consisting of $\binom{m}{2}^2\binom{n}{2}^2$ terms which can be grouped in the following nine classes of terms, involving the expectation terms shown against each class

Term	*Expectation*	*Number of terms.* $\binom{m}{2}\binom{n}{2}$ *times*
$D^2(i, j; k, l)$	p	1
$D(i, j; k, l)D(i, m; k, l)$	r	$2(m - 2)$
$D(i, j; k, l)D(i, j; k, f)$	s	$2(n - 2)$
$D(i, j; k, l)D(m, n; k, l)$	t	$\frac{1}{2}(m - 2)(m - 3)$
$D(i, j; k, l)D(i, j; f, g)$	u	$\frac{1}{2}(n - 2)(n - 3)$
$D(i, j; k, l)D(i, m; k, f)$	v	$4(m - 2)(n - 2)$
$D(i, j; k, l)D(m, n; k, f)$	a	$(m - 2)(m - 3)(n - 2)$
$D(i, j; k, l)D(i, m; f, g)$	b	$(m - 2)(n - 2)(n - 3)$
$D(i, j; k, l)D(m, n; f, g)$	p^2	$\frac{1}{4}(m - 2)(m - 3)(n - 2)(n - 3)$

(i, j, m, n all different, k, l, f, g all different.)

Collecting terms together and simplifying, we get

$$\begin{aligned}\binom{m}{2}\binom{n}{2}\sigma^2(L) &= (a - p^2)m^2 n + (b - p^2)mn^2 \\ &+ (4v + 6p^2 - 5a - 5b)mn + (\tfrac{1}{2}t + \tfrac{3}{2}p^2 - 2a)m^2 \\ &+ (\tfrac{1}{2}u + \tfrac{3}{2}p^2 - 2b)n^2 + (2r - \tfrac{5}{2}t + 10a + 6b - 8v - \tfrac{15}{2}p^2)m \\ &+ (2s - \tfrac{5}{2}u + 6a + 10b - 8v - \tfrac{15}{2}p^2)n \\ &+ (p + 3t + 3u + 16v + 9p^2 - 4r - 4s - 12a - 12b).\end{aligned} \tag{10}$$

For evaluating the parameters occurring in the above expression, it is convenient to express them in terms of the probabilities of certain ordered arrangements of a given number of X's and Y's drawn at random from the respective populations. In the following, we extend the notation of Section 2 and denote by expressions like, for example, $(xyxy)$ the event that when two X's and two Y's are drawn at random and arranged in order of magnitude, they have the indicated arrangement.

$$\begin{aligned} p &= P\{(xxyy) + (yyxx)\} \\ r &= P\{(xxxyy) + (yyxxx)\} \\ t &= P\{(xxxxyy) + (yyxxxx)\} + \tfrac{1}{3}P(xxyyxx) \\ v &= P\{(xxxyyy) + (yyyxxx)\} + \tfrac{2}{9}P\{(xxyxyy) + (yyxyxx)\} \\ a &= P\{(xxxxyyy) + (yyyxxxx)\} \\ &\quad + \tfrac{1}{3}P\{(xxxyxyy) + (yyxyxxx) + (xxyyyxx)\} \\ &\quad + \tfrac{1}{9}P\{(xxyxxyy) + (yyxxyxx) + (xxyyxxy) + (yxxyyxx) \\ &\qquad\qquad + (xxyyxyx) + (xyxyyxx)\} \\ &\quad + \tfrac{1}{18}P\{(xyxyxxy) + (yxxyxyx) + (yxxyxxy) + (xyxyxyx)\}.\end{aligned} \tag{11}$$

Similar formulae for s, u and b can be derived from those for r, t and a by interchanging x and y.

These probabilities can be evaluated very simply in the null case from the property that all permutations of the ordered sequence of x's and y's are equally probable. Then

$$p = \tfrac{1}{3}; \quad r = s = \tfrac{1}{5}; \quad t = u = \tfrac{7}{45}; \quad v = \tfrac{11}{90}; \quad a = b = \tfrac{1}{9}. \tag{12}$$

Substituting these values in (10), we find for the null case

$$\sigma^2(L) = \frac{4\{(m + n)(m + n - 1) - 2\}}{45mn(m - 1)(n - 1)} \tag{13}$$

and when $m = n$,

$$\sigma^2(L) = \frac{8(2n + 1)}{45n^2(n - 1)} \tag{14}$$

The expectation of L can be obtained in certain nonnull cases by the use of (3).

(i) *Rectangular distributions.*

(a) Difference in location. Let X be uniformly distributed in the range 0 to 1, and Y be uniformly distributed in the range Δ to $1 + \Delta$. Then it follows by simple integration that

$$P(yxy) = P(xyx) = 1/3 - \Delta^2 + \frac{2\Delta^3}{3}, \qquad (0 \leqq \Delta \leqq 1)$$

so that

$$p = 1/3 + 2\Delta^2 - \frac{4\Delta^3}{3}. \tag{15}$$

(b) Difference in scale. Let X be uniformly distributed in the range $-\frac{1}{2}$ to $+\frac{1}{2}$, and Y be uniformly distributed in the range $-\Delta$ to $+\Delta$, where $\Delta > \frac{1}{2}$. Then we have

$$P(yxy) = \tfrac{1}{2} - 1/24\Delta^2, \qquad P(xyx) = 1/6\Delta$$

so that

$$p = (12\Delta^2 - 4\Delta + 1)/24\Delta^2 \tag{16}$$

(ii) *Normal distributions.*

(a) Difference in location. Let X and Y be normally distributed with the same variance σ^2 and means μ_1 and μ_2 respectively, where $\mu_2 - \mu_1 = \delta\sigma$. If x is an observation on X and y_1 and y_2 are two observations on Y, and if we define

$$u_1 = x - y_1, \qquad u_2 = x - y_2$$

u_1 and u_2 are jointly distributed in the bivariate normal form with means $-\delta\sigma$, variances $2\sigma^2$ and correlation coefficient $\frac{1}{2}$. Then

$$\begin{aligned} P(yxy) = P(u_1 u_2 < 0) &= 2\int_{\delta/\sqrt{2}}^{\infty}\int_{-\infty}^{\delta/\sqrt{2}} \frac{1}{2\pi\sqrt{1-\rho^2}} \\ &\quad \exp\left\{-\frac{1}{2(1-\rho^2)}[t_1^2 - 2\rho t_1 t_2 + t_2^2]\right\} dt_1\, dt_2 \text{ with } \rho = \tfrac{1}{2}. \end{aligned} \tag{17}$$

We also find the same value for $P(xyx)$. These values have been tabulated for various values of δ in [3] and can be used to evaluate p.

(b) Difference in scale. Let X and Y be normally distributed with the same mean, say 0, and variances σ_x^2 and $\sigma_y^2 \neq \sigma_x^2$. If u_1 and u_2 are defined as in the previous case, they are now jointly distributed in the bivariate normal form with means 0, variances $\sigma_x^2 + \sigma_y^2$ and correlation coefficient equal to $\sigma_x^2/(\sigma_x^2 + \sigma_y^2.)$ Therefore

$$P(yxy) = P(u_1 u_2 < 0) = \tfrac{1}{2} - \frac{1}{\pi}\sin^{-1} \sigma_x^2/(\sigma_x^2 + \sigma_y^2).$$

By a similar argument, we find

$$P(xyx) = \tfrac{1}{2} - \frac{1}{\pi}\sin^{-1} \sigma_y^2/(\sigma_x^2 + \sigma_y^2).$$

Hence, we have

$$p = \frac{1}{\pi}\left\{\sin^{-1}\frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2} + \sin^{-1}\frac{\sigma_y^2}{\sigma_x^2 + \sigma_y^2}\right\}. \tag{18}$$

These methods of evaluating p can then be extended to cases where both location and scale are different in rectangular and normal populations.

4. The distribution of L. In the null case, the exact distribution of L may be computed for small samples by enumerating the whole set of equiprobable permutations. As for the limiting case, L is an extension of a U-statistic defined by Hoeffding [1] and by Lehmann's Theorem 3.2 ([2], p. 167), $\sqrt{n}(L - E(L))$ has a limiting normal distribution under the condition m/n = constant. However in the null case, the variance of L is of order n^{-2} and the limiting normal distribution of $\sqrt{n}(L - E(L))$ is singular.

Some idea of the exact distribution in the null case may be obtained from the following tables for small samples, which were obtained by complete enumeration of the various possibilities.

$m = n = 2$

x	$6P(L = x)$	$P(L \geq x)$
0	4	1.0000
1	2	0.3333

$m = n = 3$

x	$20P(9L = x)$	$P(9L \geq x)$
1	8	1.0000
3	8	0.6000
5	2	0.2000
9	2	0.1000

$m = 2; \quad n = 3$

x	$10P(3L = x)$	$P(3L \geq x)$
0	4	1.0000
1	4	0.6000
3	2	0.2000

$m = 3; \quad n = 4$

x	$35P(18L = x)$	$P(18L \geq x)$
2	4	1.0000
3	8	0.8857
4	4	0.6571
5	4	0.5429
6	5	0.4286
8	2	0.2857
9	4	0.2286
12	2	0.1143
18	2	0.0571

$m = n = 4$

x	$70P(36L = x)$	$P(36L \geq x)$
6	16	1.0000
9	24	0.7714
12	12	0.4286
15	2	0.2571
18	8	0.2286
21	4	0.1143
27	2	0.0571
36	2	0.0286

$m = 4; \quad n = 5$

x	$126P(60L = x)$	$P(60L \geqq x)$
10	4	1.0000
11	8	0.9683
12	12	0.9048
13	12	0.8095
14	4	0.7143
15	12	0.6825
16	9	0.5873
17	4	0.5159
18	12	0.4841
20	3	0.3889
21	8	0.3651
22	8	0.3016
24	2	0.2381
25	2	0.2222
26	2	0.2063
27	4	0.1905
30	4	0.1587
31	2	0.1270
33	2	0.1111
36	4	0.0952
39	2	0.0635
40	2	0.0476
48	2	0.0317
60	2	0.0159

$m = n = 5$

x	$252P(100L = x)$	$P(100L \geqq x)$
20	32	1.0000
24	64	0.8730
28	48	0.6190
32	16	0.4286
36	26	0.3651
40	24	0.2619
44	6	0.1667
48	8	0.1429
52	6	0.1111
60	10	0.0873
64	4	0.0476
72	4	0.0317
84	2	0.0159
100	2	0.0079

I am much indebted to Mr. William Kruskal for many suggestions which have greatly improved the form of this paper.

REFERENCES

[1] W. Hoeffding, "A class of statistics with asymptotically normal distributions," *Ann. Math. Stat.*, Vol. 19 (1948), pp. 293–325.

[2] E. L. Lehmann, "Consistency and unbiasedness of certain nonparametric tests," *Ann. Math. Stat.*, Vol. 22 (1951), pp. 165–179.

[3] K. Pearson, *Tables for Statisticians and Biometricians*, Part II, 1st ed. Cambridge University Press, 1931.