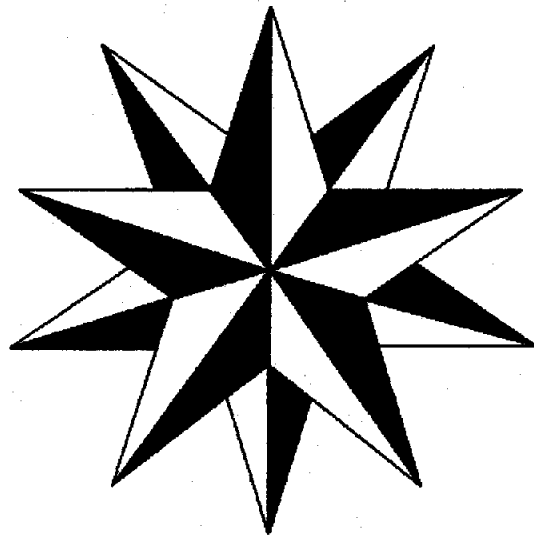


**THE PAIR CHART MACRO: A QUICK GRAPHICAL
REPRESENTATION OF THE TWO-SAMPLE PROBLEM**



**Michael Riggs
Section of Wildlife
Minnesota Department of Natural Resources**

The Pair Chart Macro: A Quick Graphical Representation
of the Two-Sample Problem

Michael Riggs

Minnesota Department of Natural Resources

Section of Wildlife

St Paul, MN 55155 (612) 296-0702

Abstract

Tests for equality of means (e.g., t-tests and f-tests) frequently are applied to data whose the residuals meet fairly restrictive assumptions. The principal use of these tests is to make an inference about the relative magnitude of the scores in one group vs. those in another. Textbooks generally illustrate this application by depicting homoscedastic normal distributions whose positions are shifted relative to one another on the horizontal axis. Rank-based tests for shifts in location can be used to examine the same kinds of hypotheses but with less restrictive assumptions. It has often been argued that the rank-based tests do not have the same intuitive appeal as their normal theory-based counterparts and that they are much harder for nonstatisticians to grasp. Quade (1973) developed a graphic device, which he called the "pair chart", for illustrating the two-sample problem for ranked data. I will demonstrate this use of the pair chart macro and will derive from it, two commonly used nonparametric test statistics. Finally, I will use a series of pair charts to compare the

results of nonparametric tests with those of parametric tests on transformed and untransformed data from increasingly dispersed distributions.

Introduction

For most students of science and engineering, the t-test for the equality of two sample means and its k-sample counterpart, the f-test, provide the formal introduction to statistical inference. For a large majority of these students, such "parametric" or, more correctly, normal theory-based tests will become the preferred test for a "statistically significant" difference in the magnitude a random variable between two or more samples. The normal theory-based tests rely on two distinguishing properties of the normal distribution in order for inference to be valid: symmetry of the distribution and the location of the mean at the center of the distribution (i.e., the mean=the median). When two such distributions have approximately the same dispersion (i.e., they are homoscedastic), a significant difference in the values of the sample means will imply a shift in the location of one or more of the samples. Such

a result indicates a general difference in magnitude of the values of the random variable between the two samples and between their respective target populations. However, when one or more of the distributions is not normal and/or they are not equally dispersed, then an observed difference in the sample means becomes difficult to interpret.

One solution to this problem is to employ a location test which is not so dependent on the form of the underlying distributions. Rank-based tests such as the Mann-Whitney-Wilcoxon (MWW) and the Kruskal-Wallis (KW) test represent two such distribution-free or "nonparametric" tests. Although the nonparametric tests are more robust and hence more reliable tests for location shifts, they seem to be grossly under-utilized in the scientific literature. It has been suggested that the main reason for this is that the rank-based tests lack the intuitive appeal of the normal-theory based tests (Draper 1988). Because the normal distribution occurs so commonly in nature, most scientists and, indeed, many nonscientists generally equate a substantial difference in means with an overall general difference in location. The same individuals are usually unaware of the fallacy of this interpretation when (for instance) the distributions have heavy tails. Likewise, it is difficult for most nonstatisticians to envision just what it is that a rank-based test is testing.

In the following discussion I shall present a simple graphical method for illustrating what a rank-based test of location is and how it can be interpreted. I will also give a brief example of how this method can be used to compare the performance of

nonparametric tests with corresponding parametric tests on transformed and untransformed data.

Constructing a Pair Chart

Suppose we have two samples X and Y with n_x and n_y observations respectively on a rankable random variable. A pair chart may be constructed as follows (Quade, 1973):

1. Pool the samples and order the observations from the smallest to the largest value, noting from which sample (X or Y) each value came.
2. Draw a rectangle of width n_x units and height n_y units.
3. Starting at the lower left corner of the rectangle (i.e., the origin), draw a line one unit to the right if the smallest value from the pooled samples is an X; if it is a Y, draw the line one unit up instead.
4. From the end of this first line draw a second line, one unit to the right if the second smallest observation is an X, one unit up if it is a Y.
5. Continue in this manner for all $(n_x + n_y)$ observations in the combined sample, until the path so formed terminates at the upper right corner of the rectangle.

The following data from Quade (1973) illustrate this process.

X	19, 25, 28, 30, 30, 36, 50, 52, 57, 67	$n_x = 10$
Y	24, 31, 33, 37, 38, 42, 49	$n_y = 7$

The ordering of the pooled sample of these data is:

X Y X X X X Y Y X Y Y Y X X X X

Thus if we follow rules 1 - 5 from above, we obtain the pair

chart shown in Fig. 1. (Example B from Quade, 1973). It may be noted that within-sample ties (e.g., two scores of 30 in sample X) do not pose any special problems for the construction of the pair chart. However, between-sample ties do complicate the picture somewhat.

Consider these data from Quade (1973, example C):

X	1,2,2,2,3,4,4,7	$n_x = 8$
Y	1,2,2,3,3,3,5,9	$n_y = 8$

The corresponding ordering of the pooled samples is:

(X Y) (X X X Y Y) (X Y Y Y) X X Y X Y

Observations within each pair of parentheses are all tied on the same value. The instructions in rules 3 and 4 (above) become ambiguous when there are between-sample ties. For example, the smallest value in the pooled sample of this data set, 1, is contained in both the X and Y samples and so it is difficult to decide whether to begin drawing the path one unit up from or one unit to the left of the origin. The solution is to start at the origin and draw a box 1 unit by 1 unit (Fig. 2). Similarly for the 5 values tied on 2 we may note that 3 of them are in the X sample and 2 are in the Y sample. Thus, we draw a box of dimensions 2x3 units with its lower left corner positioned at the upper right corner of the previous box. We continue in this manner until the path terminates in the upper right corner of the pair chart, as in the previous example.

Interpretation of the Pair Chart

The pair chart represents all the possible $n_x * n_y$ pairwise comparisons between the observations in samples X and Y. For example, in Fig 1., the first column is the comparison of each

of the 7 observations in Y (i.e., 24-49) vs. the smallest valued observation in X (i.e., 19). Since every unit in the column lies above the path of the pair chart we can conclude that the values of every observation in Y are larger than the smallest value in X. Likewise, if we examine column two, we observe that the 6 largest values in Y (31-49) which lie above the path are larger than the second smallest value in X (25); and so on. In this way, the path of the pair chart defines two areas such that the upper zone represents the proportion of the $n_x * n_y$ comparisons in which the values of sample Y exceed those of sample X, while the lower zone represents the opposite case.

If the values of one sample do not tend to be larger or smaller than those of the other sample we would expect that the path of the pair chart would divide the rectangle into two zones each with area approximately equal to $(n_x * n_y)/2$. In Fig. 1 the upper area (U_y) = the lower area (U_x) = 35 units, indicating that the values of the observations in the two samples are of comparable magnitude; i.e., the two sample distributions have equal location.

In Fig. 2, it is possible to discern three areas. In addition to the upper and lower regions there is also an area within the tie-boxes. Since the tied area constitutes the proportion of the $n_x * n_y$ comparisons in which neither the values of X nor Y is larger than the other, this area is not useful in deciding which sample tends to have the larger values. In Fig. 2 $U_y=29$, $U_x=25$ and the tied area=10 units. Again, the two samples appear to have similar locations.

Pair charts are also useful for comparing the relative dispersion of two samples. For the

comparison represented in Fig. 1, 5 of the 6 smallest values and the 4 largest values all come from sample X. This is apparent from the two long horizontal runs of the path near the origin and again near the upper right corner of the chart. Configurations of this sort can be interpreted as an indication of unequal dispersion and hence heteroscedasticity. By contrast, a pair chart whose path ascends in a regular stair stepping pattern, with few long runs early or late, is an indication of similarity in dispersion.

Derivation of the Mann-Whitney and Wilcoxon Statistics

The two-sample test for a difference in location can be expressed as the test of,

$$(1) \quad H_0: CDF_x = CDF_y \text{ vs.} \\ H_a: CDF_x \neq CDF_y$$

or, alternatively,

$$(2) \quad H_0: \text{Prob}(X < Y) = 1/2 \text{ vs.} \\ H_a: \text{Prob}(X < Y) \neq 1/2$$

Expressed in form (2), the two-sample location test is the algebraic counterpart to the pair chart. To test the hypotheses in (2) on the data in Fig. 2, we can derive either the Mann-Whitney U or the Wilcoxon T statistic as follows:

$$U = U_x + 1/2 \text{ (tie-box area)}$$

$$T = U_x + [(n_x * (n_x + 1))/2] + 1/2 \text{ (tie-box area)}$$

under H_0 , the large sample variance of both U and T is,

$$(3) \quad [n_x * n_y (n_x + n_y + 1)]/12$$

Thus, both the MW and the Wilcoxon statistics can be viewed as functions of the area below (or above) the path of the pair chart, each of which has a well defined sampling distribution

under the assumption that this area is $(n_x * n_y)/2$.

In addition to demonstrating the logic behind the use of the MW statistics to test for shifts in location, this derivation of U and T nicely illustrates the effects of ties on the power of the tests. The practice of assigning half the tied area to U_x and half to U_y is computationally equivalent to assigning to the tied scores the mean of their ranks (as is done by PROC NPAR1WAY). It is apparent from Fig 2 that this will not produce a meaningful change in the relative proportions of the areas above or below the path of the pair chart. More importantly, since each tied unit subtracts from the $(n_x * n_y)$ units which potentially are available to determine if $U_x = U_y$, the presence of tied observations reduces the effective sample size and thus the power of the test.

The Pair Chart Macro

The pair chart macro is actually a combination of three macros which together produce a GPLOT representation of Quade's pair chart for any two samples. Prior to calling the Pair macro, each sample must be output to a separate data set and two formats, \$SAMPF. and \$SAMP. must be defined. Once these are done the Pair macro will:

1. Scale the vertical and horizontal axes so that the length of the axis for the larger sample is 5.8 in. and the length of the axis for the smaller sample is proportionately shorter than 5.8 in. When the sample sizes are equal, both axes are 5.8 in.
2. Increment the axes as specified in the macro call.

3. Create formats for the axis values so that the actual value of the appropriate ordered observation is printed at each major tick mark.
4. Form the axis labels by appending the appropriate sample size values to the values of the \$SAMPF. formats.
5. Run PROC NPARIWAY to produce Wilcoxon and ANOVA tests for each pair of samples and label the output with the values from the \$SAMPF. format.
6. Draw the path of the pair chart and a diagonal reference line.
7. Compute and print in a footnote:
 - U_x = the area under the path of the pair chart
 - U_y = the area above the path of the pair chart
 - Ties = the area within the tie-boxes
 - T_x = The area below the path + $[(n_x * (n_x + 1))/2]$
8. Print titles as specified in the macro call

The following program produced Figs. 1 and 2 by calling the PAIR macro. The form of the macro call is: (name of the data set to be represented on the horizontal axis, name of the data set to be represented on the vertical axis, name of the variable common to the two data sets which is to be compared, the incremental unit of the horizontal axis, the incremental unit of the vertical axis).

```
*****
* PRODUCE THE PAIR CHARTS FROM QUADE, 1973.
*****
* CREATE A FORMAT (VALUE= $SAMPF.) FOR
* LABELING THE PAIR CHART AXES AND ANOTHER
* FORMAT (VALUE= $SAMPF.) FOR LABELING THE
* OUTPUT OF PROC NPARIWAY. NOTE: NPARIWAY
* DOES NOT PRINT FORMAT VALUES WHICH ARE
* LONGER THAN 15 CHARACTERS.
*
*****;
PROC FORMAT;
  VALUE $SAMPF 'X1'='SAMPLE X, EXAMPLE A'
               'X2'='SAMPLE X, EXAMPLE B'
               'X3'='SAMPLE X, EXAMPLE C'
               'Y1'='SAMPLE Y, EXAMPLE A'
               'Y2'='SAMPLE Y, EXAMPLE B'
               'Y3'='SAMPLE Y, EXAMPLE C';
  VALUE $SAMP 'X1'='SAMP. X EXAM. A'
               'X2'='SAMP. X EXAM. B'
               'X3'='SAMP. X EXAM. C'
               'Y1'='SAMP. Y EXAM. A'
               'Y2'='SAMP. Y EXAM. B'
               'Y3'='SAMP. Y EXAM. C';

RUN;
      /**  FETCH THE PAIR CHART MACRO  ***/

%INCLUDE PAIR / NOSOURCEZ;
*****;
* READ IN THE DATA FOR THE 3 EXAMPLES FROM
* QUADE, 1973.
*
*****;
DATA X1 Y1
      X2 Y2
      X3 Y3;
  INPUT SAMPLE $ VALUE @@;
  K=1 ;      /** K=1 required for drawing
              ref. diagonal **/
  IF SAMPLE='X1' THEN OUTPUT X1 ;
  ELSE IF SAMPLE='Y1' THEN OUTPUT Y1 ;
  ELSE IF SAMPLE='X2' THEN OUTPUT X2 ;
  ELSE IF SAMPLE='Y2' THEN OUTPUT Y2 ;
  ELSE IF SAMPLE='X3' THEN OUTPUT X3 ;
  ELSE IF SAMPLE='Y3' THEN OUTPUT Y3 ;

CARDS ;
X1 18  X2 19  X3 1
X1 20  X2 25  Y3 1
X1 30  X2 28  X3 2
X1 32  X2 30  X3 2
X1 36  X2 30  X3 2
X1 38  X2 36  Y3 2
X1 39  X2 50  Y3 2
X1 41  X2 52  X3 3
X1 51  X2 57  Y3 3
X1 70  X2 67  Y3 3
Y1 28  Y2 24  Y3 3
Y1 43  Y2 31  X3 4
Y1 46  Y2 33  X3 4
Y1 46  Y2 37  Y3 5
Y1 50  Y2 38  X3 7
Y1 56  Y2 42  Y3 9
Y1 64  Y2 49
Y1 79
; RUN;
```

```

*****;
* CALL THE PAIR MACRO 3 TIMES, TO
  PRODUCE THE 3 PAIR CHARTS
*
*****;
ZLET TIT1='PAIR CHARTS COMPARING
3 PAIRS OF SAMPLES (QUADE, 1973)';
ZLET TIT2='';

ZPAIR (X1,Y1,VALUE,1,1) ;
  /** pair chart not shown **/
RUN;
ZPAIR (X2,Y2,VALUE,1,1) ;
  /** pair chart in Fig. 1 **/
RUN;
ZPAIR (X3,Y3,VALUE,1,1) ;
  /** pair chart in Fig. 2 **/
RUN;

```

Some Practical Applications

Although the MWW tests frequently are used *whenever* the assumptions of the normal theory-based tests are not met, inference is *not* justified when the sample variances are grossly unequal. When sample variances are unequal, the variance estimator of the MWW statistics is much more complex than expression (3), above (Zaremba, 1962). In such cases, a p-value based on (3) will be incorrect. Interpretation of location tests in these circumstances is called the Behrens-Fisher Problem and has been the subject of considerable research. I have found the construction of pair charts to be a useful tool in interpreting the results of rank-based location tests when the scale parameters are unequal. Figure 3 is an example of this application. The two samples were generated from parent distributions with equal means (28) but unequal variances (1000 vs. 12300). It is clear from the pair chart that the two heavy-tailed distributions have different locations despite having come from populations with the same mean. Given the sample sizes, It does *not* seem very likely that these two samples come from populations which have similar locations. Based on these results, I would be inclined to concur with the results of the MWW test which rejects the hypothesis of equal locations ($T_x = 2250 + 180/2 = 2340$; $P < .0001$).

I have attempted to investigate the Behrens-Fisher problem further by comparing the performance of t-tests on log-transformed and untransformed data vs. that of the MWW tests vs. the corresponding pair charts. I generated samples from binomial, Poisson and negative binomial populations with equal means but with variances ranging from 14 - 28000 and performed all possible pairwise tests among them. The data in Fig. 3 represent one such pair. Despite the severe violations of the assumptions, the t-tests on the untransformed data did what they were supposed to do. In nearly every case they indicated that the samples came from populations with the same means ($P > .05$). The t-tests on the transformed data and the MWW tests, on the other hand, uniformly rejected the null hypotheses of (respectively) equal *geometric* means, and equal locations. These later results were in accordance with the patterns exhibited in the corresponding pair charts. Interestingly, the log transformation did not consistently correct for heteroscedasticity in the data (based on Levene's test). Thus, neither the transformation nor the rank-tests produced p-values that were, in a strict sense, any more correct than the t-tests on the untransformed data. Nonetheless, the pair charts confirm that the t-test on the log-transformed data and the MWW test consistently provide reasonable tests of location even in the face of severe violations of their underlying assumptions. By contrast, the t-test (even with Welch's adjustment), on the untransformed heavy-tailed data do not provide a very useful result.

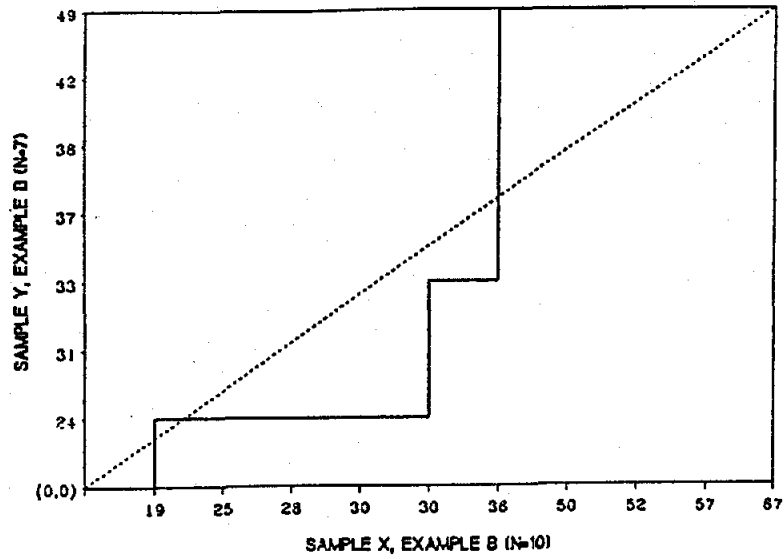
References

Draper, D. 1988. Rank-based robust analysis of linear models: I. exposition and review. Stat. Sci. 3 (2):239-271.

Quade D, 1973. The pair chart. Stat. Neerland. 27 (1):29-45.

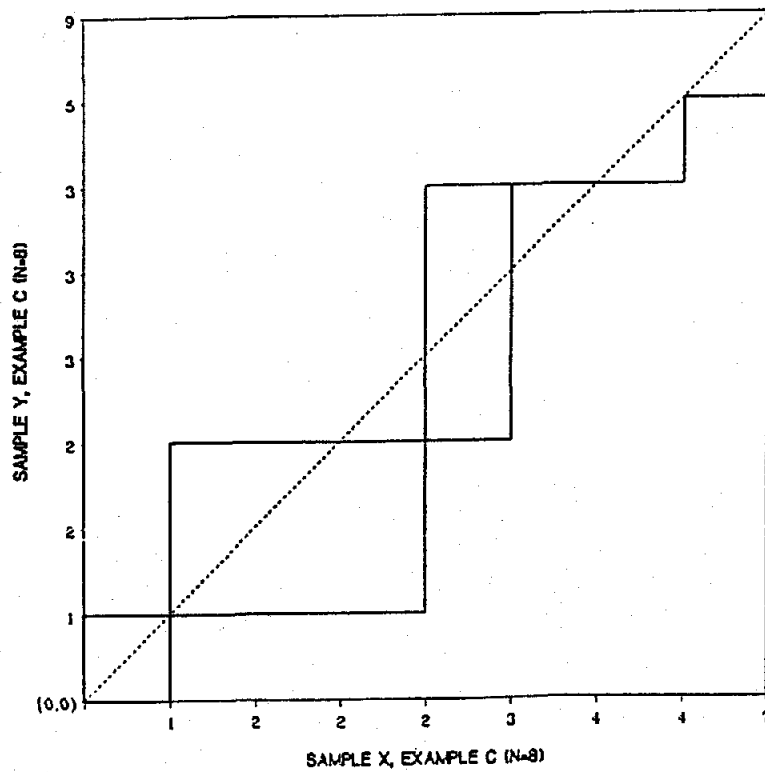
Zaremba, S.K. 1962. A generalization of Wilcoxon's test. Monat. fur Math. 66:359-370.

FIGURE 1.
PAIR CHARTS COMPARING 3 PAIRS OF SAMPLES (QUADE, 1973)



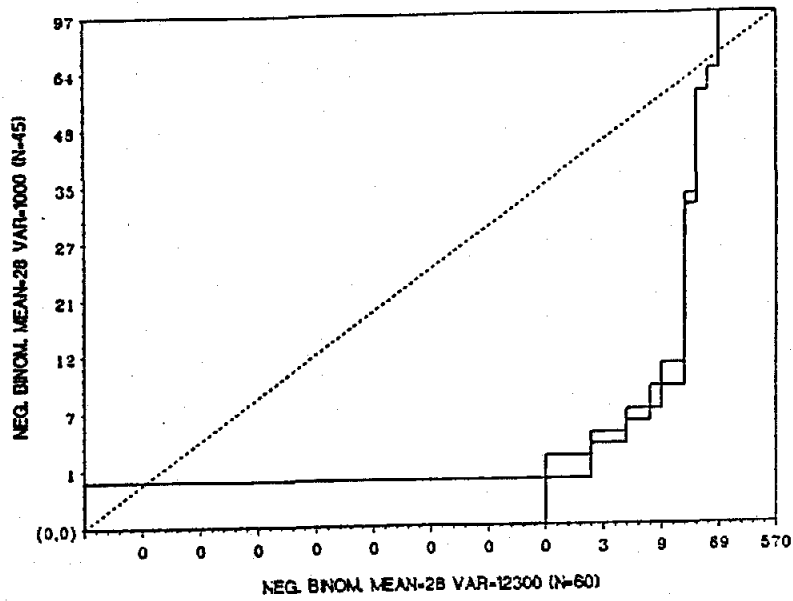
UY = 35 UX = 35 TIES = 0 TX = 90

FIGURE 2.
PAIR CHARTS COMPARING 3 PAIRS OF SAMPLES (QUADE, 1973)



UY = 29 UX = 25 TIES = 10 TX = 61

FIGURE 3.
PAIR CHARTS COMPARING NEGATIVE BINOMIAL VS. POSITIVE BINOMIAL VS.
POISSON DISTRIBUTIONS WITH EQUAL MEANS BUT UNEQUAL VARIANCES



UY = 2100 UX = 420 TIES = 180 TX = 2250