

Comparing Dispersions

Phillip Good
205 W. Utica Ave.
Huntington Beach CA 92648
drgood@statcourse.com

Permutation methods provide both exact and more powerful tests for comparing dispersions from two or many populations with immediate applications in quality control and as a pre-test for homoscedasticity before applying the analysis of variance. The results can be extended to obtain a more powerful test for treatment effects when non-responders are present.

Keywords and Phrases: dispersion; permutation tests; two-sample comparisons; k-sample comparisons; homoscedasticity; non-responders; power.

1. MOTIVATION

Precision is essential in a manufacturing process. Items that are too far out of tolerance must be discarded and an entire production line brought to a halt if too many items exceed (or fall below) designated specifications. With some testing equipment, such as that used in hospitals, precision can be more important than accuracy. For *accuracy* (closeness to the correct value) can always be achieved

through the use of standards with known values, while a lack of *precision* may render an entire sequence of tests invalid. Thus tests for consistency in dispersion are essential.

The p-values obtained from an analysis of variance can be heavily distorted by inequality of the variances; a preliminary test for equality of variances with exact significance level is essential.

2. THE ASYMPTOTIC APPROACH

There is no shortage of methods to test the hypothesis that two samples come from populations with the same inherent variability. Sukhatme [1958] lists four alternative approaches and adds a fifth of his own; Miller [1968] lists ten alternatives and compares four of these with a new test of his own; Conover, Johnson and Johnson [1981] list and compare 56 tests; and Balakrishnan and Ma [1990] list and compare nine tests with one of their own.

None of these tests can be relied on. Many promise an error rate or significance level of 5% but in reality make errors as frequently as 8% to 20% of the time. For example, with the test proposed by Miller [1968], a 10% test with samples of size 12 and 8 taken from normal populations yielded Type I errors 14% of the time.

This is because even the so-called non-parametric methods make use of asymptotic parametric approximations to derive their cut-off values. Consequently, the claimed power for small and mid-sized

samples is excessive as it is based on cut-off values that are accurate only for very large samples. To reassess the value of these methods our own simulations utilized a two-stage procedure:

- First, the necessary cut-off values were determined by simulation under the null hypothesis.
- These values were employed in subsequent power calculations under the alternative(s) of interest.

None of the methods cited above proved at all powerful when the correct significance levels were employed. Indeed, the actual power was only 70% to 85% of the claimed power when testing against normal and double exponential distributions, for sample sizes of 6 to 12 in number, and a claimed power of 80% or less. In other words, the power claimed for these tests had been obtained at the expense of Type I errors in excess of that specified and desired.

All the tests including those proposed here require that the observations be independent or, at least, exchangeable. The F-ratio test (Fisher, 1925) is exact only if the observations come from a normal distribution, and unlike the t-test, is very sensitive to deviations from normality.

The other previously-proposed tests have more restrictions. Each requires that two or more of the following four conditions be satisfied:

1. The observations be normally distributed.
2. The location parameters of the two distributions be the same or differ by a known quantity, Duran[1976].

3. The two samples be equal in size.
4. The samples be large enough that asymptotic approximations to the distribution of the test statistic are valid.

As an example, the first published solution to this classic testing problem is the z-test proposed by Welch [1937] based on the ratio of the two sample variances. If the observations are normally distributed, this ratio has the F-distribution, and the test whose critical values are determined by the F-distribution is uniformly most powerful among all unbiased tests (Lehmann, 1986, section 5.3). But with even small deviations from normality, significance levels based on the F-distribution are grossly in error, (Lehmann, 1986, page 207); the magnitude of the error will depend on the 4th moment of the distribution from which the samples are drawn.

Box and Anderson [1955] propose a correction to the F-distribution for "almost" normal data, based on an asymptotic approximation to the permutation distribution of the F-ratio. Not surprisingly, their approximation is close to correct only for normally distributed data or for very large samples. The Box-Anderson statistic results in an Type I error rate of 21%, twice the claimed and desired value of 10%, when two samples of size 15 are drawn from a gamma distribution with four degrees of freedom.

3. THE PERMUTATION APPROACH

At first glance, the permutation test for comparing the variances of two populations would appear to be an immediate extension of the test used for comparing location parameters, the distinction being that we make use of the squares of the observations rather than the observations themselves. But these squares are actually the sum of two components, one of which depends upon the unknown variance, the other upon the unknown location parameter. That is, $E(X-\mu+\mu)^2 = E(X-\mu)^2 + 2\mu E(X-\mu) + \mu^2 = \sigma^2 + 0 + \mu^2$. A permutation test based upon the squares of the observations is appropriate only if the location parameters of the two populations are known or are known to be equal (Hayes, 1997).

We cannot eliminate the effects of the location parameters by working with the deviations about each sample mean as these deviations are interdependent (Maritz, 1981).

3.1. Aly's Test Statistic

Good[2000] proposed a test based on the permutation distribution of the statistic described by Aly[1990],

$$S_A = \sum_{i=1}^{m-1} i(m-i)(X_{(i+1)} - X_{(i)})$$

where $X_{(1)} < X_{(2)} < \dots < X_{(m)}$ are the order statistics of the first sample.

That is, $X_{(1)}$ is the smallest of the observations in the first sample (the

minimum), $X_{(2)}$ is the second smallest and so forth, up to $X_{(m)}$ the maximum.

As S_A puts its greatest weight on differences in the center of the distribution the effect of outliers is minimized.

Aly[1990] made use of a far-from-exact asymptotic parametric approximation to the distribution of S_A . Fortunately, the test based upon the permutation distribution of Aly's statistic is exact and is unbiased when testing within the family of distributions which differ only in their location and scale parameters, $F_F = \{F[(x-\mu)/\sigma]\}$.

To illustrate the application of Aly's statistic, suppose the first sample consists of the measurements 121, 123, 126, 128.5, 129 and the second sample of the measurements 153, 154, 155, 156, 158. $X_{(1)}=121$, $X_{(2)}=123$ and so forth.

Set $\{z_{1i}\}$ equal to the differences between successive order values in the first sample, $z_{1i} = X_{(i+1)} - X_{(i)}$ for $i = 1, \dots, 4$. In this instance, $z_{11} = 123 - 121 = 2$, $z_{12} = 3$, $z_{13} = 2.5$, $z_{14} = 0.5$.

The original value of Aly's test statistic is $8 + 18 + 15 + 2 = 38$. To compute this test statistic for other arrangements of the labels on the observations, we also need to know the differences $z_{2i} = Y_{(i+1)} - Y_{(i)}$ for the second sample; $z_{21}=1$, $z_{22}=1$, $z_{23}=1$, $z_{24}=2$.

Only certain exchanges are possible. Rearrangements are formed by first choosing either z_{11} or z_{21} , next either z_{12} or z_{22} , and so forth until we have a set of four differences.

One possible rearrangement is $\{2, 1, 1, 2\}$ which yields a value of $S_A = 20$. There are $2^4 = 16$ rearrangements in all, of which only one $\{2, 3, 2.5, 2\}$ yields a more extreme value of the test statistic than our original observations. With two out of 16 rearrangements yielding values of the statistic as or more extreme than the original, we should accept the null hypothesis. (Better still, given the limited number of possible rearrangements, we should gather more data before we make a decision.)

But the test we've described is restricted to two equal-sized samples and with missing data nearly inevitable may not always be applicable.

If our second sample is larger than the first, we may still resample in two stages: First, we select a subset of m values $\{Y_i^*, I=1, \dots, m\}$ without replacement from the n observations in the second sample, and compute the order statistics $Y_{(1)}^* < Y_{(2)}^* < \dots < Y_{(m)}^*$ and their differences $\{z_{2i}^*\}$. Next, we examine all possible values of Aly's measure of dispersion for permutations of the combined sample $\{\{z_{1i}^*\}, \{z_{2i}^*\}\}$ and compare Aly's measure for the original observations with this distribution. Repeating the two steps for several hundred random subsets we obtain a bootstrap confidence interval for the p-value.

3.2. Deviations About The Median

Good [1994] proposed a permutation test based on the sum of the absolute values of the deviations. First, we compute the median for

each sample; next, we replace each of the remaining observations by the square of its deviation about its sample median; last, in contrast to the test proposed by Brown and Forsythe [1974], we discard the redundant linearly-dependent value from each sample.

Suppose the first sample contains the observations x_{11}, \dots, x_{1n_1} whose median is M_1 ; we begin by forming the deviates $x'_{1j} = |x_{1j} - M_1|$ for $j = 1, \dots, n_1$. Similarly, we form the set of deviates $\{x'_{2j}\}$ using the observations in the second sample and their median.

If there are an odd number of observations in the sample, then one of these deviates must be zero. We can't get any information out of a zero, so we throw it away. In the event of ties, should there be more than one zero, we still throw only one away. If there is an even number of observations in the sample, then two of these deviates (the two smallest ones) must be equal. We can't get any information out of the second one that we didn't already get from the first, so we throw it away.

Our new test statistic S_G is the sum of the remaining $n_1 - 1$ deviations in the first sample, that is,

$$S_G = \sum_{j=1}^{n_1-1} x'_{1j}.$$

We obtain the permutation distribution for S_G and the cut-off point for the test by considering all possible rearrangements of the remaining deviations between the first and second samples.

To illustrate the application of this method, suppose the first sample consists of the measurements 121, 123, 126, 128.5, 129.1 and the second sample of the measurements 153, 154, 155, 156, 158.

Thus, after eliminating the zero value, $x'_{11}=5$, $x'_{12}=3$, $x'_{13}=2.5$, $x'_{14}=3.1$, and $S_G = 13.6$. For the second sample $x'_{21}=2$, $x'_{22}=1$, $x'_{23}=1$, $x'_{24}=3$.

There are $\binom{8}{4}$ arrangements in all of which only three yield values of the test statistic as or more extreme than our original value. $3/70=0.043$ and we conclude that the difference between the dispersions of the two manufacturing processes is statistically significant at the 5% level.

As there is still a weak dependency among the remaining deviates within each sample, they are only asymptotically exchangeable. Tests based on S_G are alternately conservative and liberal according to Baker [1995] in part because of the discrete nature of the permutation distribution unless

- a. The ratio of the sample sizes n , m is close to 1;
- b. The only other difference between the two populations from which the samples are drawn is that they might have different means, that is, $F_2[x] = F_1[(x-\delta)/\sigma]$.

We were unable to confirm her results in our own simulations (the R code for which may be obtained from the author at pigood@verizon.net). We found tests based on S_G to be uniformly conservative for samples of sizes 4 and above. Our simulations employed either normally-distributed data or mixed normal data. To avoid randomizing on the boundary in our simulations, we set the alpha level in each instance to correspond to one of the discrete levels

available for the permutation distribution. Chernick and Liu [2002] describe the necessity of such a procedure.

3.3. Paired Deviations

When we have the same number of observations in each sample, an alternate method of rearrangement suggests itself. Suppose we pair the deviations according to their magnitude within each sample, form a rearrangement by selecting one member of each pair, and again compute the sum.

For example, using the measurement data, we would form the pairs (5,3), (3.1,2), (3,1) and (2.5,1). Now there are just 16 possible rearrangements with the sum of the observations as originally labeled being the largest possible value. We obtain a p-value of 0.0625, not because this method is less powerful than the preceding one, but because we have severely restricted the number of possible rearrangements. With two samples of size n , there are only $n-1$ pairs and 2^{n-1} possible rearrangements.

If we draw six observations from a $N(0,1)$ population and six from a $N(0,2)$ population, and test at the $2/32=0.0625$ significance level, the first method using deviations about the sample medians has a power of 33% and the second of 60%. The first test is conservative, with an actual p-value of less than 6%. The second test is exact. Aly's permutation method, which is also exact for this p-value, has a power

of only 24%. Similar results for the three tests were obtained in a comparison of Gamma(4) distributions with two samples of size 6.

Of course, if we insist on using a significance level of 5%, then either we must sacrifice power due to the discrete nature of the permutation distribution, or worse, leave decision making on the boundary up to a chance device (see, for example, Lehmann, 1986, p75).

4. K-SAMPLE PERMUTATION TEST

The preceding tests based upon the absolute deviations about the sample medians are easily generalized to the case of K-samples from K-populations. Such a test would be of value, for example, as a test for homoscedacity as a preliminary to a k-sample analysis for a difference in means among test groups.

First, we create K sets of deviations about the sample medians and make use of the test statistic

$$S = \sum_{k=1}^K (\sum_{j=1}^{n_k-1} x'_{kj})^2$$

The choice of the square of the inner sum ensures that this statistic takes its largest value when the largest deviations are all together in one sample after relabeling.

To generate the permutation distribution of S, we again have two choices. We may consider all possible rearrangements of the sample

labels over the K sets of deviations. Or, if the samples are equal in size, we may first order the deviations within each sample, group them according to rank, and then rearrange the labels within each ranking.

Again, this latter method is directly applicable only if the K samples are equal in size, and, again, this is unlikely to occur in practice. We will have to determine a confidence interval for the p-value for the second method via a bootstrap in which we first select samples from samples (without replacement) so that all samples are equal in size. While I wouldn't recommend doing this test by hand, once programmed, it still takes less than a second on last year's desktop.

5. TESTING WHEN NON-RESPONDERS ARE PRESENT

In testing for a response to drug treatment, it is common to encounter a response threshold peculiar to each individual, such that some individuals respond to drug treatment and some do not. If the treatment is effective, one expects both the mean and the variance of the treated population to be larger than those of the control population. This suggests that a test for simultaneous changes in expectation and variance would be more powerful than one that tests for changes in expectation alone.

We wish to test the hypothesis $H: F_2[x]=F_1[x]$ against the alternative $K: F_2[x]=pF_1[x] + (1-p) F_1[(x-d)/s]; \quad 0 < p < 1; d>0 ; s \geq 1.$

Good[1979] proposed the test statistic

$$v = u(\bar{X}_T - \bar{X}_C)^2 + (1-u)S_T^2$$

where the first term is proportional to the difference in means of the two samples and the second to the variance of the treatment sample. Rearranging the labels between the two sets of observations generates its permutation distribution. But, alas, its power is only marginally better than the t-test.

We suggest instead the test statistic $T = u(\bar{X}_\pi - \bar{X}_O) + (1-u)(S_\pi - S_O)$ where S is the sum of the deviations about the median as defined in Section 3, and the subscripts O and π refer to the original data and the data after rearranging sample labels.

Care must be taken in generating the rearrangements as the first part of our test statistic is based on one more value than the second. To accomplish the desired result, we first select $n - 1$ observations at random from the reduced data set to use in forming S and then one more observation (of those not already selected) to calculate the mean.

Ideally, the parameter u would be chosen equal to p , but typically, p is not known. In our simulations, we used a value of $u=0.67$ and selected data from an $N(1,1)$ population for controls and from a mixture of 50% $N(1,1)$ and 50% $N(2,2)$ for the treated group. At a significance level of 10%, and using two samples of size 5 respectively, the t-test yielded power of 20%, a permutation test using Student's t as its test statistic had a power of 21%, and a permutation test that made use of our new statistic had a power of 33%.

REFERENCES

- Aly E-E AA. (1990) Simple tests for dispersive ordering. *Stat. Prob. Ltr.* 9: 323–325.
- Baker RD. (1995) Two permutation tests of equality of variance. *Statist. Comput.* 5(4): 289–96.
- Balakrishnan N; Ma CW. (1990) A comparative study of various tests for the equality of two population variances. *Statist. Comp. Simul.* 35: 41-89.
- Box GEP; Anderson SL.(1955) Permutation theory in the development of robust criteria and the study of departures from assumptions. *J. Roy. Statist. Soc B.* 17: 1-34 (with discussion).
- Brown MB; Forsythe AB. (1974) Robust tests for equality of variances. *JASA.* 69: 364-367.
- Chernick MR; Liu CY. (2002) The saw-toothed behavior of power versus sample size and software solutions: single binomial proportion using exact methods. *Amer. Statist.* 56:149-155.
- Conover WJ; Johnson ME; Johnson MM. (1981) Comparative study of tests for homogeneity of variances: with applications to the outer continental shelf bidding data. *Technometrics.* 23: 351-361.
- Conover WJ; Salsburg D. (1988) Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to "respond" to treatment. *Biometrics.* 44: 189-196.
- Duran BS. (1976) A survey of nonparametric tests for scale. *Commun. Statist. Theor-Meth.* A5:338-370.
- Fisher RA. (1925) *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd; 1st ed.

- Good PI. (1979) Detection of a treatment effect when not all experimental subjects respond to treatment. *Biometrics*. 35:483-489.
- Good PI. (1994) *Permutation Tests*. New York: Springer-Verlag. 1st ed.
- Hayes AF. (1997) Cautions in testing variance equality with randomization tests. *J. Statist. Compu. Simul.* 59:25-31.
- Lehmann EL. (1986) *Testing Statistical Hypotheses*. 2nd ed. New York: John Wiley and Sons.
- Maritz JS. (1996) *Distribution Free Statistical Methods*. 2nd ed. London: Chapman and Hall.
- Miller RG. (1968) Jackknifing variances. *Annals Math. Statist.* 39: 567-582.
- Sukhatme BV. (1958) A two sample distribution free test for comparing variances: *Biometrika*. 45: 544-8.
- Welch BL. (1937) On the z-test in randomized blocks and Latin squares. *Biometrika*. 29: 21-52.