Making the Best of Limited Global Data

C. K. Chauhan - Indiana Purdue University
Y. M. Zubovic - Indiana Purdue University

Abstract:
This paper explores a method of conducting a statistical test to compare the mean values of two or more populations when complete information over a given period of time is unavailable for all the populations of interest. For example, the real wages of computer programmers may be available for both USA and Korea for a certain time period. However, the real wages for Korea alone may be available for an additional time period, and the real wages for USA alone may be available for another time period. Our proposed statistical test utilizes all the available data to test for the equality of the two means.

Introduction:
Conducting a statistical analysis on real data is exciting. However, finding complete data for different populations can sometimes be more challenging and time consuming than the actual analysis. Consider the following scenario when comparing a value such as the unemployment rate or the wages of two countries:

1980–1995: Data available for both countries
1996- 2000: Data available for country 1
2001-2006 : Data available for country 2

In another example, suppose in a group of patients, some patients have a certain type of symptom in both eyes and some have it in either the left or the right eye. To compare the performances of two types of eye medication, we have complete data for some patients (who use one treatment in one eye and the other treatment in the other eye) and incomplete data for other patients.

If complete information is available for both populations, a simple paired data t test can be used. However, the situation discussed above demands a careful statistical analysis. One easy but inefficient solution to analyze above data is to consider the data from 1980 to 1995 and ignore the rest of the data (or equivalently, only consider the patients who have symptoms in both eyes). In this approach, one may be throwing away significant amount of useful data. Another method may include estimating the unknown information. For example, the unavailable data for a particular country may be estimated from the available data for that country. This may be complicated as the estimated values may depend on the choice of the model and the predictors. Moreover, the error of estimation may reduce the power of the test.

Our proposed method neither ignores the available data nor estimates the unavailable data. The proposed test is simple to apply and it has a reasonably high power.

Notation: Consider two populations of interest.
Let (X_i, Y_i) be the paired data for the i^{th} year (or subject), where X is the value of interest from

the first population and Y from the second.
Let n= number of years for which information, (X_i, Y_i), is available for both populations.
m_1= number of years for which information, Xj, is available from population1.
m_2= number of years for which information, Yj, is available from population 2.

Objective: use all $n + m_1 + m_2$ data to test the following hypothesis:

Ho: $\mu_x = \mu_y$ versus H_1: $\mu_x > \mu_y$, where μ represents the true mean of a population.
(Note: the proposed test applies to H_1: $\mu_x < \mu_y$ as well).

Statistical Assumption:
(X, Y) ~ Normal (μ_x, σ_y, μ_x, σ_y, ϱ), where μ, σ represent the mean and the standard deviation and ϱ represents the correlation coefficient between X and Y.

Statistical Background:

1) Let T_1 be the usual paired t test on the n complete pairs, Di= Xi-Yi, where

$T_1 = \sqrt{n}.\ D/Sd$, where D and Sd are the mean and standard deviations of Di.

2) Consider another test statistic, T_2, which is based on $n + m_1 + m_2$ data. This test statistic depends on the means, standard deviations, correlation coefficient of the complete n pairs, and the means and the standard deviations of incomplete data sets. The formula of T_2 is available in the Appendix.

3) For each of two tests, T_1 and T_2, calculate the corresponding p values, p_1 and p_2. Recall a p value is also known as the observed significance level of a test. A null hypothesis is rejected if the p value is small, say less than 5%.

Prior Work and their Limitations:

Several authors have studied the problem of combining either the two test statistics, or the two p values. A test based on both T_1 and T_2 (or both p_1 and p_2) performs better than a test based on only one test statistic (or only one p value). A number of different criterion have been considered to combine T_1 and T_2. These criterion include, but are not limited to, combining T_1 and T_2 in proportion of the sample sizes or in the inverse proportion of the standard deviations of T_1 and T_2. One such test is proposed by Bhoj. Bhoj uses a complicated method to transform T_1 and T_2 to obtain two variables (U_1 and U_2) having approximately normal distributions. The resultant test, Zb, is a linear combination of U_1 and U_2. Fisher, on the other hand, proposed combining the p values. Bhoj's test is tedious and the distribution of the test is approximate. Fisher's test does not always perform well.

The Proposed Test

We propose a test which is easy to apply and also performs well. The proposed test is explained as follows:

1) Let Ti and p_i be defined as above, for i=1, 2.

2) For each p i ($0<p_i<1$), find the corresponding Z score, Zi, such that $Pr(Z>Zi)= p_i$.
For example, if $p_i = .03$, then Z i=1.88.

3) We now combine Z_1 and Z_2 to form one Z value. Below are two ways to combine the Z values:

$$Z.s= (Z_1 +Z_2)\sum\sum 2 \qquad Z.w = (w_1 Z_1 + w_2 Z_2)\sum\sum(w_1^2 +w_2^2)$$

Note that the particular choice of the divisor in each case ensures that the resultant Z score has a mean 0 and standard deviation 1. Here, w_i are the weights assigned to each Zi value. The weights may either be in proportion to the sample sizes or in proportion to the standard deviations of the test statistics. For ease of computation, let the weights be equal to the corresponding sample sizes. Thus,

$$Zw = (n_1 Z_1 +n_2 Z_2)\sum\sum(n_1^2 +n_2^2).$$

4) It can be shown that if two population means are equal, Zw and Zs, each has a standard normal distribution. For $H_1: \mu_x > \mu_y$, the null hypothesis is rejected if Z.s (or Z.w) is significantly large. For $H_1: \mu_x < \mu_y$, the null hypothesis is rejected if Z.s (or Z.w) is significantly small.

Example: The following real life data resulted from a study conducted to compare the average performance of two types of eye medication, known as B-G and RK. Forty patients participated in the study. The values represent a score given to a patient after he or she had applied the medication for a certain period of time. Twenty patients provided data for both types of medications, 10 additional patients provided data for only B-G, while another 10 patients provided data for only RK. Although the data was collected from a health related study, this example serves a model for a variety of situations in which two means are compared from limited data.

Both eyes treated:

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13
B-G	4	69	87	35	39	79	31	79	65	95	68	62	70
RK	62	80	82	83	0	81	28	69	48	90	63	77	0

Patient	14	15	16	17	18	19	20
B-G	80	84	79	66	75	59	77
RK	55	83	85	54	72	58	68

One Eye Treated:

B-G	36	86	39	85	74	72	69	85	85	72
RK	88	83	78	30	58	45	78	64	87	65

Analysis:

Since the data are not normally distributed, Y= ln(100-X) transformation was deemed suitable. The rest of the calculations are based on the transformed data.

a) Based on the 20 pairs, T_1= .641 with p_1 = .264, Z_1= .63

b) Based on all the data, T_2= .294 with p_2 =.385, Z_2 = .291

(Note: Large p values are indicative of the lack of evidence to reject H_0)

c) Z_s = .65 with p value .257

d) Z_w = .547 with p value .292

Conclusion: Based on both proposed test statistics, we conclude that there is not sufficient evidence to conclude that the two treatment means are different, since each p value is greater than 5%. If both p values provide conflicting conclusions, one may decide to either take additional data, or select either Zs or Zw for inference. Our simulation study indicates that Zs is more likely to perform better than Zw. The detailed discussion on this follows in a later section.

Comparing our proposed test with other existing methods

In this paper, we compare our tests with two other tests; one proposed by Fisher, and Zb, proposed by Bhoj. Several simulation studies were conducted and the corresponding graphs were plotted. In this paper only two graphs are included. However, a detailed discussion of the comparison is given in the summary. Consider the following graphs:

In graph, the level of significance (Probability of Type 1 error) was calculated for Zb and for Zs, one of our proposed statistics. The horizontal axis represents various values of the correlation coefficient, rho,between the paired data. The graph was plotted for three different groups of data: (1) $\mu_x = \mu_y$, (2) $\mu_x < \mu_y$ and (3) $\mu_x, > \mu_y$.

The graph indicates that both tests attain the level of significance close to 5% when the standard deviations are equal. Our proposed test performs slightly better than Zb in terms of attaining 5% level of significance for each group.

In graph 2, the test Zs is compared with the test proposed by Fisher. In this graph, the powers of both the tests were calculated under different alternatives. Recall the alternative hypothesis is that $\mu_x > \mu_y$. Thus, on the horizontal axis, some selected values of $\mu_x - \mu_y$ are plotted. These values are labeled as delta. The values on the horizontal axis represent the power, in percent, of the corresponding tests. Since the power is a function of correlation coefficient, rho, two values of rho, -.5 and .1 are selected. The graph clearly indicates that for both values of rho, the power of Zs is higher than that of Fisher's test.

Interesting observations

The following observations are based on the simulation study conducted for several different values of the sample sizes, correlation coefficients, and standard deviations. The power of the tests and the

empirical level of the probability of type 1 error were calculated. The following observations were made.

1) **Regarding attaining the level of significance:**
Consider sample sizes $n = m_1 + m_2$ where $m_1 = m_2$. In terms of attaining the level of significance, Fisher's test, Zs, and Zw performed similar to Zb. No apparent relationship was found between the level of significance and the strength or the direction of the correlation coefficient.

For unequal sample sizes, the level of significance was sensitive to deviation from $\mu_X = \mu_Y$ for Zb, Fisher's test , Zs, and Zw (worse if smaller sample had larger variance). The level was related to correlation coefficient, both in direction and strength. Fisher's test and Zs performed better than Zb while all three performed better than Zw.

2) **Regarding Power of the test:**
Zs, Fisher's test, and Zb had higher power than Zw, while Zs tended to perform better than Fisher's test. For a given value of $\mu = \mu_x - \mu_y$, the power increased with correlation coefficient for all tests. The gain in power for $\alpha = 0.4$ to 0.8 tended to be higher than the gain for $\alpha = 0.8$ to 1.2. For all tests, the power was higher for equal variance case than with the unequal variance.

Summary:
Our proposed test to analyze an incomplete data is based on the linear combination of the z transformed p values associated with two test statistics s. The proposed test statistic has a normal distribution. The simulation study shows that the test performs as well as or better than some existing tests.

Appendix: Define the following:

$$a_{11} = \sum_{i=} \qquad \sum_{=} \qquad \sum_{=}$$

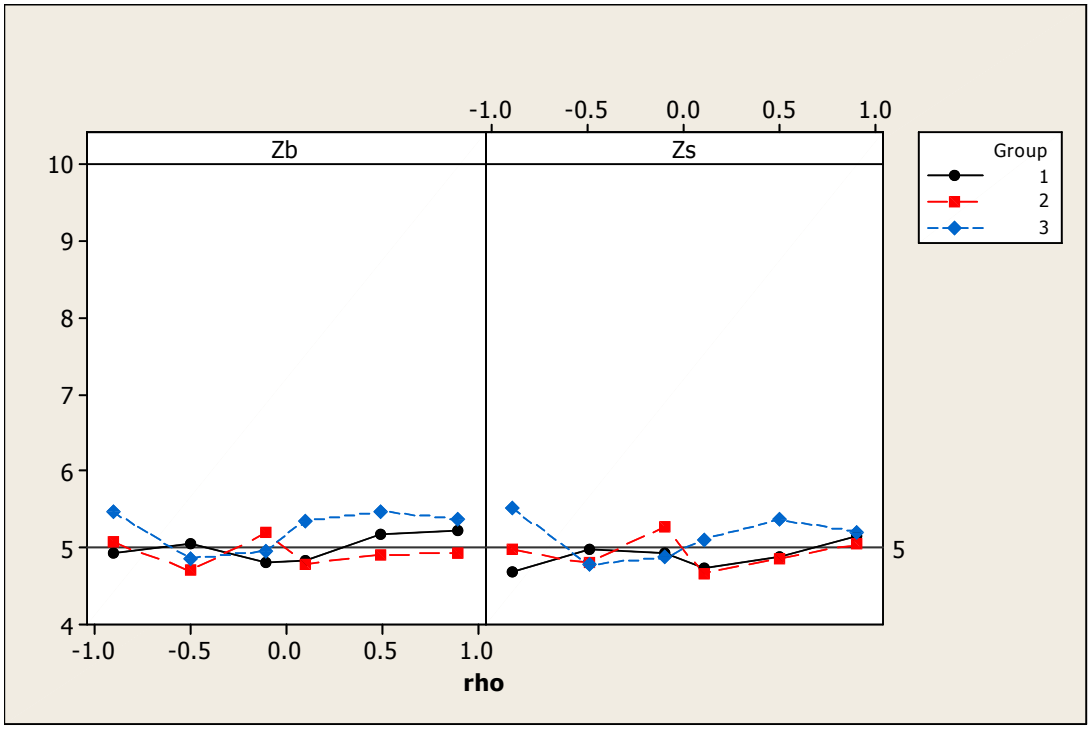
$$b_1 = \sum_{j= +}^{\cdots} \qquad \sum_{= +}^{\cdots}$$

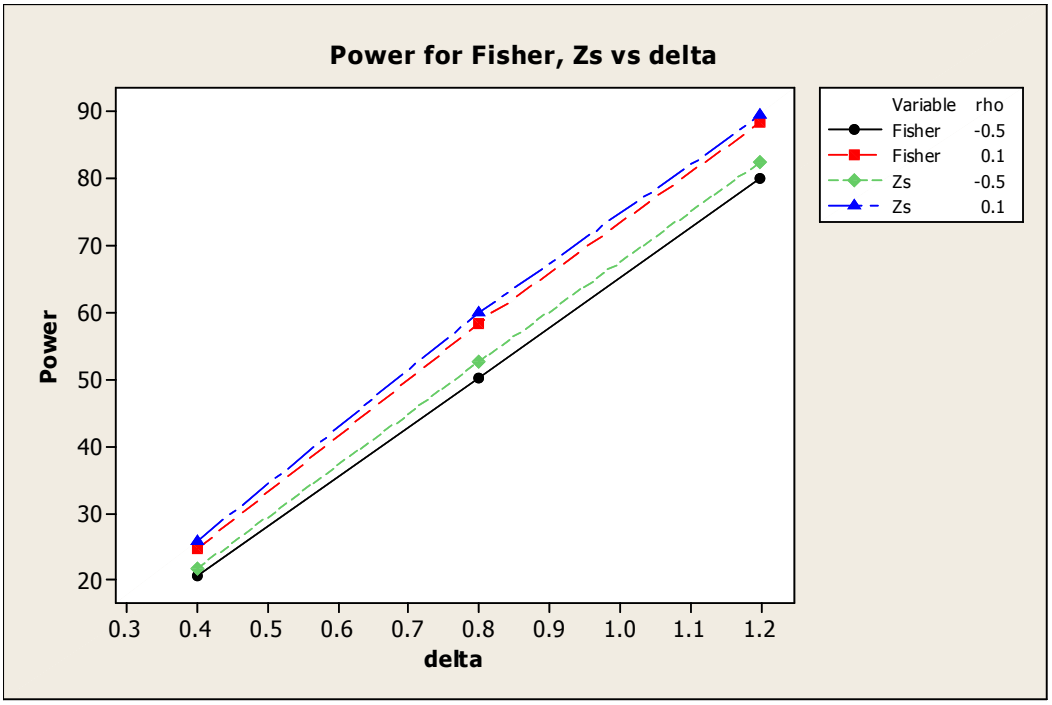
$$\omega = \frac{m \left[\qquad \right]}{n \;\; m_1 + \quad + \quad +}$$

The test statistic defined by Bhoj(1989) is as follows:

$$T_2 = \frac{\omega \quad - \quad - \quad + \quad - \qquad + \quad -}{\sqrt{\left[\quad + \quad + \quad + \quad + \quad + \qquad - \qquad \right]}}$$

Graphs


Zb
Zs
-1.0
-0.5
0.0
0.5
1.0
10
9
8
7
6
5
4
5
rho
Group
1
2
3


Power for Fisher, Zs vs delta
Power
delta
90
80
70
60
50
40
30
20
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
Variable rho
Fisher -0.5
Fisher 0.1
Zs -0.5
Zs 0.1

e

References

Bhoj, D.S., 1978: Testing equality of means of correlated variates with missing data on both responses. Biometrika, 65, 225-228.
Bhoj, D.S.,1989: On comparing correlated means in the presence of incomplete data. Biom. Journal, 81, 279-288.