The Korean Communications
in Statistics Vol. 6, No. 3, 1999
pp. 659-666

An Influence Measure in Comparing Two Population Means[1)]

Whasoo Bae[2)]

Abstract

In comparing two population means, the test statistic depends on the sample means and the variances, which are very sensitive to the extremely large or small values. This paper aims at examining the behavior of such observations using proper criterion which can measure the influence of them. We derive a computationally feasible statistic which can detect influential observations on the two-sample t-statistic.

1. Introduction

It is known that some extremely large or small valued observations compared to the other values may give so much influence that they can change the considered statistical model and the conclusion of test. However, most works have been done in the area of regression analysis. Cook(1977) detected the influential data point using Cook's distance in linear regression analysis, and good references in this area are Belsley, Kuh and Welsch(1980), Cook and Weisberg(1982), and Chatterjee and Hadi(1986). Pregibon(1981) suggested one-step estimator in the logistic regression diagnostics. In Box-Cox transformation model, Cook and Wang(1983), Hinkley and Wang(1988), Tsai and Wu(1990), Kim, Storer, and Jeong(1996) studied the influence on the transformation parameter. Regression diagnostics in nonparametric regression models are studied by Eubank(1985), Silverman(1985), Thomas(1991), and Kim(1996). Also, in the area of multivariate analysis, such as the discriminant analysis, Campbell(1978) and Fung(1995) studied identification of influential observations and suggested some basic building blocks.

In testing problem, the testing result also will be changed by one or few influential data points. But it is very rare to find a work about the influence of the point in testing mean. In testing population mean, the testing statistic is based on the sample mean which is very sensitive to the extreme values. If some observations are extremely large or small compared to the other values, the testing result might be changed by these influential points. Hence it might be interesting to see how these observations give effect to testing result.

---

1) This work was supported by the Inje Scholarship Foundation in 1997.

2) Assistant Professor, Department of Applied Statistics, Inje University, Kimhae, 621-749, KOREA

In section 2 a statistic to measure the influence of cases on the two-sample t-statistic is given and section 3 gives a behavior of p-value of test statistic by infinitesimal perturbation of observation. An example using an artificial data set with the masking effect is in section 4.

2. An Influence Statistic

Let $X_1, X_2, \cdots, X_m$ be random sample from $N(\mu_1, \sigma^2)$ and $Y_1, Y_2, \cdots, Y_n$ be random sample from $N(\mu_2, \sigma^2)$. Also, assume that these two samples are independent. Note that these are the usual set-up in two sample comparison problem. When we test $H_0 : \mu_1 = \mu_2$ the usual two-sample t-statistic is

$$t = \frac{\overline{X} - \overline{Y}}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

and it follows the t-distribution with $m+n-2$ degree of freedom, where $\overline{X}$ and $\overline{Y}$ are sample means and s_p is the pooled sample standard deviation.

If some observations are considered to be very extreme and influential to testing result, then we can check the influence of these data by comparing the value of testing statistic with them in and the one without them. That is, the influence of such observations can be measured by investigating the difference between these two values of test statistic. This approach is so called the deletion method. The well known Cook's distance belongs to this approach.

Let $K=\{i_1, i_2, \cdots, i_k\}$ and $L=\{j_1, j_2, \cdots, j_l\}$ be two index sets and we will delete k and l cases from m and n observations, respectively. Also let $\overline{X}_{(K)}$ and $\overline{Y}_{(L)}$ be the sample means based on $m-k$ and $n-l$ observations after deleting k and l data points and $s_{p(K,L)}$ be the corresponding pooled sample standard deviation. Then the test statistic without these data points is as follows ;

$$t_{(K,L)} = \frac{\overline{X}_{(K)} - \overline{Y}_{(L)}}{s_{p(K,L)} \sqrt{\frac{1}{m-k} + \frac{1}{n-l}}}$$

$$= q_{(K,L)} \{ t - r_{(K,L)} \}, \qquad (1)$$

where

$$q_{(K,L)} = \frac{\sqrt{(m-k)(n-l)(m+n)(m+n-k-l-2)}}{\sqrt{mn(m+n-2)(m+n-k-l)}} \frac{1}{\sqrt{1-\frac{d_{ij}}{(m+n-2)\,s_p^2}}} \tag{2}$$

with

$$d_{ij} = \sum_{i\in K}\delta x_i^2 + \frac{1}{m-k}\left(\sum_{i\in K}\delta x_i\right)^2 + \sum_{j\in L}\delta y_j^2 + \frac{1}{n-l}\left(\sum_{j\in L}\delta y_j\right)^2 \tag{3}$$

and

$$r_{(K,L)} = \frac{\frac{1}{(m-k)}\sum_{i\in K}\delta x_i - \frac{1}{(n-l)}\sum_{j\in L}\delta y_j}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}} . \tag{4}$$

where

$$\delta x_i = X_i - \overline{X}\,, \qquad \delta y_j = Y_j - \overline{Y} \tag{5}$$

See the Appendix for the proof. Note that Eq.(1) is very useful and computationally feasible because $t_{(K,L)}$ is expressed as the original test statistic t. Therefore, if $t_{(K,L)}$ is very different from t, then cases in K and L can be regarded as influential. To make the difference statistic have better interpretation, we can express it as a relative change, i.e.,

$$rt_{(K,L)} = \left|\frac{t - t_{(K,L)}}{t}\right| = \left|1 - \frac{t_{(K,L)}}{t}\right|$$

So, if $rt_{(K,L)}$ is away from 0 cases in K and L can be regarded as influential on the test statistic. However, it seems to be very difficult to find the reference distribution for $|t - t_{(K,L)}|$.

3. Numerical Perturbation

To see how one or few cases affect the test statistic we generate an artificial data set A given in Table 1. Based on these data the p-value of the t-statistic is 0.0515. Now we add one case $30+k$ to sample X and varies k from 0 to 28, and the resulting p-value for each k is given in Figure 1. Note that the added case is between minimum and maximum of the sample X, and the p-value changes from 0.125 to 0.038. i.e., one case can change the result of test. Also, we note that even though a case is not away from the sample mean it can be influential observation on the test statistic. For example, in data A, if 50 is added to sample

X, then the null hypothesis is rejected at the level of significance 0.05. Therefore we must be cautious when detecting influential observations in this problem.

Table 1. Artificial data A

X	30	35	38	41	44	44	46	46	47	49	53	56	58
Y	29	33	34	38	38	39	40	40	41	42	47	51	

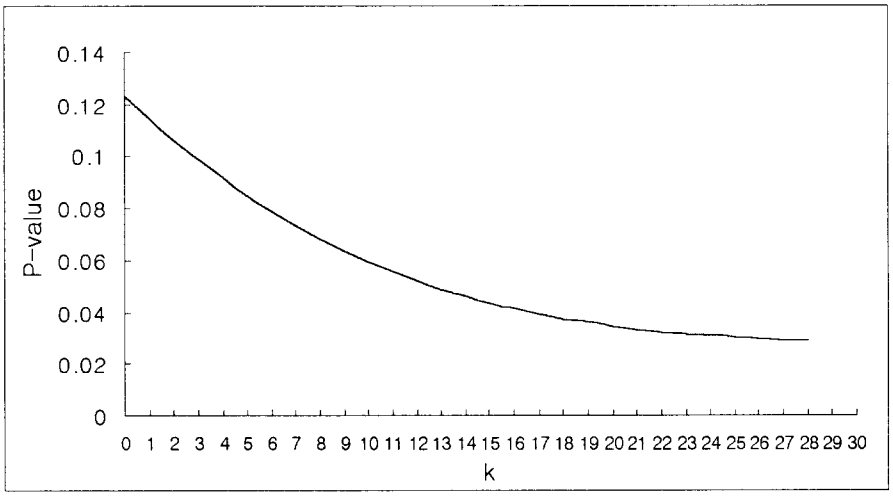

Figure 1. Plot of p-values vs k.
One case $30+k$, $k=0, 1, \cdots, 28$ is added to sample X.

One can guess that if δx_i or δy_j is large then the statistic would change a lot, however, it is not always true. For the data set B in Table 2, we make the case 58 (maximum of sample X) larger and larger, i.e., $58+k$, $k=1,2,3,\ldots$. Then, one might expect that the resulting p-value of the test statistic would be smaller and smaller. However, as shown in Figure 2, p-value becomes smaller up to some k (about 30) and remains almost constant for $k>30$. If there is one extreme outlier it affects not only to the sample mean but also to the pooled sample variance. Therefore, an outlier is not necessarily an influential case, and we should be cautious in interpreting an outlier.

Table 2. Artificial data B

X	30	35	38	41	44	44	46	46	47	49	53	56	58
Y	29	32	35	39	39	40	41	42	47	51	55	57	

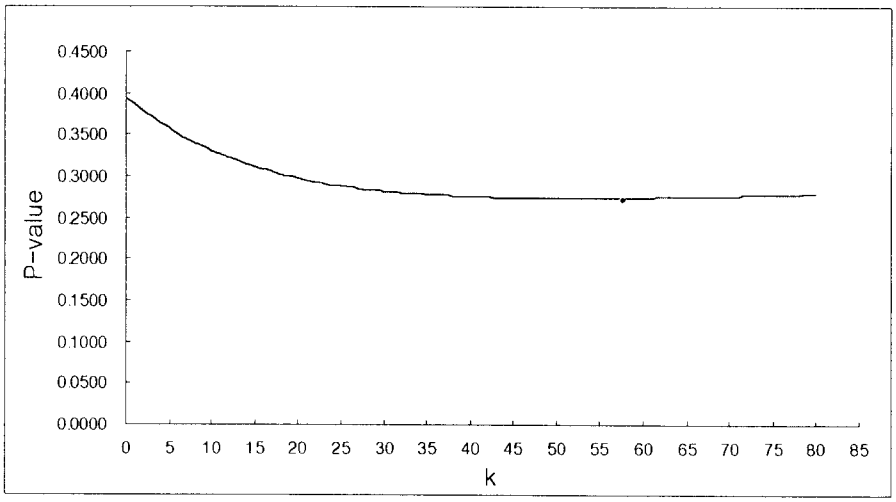

Figure 2. Plot of p-values vs k.
Case 58 is replaced by $58+k$, $k=0, 1, \cdots, 80$.

4. Example for the Masking Effect

When a set of observations (i, j) is influential and they are not individually influential, we say the set (i, j) has a masking effect. This is why the single case deletion is not enough and multiple case deletion is necessary.

As an illustrative example for the masking effect, use an artificial data set C given in Table 3. For the data set C, p-value is 0.015, i.e., the null hypothesis is rejected at the level of significance 0.05. If we delete case 58 from sample X, the resulting p-value is 0.029, and if we delete case 29 from sample Y, the resulting p-value is 0.028. Therefore, deletion of one case from either sample does not change the result of test. But, if we delete both cases (case 58 from sample X and case 29 from sample Y), the resulting p-value is 0.052, i.e., the null hypothesis is not rejected.

Table 3. Artificial data C

X	30	35	38	41	44	44	46	46	47	49	53	56	58	58
Y	29	29	33	34	38	38	39	40	40	41	42	47	51	

Appendix

First, we note that

$$\begin{aligned}\bar{X}-\bar{X}_{(K)} &= \bar{X}-\frac{1}{m-k}\sum_{i\notin K}X_i \\ &= \bar{X}-\frac{1}{m-k}(m\bar{X}-\sum_{i\in K}X_i) \\ &= \frac{1}{m-k}\sum_{i\in K}(X_i-\bar{X}) \qquad (A.1)\end{aligned}$$

i.e., $\bar{X}-\bar{X}_{(K)}=\sum_{i\in K}\delta x_i/(m-k)$ and similarly, $\bar{Y}-\bar{Y}_{(L)}=\sum_{j\in L}\delta y_i/(n-l)$, where δx_i and δy_i are defined in Eq.(5). Let s_X^2 , s_Y^2 be sample variances based on m and n observations, respectively. Also, let $s_{X(K)}^2$, $s_{Y(L)}^2$ be sample variances based on $m-k$ and $n-l$ observations, respectively. Then, by $(A.1)$, can easily show that

$$\begin{aligned}(m-k-1)s_{X(K)}^2 &= \sum_{i\notin K}(X_i-\bar{X}_{(K)})^2 \\ &= \sum_{i=1}^{m}(X_i-\bar{X}_{(K)})^2-\sum_{i\in K}(X_i-\bar{X}_{(K)})^2 \\ &= \sum_{i=1}^{m}(X_i-\bar{X}+\bar{X}-\bar{X}_{(K)})^2-\sum_{i\in K}(X_i-\bar{X}+\bar{X}-\bar{X}_{(K)})^2 \\ &= (m-1)s_X^2+(m-k)(\bar{X}-\bar{X}_{(K)})^2-\sum_{i\in K}(X_i-\bar{X})^2 \\ &\quad -2(m-k)(\bar{X}-\bar{X}_{(K)})^2 \\ &= (m-1)s_X^2-(m-k)(\bar{X}-\bar{X}_{(K)})^2-\sum_{i\in K}(X_i-\bar{X})^2 \\ &= (m-1)s_X^2-\sum_{i\in K}\delta x_i^2-\frac{1}{m-k}(\sum_{i\in K}\delta x_i)^2 \qquad (A.2)\end{aligned}$$

and similarly

$$(n-l-1)s_{Y(L)}^2=(n-1)s_Y^2-\sum_{j\in L}\delta y_i^2-\frac{1}{n-l}(\sum_{j\in L}\delta y_i)^2 \ . \qquad (A.3)$$

Therefore, by $(A.2)$ and $(A.3)$,

$$\begin{aligned}(m+n-k-l-2)s_{p(K,L)}^2 &= (m-k-1)s_{X(K)}^2+(n-l-1)s_{Y(L)}^2 \\ &= (m-1)s_X^2+(n-1)s_Y^2 \\ &\quad -\left\{\sum\delta x_i^2+\frac{1}{m-k}(\sum\delta x_i)^2+\sum\delta y_i^2+\frac{1}{n-l}(\sum\delta y_i)^2\right\} \\ &= (m+n-2)s_p^2-d_{ij} \qquad (A.4)\end{aligned}$$

, where d_{ij} is given in Eq.(3). Finally by $(A.4)$, we have

$$t_{(K,L)} = \frac{\overline{X}_{(K)} - \overline{Y}_{(K)}}{s_{p(K,L)}\sqrt{\frac{1}{m-k}+\frac{1}{n-l}}}$$

$$= \frac{\overline{X} - \overline{Y} - \left(\frac{1}{m-k}\sum \delta x_i - \frac{1}{n-l}\sum \delta y_i\right)}{s_p\sqrt{\frac{1}{m+n-k-l-2}}\sqrt{(m+n-2)-\frac{d_{ij}}{s_p^2}}\sqrt{\frac{1}{m-k}+\frac{1}{n-l}}}$$

$$= \frac{\overline{X} - \overline{Y} - \left(\frac{1}{m-k}\sum \delta x_i - \frac{1}{n-l}\sum \delta y_i\right)}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}\sqrt{\frac{m+n-2}{m+n-k-l-2}}\sqrt{\frac{m+n-k-l}{(m-k)(n-l)}}\sqrt{\frac{mn}{m+n}}\sqrt{1-\frac{d_{ij}}{(m+n-2)s_p^2}}}$$

and we get Eq.(1)

References

[1] Belsley, D. A., Kuh, E. and Welsch, R. E. (1980) *Regression Diagnostics : Identifying Influential Data and Source of Collinearity*, Wiley, New York.

[2] Campbell, N. A. (1978) The influence as an aid in outlier detection in discriminant analysis, *Applied Statistics*, 27, 251-258

[3] Chatterjee, S. and Hadi, A. S. (1986) Influential observations, high leverage points, and outliers in linear regression (with discussion), *Statistical Science*, 1, 379-416.

[4] Cook, R. D. (1977) Detection of influential observations in linear regression, *Technometrics*, 22, 494-508.

[5] Cook, R. D. and Wang, P. C. (1983) Transformation and influential cases in regression, *Technometrics*, 25, 337-343.

[6] Cook, R. D. and Weisberg, S. (1982) *Residuals and Influence in Regression*, Chapman and Hall, New York.

[7] Eubank, R. L. (1985) Diagnostics for smoothing splines, *Journal of the Royal Statistical Society*, Ser. B, 47, 332-341.

[8] Fung, W. K. (1995) Diagnostics in linear discriminant analysis, *Journal of the American Statistical Association*, 90, 952-956.

[9] Hinkley, D. V. and Wang, S. (1988) More about transformations and influential cases in regression, *Technometrics*, 30, 435-440.

[10] Kim, C. (1996) Cook's distance in spline smoothing, *Statistics and Probability Letters*, 31, 139-144.

[11] Kim, C., Storer, B. E. and Jeong, M. (1996) A note on Box-Cox transformation diagnostics, *Technometrics*, 38, 178-180.

[12] Pregibon, D. (1981) Logistic regression diagnostics, *The Annals of Statistics*. 9, 705-724.

[13] Silverman, B. W. (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 47, 1-52.

[14] Tsai, C. L. and Wu, X. (1990) Diagnostics in transformation and weighted regression, *Technometrics*, 32, 315-322.

[15] Thomas, W. (1991) Influence diagnostics for cross-validated smoothing parameter in spline smoothing, *Journal of the American Statistical Association*, 86, 693-698.