A Step-wise Test for Identical Normal Distributions

Khairul Islam[1], Mian Arif Shams Adnan[2], Tanweer J Shapla[3]

[1,3]Department of Mathematics, Eastern Michigan University, Ypsilanti, Michigan 48197

[2]Department of Mathematical Sciences, Ball State University, Muncie, IN 47304

Abstract

Given two samples drawn from two normal populations, we wish to test if two populations are identical. The traditional test of the equality of two means requires the assumption of the equality of two variances. But the variances that are assumed equal are less pragmatic to be equal in real life. Since a normal distribution is characterized by two parameters mean and variance, the test of identical normal distributions might be carried out using two means and variances, simultaneously. As such, several tests have been attempted incorporating means and variances, simultaneously. The power of the underlying tests will be obtained using a Monte Carlo simulation from two populations.

Key Words: Trimmed mean, power, level of significance

1. Introduction

Let x and y represent the measurement of the same variable for two populations or two different groups. Let x and y be both distributed as normal. For example, x and y could be intelligence quotient (IQ) for male and female, respectively, which are normal. Given x and y follow normal distributions, we would like to investigate if x and y follow an identical normal distribution. Since a normal distribution is characterized by two parameters mean and variance, it seems ideal to test for the identity of the two normal populations by considering means and variances, simultaneously. As of now, we test equality of two means and variances by employing two separate tests. However, the classical test of equality of two means depends on the equality of two variances. In this paper, we wish to test identity of two normal populations on the basis of tests involving simultaneously both means and variances. The estimated power and level of significance will be investigated via simulation by noting consistent behavior of the test statistics under alternative and null distributions.

2. Notations

Let $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ be two samples from two normal populations, and $\bar{x}$ and $\bar{y}$ be their respective sample means. Let s_x^2 and s_y^2 be sample variances of x and y. Let $\bar{x}(p)$ and $\bar{y}(p)$ be the sample means with p% data values trimmed from both tails of the samples. Let $s_x^2(p)$ and $s_y^2(p)$ be the sample variances of x and y after trimming p% of data values.

kislam@emich.edu

Given x and y, we wish to test

$$\mathrm{H}_0: x \text{ and y have an indentical normal distribution}$$

against

$$\mathrm{H}_a: x \text{ and y do not have an indentical normal distribution}$$

3. Methods

In classical t-test, we assume that the two populations have an identical variances to test the equality of two population means. For the test of equality of two population variances, we employ an F test. In our study, we consider tests for means and variances simultaneously to determine the identity of two normal distributions. To this end, we investigate a number of approaches and their performances in terms of the estimated level of significance and power resulting from the simulations. For each approach, we wish to test the same null and alternative hypotheses:

$$\mathrm{H}_0: x \text{ and y have an indentical normal distribution}$$

against

$$\mathrm{H}_a: x \text{ and y do not have an indentical normal distribution}$$

Below we address briefly the methods that will be investigated for their performances empirically via simulation and/or using examples.

3.1 Classical t and F tests

In order to test null hypothesis H_0, we use classical t- and F-test simultaneously, given respectively by:

$$T_1 = \frac{\bar{x} - \bar{y}}{\widehat{\sigma_p}\sqrt{2/n}} \sim t_{2(n-1)}$$

and

$$F_1 = \frac{s_x^2}{s_y^2} \sim F(n-1, n-1)$$

where

$$\widehat{\sigma_p} = \frac{(n-1)\left(s_x^2 + s_y^2\right)}{2(n-1)}$$

We reject the identity of the two normal distributions if either of the two tests T_1 and F_1 results in the rejection after the Bonferroni type of adjustment for a given significance level α, and otherwise, we accept the identity of the two normal distributions.

3.2 Pooled Test

In order to test null hypothesis H_0, we use pooled sample mean and pooled sample variance to form t- and F-test. Under the null hypothesis, it is reasonable to estimate mean and variance using the combined sample given by:

$$\widehat{\mu_p} = \frac{n(\bar{x} + \bar{y})}{2n}$$

$$\widehat{\sigma_p} = \frac{(n-1)\left(s_x^2 + s_y^2\right)}{2(n-1)}$$

We then define $Z_x = \frac{X - \widehat{\mu_p}}{\widehat{\sigma_p}}$ and $Z_y = \frac{Y - \widehat{\mu_p}}{\widehat{\sigma_p}}$.

We wish to test

$$H_0: x \text{ and y have an indentical normal distribution}$$

against

$$H_a: x \text{ and y do not have an indentical normal distribution}$$

by utilizing the test statistics

$$T_2 = \frac{\overline{Z_x} - \overline{Z_y}}{\widehat{\sigma_p}\sqrt{2/n}} \sim t_{2(n-1)}$$

and

$$F_2 = \frac{s_{z_x}^2}{s_{z_y}^2} \sim F(n-1, n-1)$$

simultaneously, to test for equality of means ($\mu_x = \mu_y$) and equality of variances ($\sigma_x^2 = \sigma_y^2$).

We reject the identity of the two normal distributions if either of the two tests T_2 and F_2 results in the rejection after the Bonferroni type of adjustment for a given significance level α, and otherwise, we accept the identity of the two normal distributions.

3.3 Trimmed Test

Under this approach, we test identity of the two normal distributions by simultaneously applying t- and F- test to the trimmed x and y for a given value of trimming (p). Let $\mu_x(p)$ and $\mu_y(p)$ be two population means with $p\%$ of data points trimmed from both tails of the distributions of x and y, respectively.

We wish to test

$$H_0: x \text{ and y have an indentical normal distribution}$$

against

$$H_a: x \text{ and y do not have an indentical normal distribution}$$

by utilizing the test statistics

$$T_3 = \frac{\bar{x}(p) - \bar{y}(p) - \left(\mu_x(p) - \mu_y(p)\right)}{\hat{\sigma}(\bar{x}(p) - \bar{y}(p))}$$

and

$$F_3 = \frac{s_x^2(p)}{s_y^2(p)}$$

simultaneously, for a given value of p, to test for equality of means ($\mu_x = \mu_y$) and equality of variances ($\sigma_x^2 = \sigma_y^2$).

Under H_0, T_3 is expected to follow a t-distribution with degrees of freedom $t_{2n-4\lfloor np\rfloor-2}$, and F_3 as $F(\nu_1, \nu_2)$-distribution with degrees of freedom ν_1 and ν_2, where $\nu_1 = \nu_2 = n-2\lfloor np\rfloor-1$ and for a given value of p.

One can estimate $\hat{\sigma}(\bar{x}(p) - \bar{y}(p))$ by either

$$\hat{\sigma}^2(\bar{x}(p) - \bar{y}(p)) = \frac{1}{2n - 4\lfloor np\rfloor - 2}\left[\sum_{i=\lfloor np\rfloor+1}^{n-\lfloor np\rfloor}\left(x_{(i)} - \bar{x}(p)\right)^2 + \sum_{i=\lfloor np\rfloor+1}^{n-\lfloor np\rfloor}\left(y_{(i)} - \bar{y}(p)\right)^2\right]$$

or,

$$\hat{\sigma}^2(\bar{x}(p) - \bar{y}(p)) = \frac{\sum_{i=\lfloor np\rfloor+1}^{n-\lfloor np\rfloor}\left(x_{(i)} - \bar{x}(p)\right)^2}{n - 2\lfloor np\rfloor - 2} + \frac{\sum_{i=\lfloor np\rfloor+1}^{n-\lfloor np\rfloor}\left(y_{(i)} - \bar{y}(p)\right)^2}{n - 2\lfloor np\rfloor - 2}$$

where $\lfloor np\rfloor$ =floor of $(n \times p/100)$ for p% trimmed sample.

An estimate of $\hat{\sigma}^2(\bar{x}(p) - \bar{y}(p))$ can also be obtained using the bootstrap procedure. An algorithm for estimating $\hat{\sigma}^2(\bar{x}(p) - \bar{y}(p))$ using bootstrap procedure is given below:

Given x and y, generate B bootstrap samples. For each bootstrap sample and given p, compute $d^{j*}(p) = \bar{x}^{j*}(p) - \bar{y}^{j*}(p)$, for $j = 1,2, \ldots, B$, where B is the desired value of the bootstrap replication size. An estimate of $\hat{\sigma}^2(\bar{x}(p) - \bar{y}(p))$ using bootstrap replications is given by:

$$\widehat{\sigma^*} = \frac{1}{B-1}\sum_{j=1}^{B}\left(d^{j*}(p) - \bar{d}^*(p)\right)^2$$

where $\bar{d}^*(p) = \sum_{j=1}^{B}\left(\bar{x}^{j*}(p) - \bar{y}^{j*}(p)\right)/B$.

The choice of bootstrap replication size can be considered following Efron (1987), Booth and Sarker (1998), Hall (1992), etc. In this paper, however, we utilized Yuen's trimmed mean test, Yuen (1974), available in R via PairedData package, and F-test simultaneously to test for equality of means ($\mu_x = \mu_y$) and equality of variances ($\sigma_x^2 = \sigma_y^2$).

We reject the identity of the two normal distributions if either of the two tests T_3 and F_3 results in the rejection after the Bonferroni type of adjustment for a given significance level α, and otherwise, we accept the identity of the two normal distributions.

3.4 Kolmogorov-Smirnov Test

Given two normal distributions, to test H_0: The two normal distributions are identical against H_1: The two normal distributions are not identical, we also consider the Kolmogorov-Smirnov test given by

$$D_n = \sup_x \left| F_{X,n}(x) - F_{Y,n}(x) \right|$$

where $F_{X,n}$ and $F_{Y,n}$ are the empirical distribution functions, also termed as cumulative distribution functions (CDFs) of the first and the second sample, respectively, and sup is the supremum function. We implement this test in R and compute the p-value to decide so as to accept or reject the null hypothesis. Considering this test would allow us to justify and compare the performances of other three tests as reference to this test.

4. Simulation

In this section, we consider simulations from selected normal distributions to investigate the testing power and size of the underlying tests. The estimated power and size are the rejection rates of identity of normal distributions under alternative and null models, respectively, over all simulations for a Monte Carlo size of $M = 1000$. Let $x \sim N(\mu_x, \sigma_x^2)$ and $y \sim N(\mu_y, \sigma_y^2)$. We consider the following two forms of alternative models:

(i) $M_1: \mu_x = \mu_y + \Delta, \Delta \neq 0$

(ii) $M_2: \sigma_x^2 = k\sigma_y^2, k \neq 0, 1$

For estimating the level of significance, we consider distributions of x and y under the null model ($\mu_x = \mu_y,\ \sigma_x^2 = \sigma_y^2$ and $x, y \sim N(\mu_x, \sigma_x^2)$) and estimate the proportion of rejection over all simulation.

In simulation, we set values of Δ arbitrarily equal to 0.25, 0.50, 0.75, 1, 1.50, and values of k arbitrarily equal to 0.5, 2, 2.5, 3 to determine the effect of Δ and k on underlying tests. We also consider the sample size arbitrarily equal to 10, 15, 20, 25, 30, so as to understand the finite sample performance of underlying tests measured by the testing power and estimated level of significance.

The performance of the simulation study under models (i) and (ii) are reported in Tables 1 and 2, in terms of the testing power, for varying values of the sample size. The estimated level of significance under the null model has been reported in Table 3. The simultaneous performance of classical t- and F-test are reported under the heading (T_1, F_1), the simultaneous performance of pooled t- and F-test are reported under the heading (T_2, F_2), the simultaneous performance of trimmed t- and F-test are reported under the heading (T_3, F_3), and the performance of Kolmogorov-Smirnov test are reported under the heading $(K - S)$.

Table 1: Estimated power for varying values of Δ and sample size n

Δ	(T_1, F_1)	(T_2, F_2)	(T_3, F_3)	$(K-S)$
		$n = 10$		
0.25	0.073	0.073	0.085	0.025
0.5	0.123	0.123	0.125	0.049
0.75	0.262	0.262	0.258	0.123
1	0.455	0.455	0.453	0.245
1.5	0.813	0.813	0.810	0.578
		$n = 15$		
0.25	0.089	0.089	0.091	0.037
0.5	0.208	0.208	0.221	0.140
0.75	0.397	0.397	0.398	0.300
1	0.650	0.650	0.649	0.522
1.5	0.939	0.939	0.938	0.879
		$n = 20$		
0.25	0.091	0.091	0.101	0.073
0.5	0.245	0.245	0.244	0.227
0.75	0.53	0.530	0.511	0.449
1	0.804	0.804	0.786	0.728
1.5	0.992	0.992	0.988	0.976
		$n = 25$		
0.25	0.091	0.091	0.106	0.083
0.5	0.320	0.320	0.316	0.259
0.75	0.650	0.650	0.628	0.548
1	0.890	0.890	0.879	0.826
1.5	0.999	0.999	0.998	0.989
		$n = 30$		
0.25	0.119	0.119	0.123	0.098
0.5	0.352	0.352	0.351	0.299
0.75	0.711	0.711	0.696	0.613
1	0.933	0.933	0.928	0.878
1.5	1.000	1.000	1.000	0.997
		$n = 50$		
0.25	0.164	0.164	0.167	0.144
0.5	0.591	0.591	0.588	0.520
0.75	0.939	0.939	0.940	0.887
1	0.995	0.995	0.998	0.992
1.5	1.000	1.000	1.000	1.000

Table 2: Estimated power for varying values of k and sample size n

k	(T_1, F_1)	(T_2, F_2)	(T_3, F_3)	$(K-S)$
		$n = 10$		
0.5	0.119	0.119	0.111	0.017
2	0.111	0.111	0.113	0.010
2.5	0.202	0.202	0.179	0.020
3	0.293	0.293	0.258	0.026
	15	$n = 15$		
0.5	0.152	0.152	0.154	0.024
2	0.177	0.177	0.173	0.054
2.5	0.257	0.257	0.240	0.041
3	0.417	0.417	0.371	0.051
	20	$n = 20$		
0.5	0.240	0.240	0.229	0.049
2	0.237	0.237	0.211	0.052
2.5	0.387	0.387	0.356	0.065
3	0.548	0.548	0.498	0.083
	25	$n = 25$		
0.5	0.298	0.298	0.285	0.073
2	0.335	0.335	0.291	0.073
2.5	0.487	0.487	0.412	0.090
3	0.652	0.652	0.561	0.112
	30	$n = 30$		
0.5	0.361	0.361	0.329	0.060
2	0.373	0.373	0.349	0.062
2.5	0.586	0.586	0.523	0.091
3	0.731	0.731	0.666	0.130
	50	$n = 50$		
0.5	0.585	0.585	0.530	0.092
2	0.598	0.598	0.551	0.090
2.5	0.838	0.838	0.763	0.150
3	0.921	0.921	0.905	0.239

Table 3: Estimated level of significance ($\alpha = 0.05$) for varying value of the sample size n under the null model

n	(T_1, F_1)	(T_2, F_2)	(T_3, F_3)	$(K-S)$
10	0.042	0.042	0.047	0.014
15	0.048	0.048	0.049	0.023
20	0.044	0.044	0.044	0.033
25	0.049	0.049	0.057	0.039
30	0.053	0.053	0.058	0.046
50	0.054	0.054	0.055	0.035

5. Results and Discussion

Estimated testing power for underlying tests have been reported in Tables 1 and 2 for varying values of Δ and k, and an arbitrarily chosen set of sample size. The estimated level of significance (at 5% level of significance) has been reported in Table 3 for varying sample size. It appears that the performance of (T_1, F_1) and (T_2, F_2) are identical with respect to estimated power and level of significance. The estimated power of three tests (T_1, F_1), (T_2, F_2) and (T_3, F_3) are much higher than $(K-S)$.

It also follows that, under the model (i), the power of all tests increases significantly as the value of the mean difference Δ increases. The estimated power also increases with the increasing values of the sample size n. It also follows that for lower mean difference ($\Delta = 0.25$), the performance of (T_3, F_3) seems little better than that of (T_1, F_1) or (T_2, F_2). While the estimated testing power of $(K-S)$ is lower than other three tests for lower sample size, the power is comparable for large n.

Under the model (ii), the power of all tests increases significantly as the value of the variance ratio k between the two population increases. The sample size also has an increasing effect on the estimated power. Interestingly, while the performance of (T_1, F_1), (T_2, F_2) and (T_3, F_3) seems to be comparable with increasing k, the power of $(K-S)$ test breaks down completely with much lower power than is expected for a 5% level of significance.

As far as the estimated level of significance is concern, the $(K-S)$ test provides lower rate of rejection, below the nominal level of 5% even for large sample size. The performance of (T_1, F_1) are (T_2, F_2) are identical and is comparable with (T_3, F_3).

6. Conclusion

In this paper, we sought for testing identity of two normal populations by employing three tests incorporating means and variances, simultaneously, and using the Kolmogorov-Smirnov test. The idea is to simultaneously test the equality of two means and variances. In order to control for Type I error rate (α), we used Bonferroni type adjustment (Bonferroni, 1936; Holm, 1979; Miller, 1991) to simultaneously implement t and F tests. The results of simulation suggest that the three simultaneously performed (t, F)-tests (T_1, F_1), (T_2, F_2) and (T_3, F_3) provided better power as compared to the Kolmogorov-

Smirnov test. The power of all tests increases as the mean-difference, variance-ratio difference, and the sample size increase. However, $(K-S)$ test provided lower power consistently. Even though the estimated level of significance of all tests seem to be satisfactory for controlling Type I error rate, the $(K-S)$ test seems provide an under estimate of true level of significance. Given these facts, we recommend the use of either of the three tests (T_1, F_1), (T_2, F_2) or (T_3, F_3) for testing the identity of two normal populations so as to achieve a better performance than the traditional use of the $(K-S)$ test.

References

Yuen, K.K. (1974). The two-sample trimmed t for unequal population variances. Biometrika, **61**, 165-170.

Bonferroni, C. E. (1936) "Teoria statistica delle classi e calcolo delle probabilità." *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **8**, 3-62.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65-70.

Miller, R. G. Jr. *Simultaneous Statistical Inference.* New York: Springer-Verlag, 1991.

Hall, P. The Bootstrap and Edgeworth Expansion. Springer-Verlag: New York, 1992.

Booth, JG and Sarker, S. (1998) Monte Carlo approximation of bootstrap variances. Annals of Statistics, **52**: 354-357.

Efron, B. (1987). Better bootstrap confidence intervals (with discussion). Journal of American Statistical Association, **82**: 171-200.