1 Boxplot Slippage Tests

Abstract

Two hypothesis tests are presented to test for a difference in location between two independent samples. The tests are performed by inspecting boxplots constructed from samples drawn from the two populations of interest. The populations being sampled should be symmetric and have equal variances. The decision to accept or reject the hypothesis that the population means are equal is based on the relative positions of features of the two boxplots.

The first boxplot slippage test requires that the samples are of approximately equal size with $n \geq 5$. The claim of equal treatment means can be rejected at $\alpha = 0.05$ if the two boxes are completely slipped from each other. The second boxplot slippage test requires that the samples are of approximately equal size with $n \geq 30$. The claim of equal treatment means can be rejected at $\alpha = 0.05$ if one of both of the boxplot medians falls outside of the other sample's box. The power of the two tests is relatively poor compared to the power of Tukey's quick test or the two-sample t test, however, the boxplot slippage tests are very easy to use and have good to excellent protection against type 1 errors.

Introduction

The two-sample t test to compare two independent samples for a difference in location is a powerful method, but as a minimum it requires a calculator and tables of critical values or appropriate statistical software. Many two-sample location problems can be addressed with sufficient rigor using simpler methods of analysis that only require simple inspection of the data, the classic example being Tukey's quick test. The purpose of this method is to present another quick test to analyze the two independent sample location problem using boxplots.

Where the Technique is Used

The boxplot slippage tests are used to test two independent samples for a difference in location. The populations being sampled should be symmetric and have equal variance. The boxplot slippage tests can be used any time that the two independent sample t test or Tukey's quick test are appropriate.

Data

1. The data consist of two independent random samples drawn from continuous populations.
2. The sample sizes should be at $n \geq 5$ for the first boxplot slippage test and $n \geq 30$ for the second boxplot slippage test.
3. The ratio of the sample sizes (bigger/smaller) should be less than 1.33.
4. The data can be quantitative but must at least be ordinal (i.e. capable of being ordered by size).

Assumptions

1. The samples are independent and random.
2. The two samples are measured (or ordered) on the same scale with equal accuracy and precision.
3. The distributions of the populations being sampled are the same (i.e. in terms of shape and variation) except for a possible difference in location.

Hypotheses Tested

The hypotheses to be tested are: $H_0 : \tilde{\mu}_1 = \tilde{\mu}_2$ vs. $H_A : \tilde{\mu}_1 \neq \tilde{\mu}_2$ where $\tilde{\mu}$ indicates the median. When the distributions being studied are symmetric then the parameters tested are population means.

Procedure

1. Collect independent random samples of approximately the same size from the two populations.
2. Construct boxplots of the data using the sample measurement scale. The boxplots should use quartiles to determine the box ends.
3. Inspect the two boxplots for symmetry and equal variance. If there is evidence that one or both distributions are asymmetric or if the samples might come from populations with different variances then do not proceed with the tests.
4. If the sample size is $n \geq 5$ then accept H_0 if the boxes are overlapped and reject H_0 if the boxes are slipped from each other.
5. If the sample size is $n \geq 30$ then accept H_0 if both medians fall inside of the other samples's boxes. Reject H_0 if one of both of the medians falls outside of the other sample's box.

Example: *Two independent random samples of size $n = 10$ were drawn to test for a difference in location. Boxplots were constructed from the sample data and are shown in Figure 1A. Is there evidence that the samples come from populations with different locations?*
Solution: *The first boxplot slippage test is appropriate because: the sample sizes are equal, they meet the minimum sample size condition, and the boxplots appear to be reasonably symmetric and display at least approximately equal variance. The two boxes are slipped from each other so we must conclude that there is a difference in location between the two populations.*

Example: *Two independent random samples of size $n = 10$ were drawn to test for a difference in location. Boxplots were constructed from the sample data and are shown in Figure 1B. Is there evidence that the samples come from populations with different locations?*
Solution: *The sample size and distributional assumptions appear to satisfy the requirements of the first boxplot slippage test. The two boxes are overlapped so we can't reject H_0:$\mu_1 = \mu_2$, however, a more powerful test like the two-sample t test might be able to detect a difference between the two populations.*

Example: *Two independent random samples of size $n = 35$ were drawn to test for a difference in location. Boxplots were constructed from the sample data and are shown in Figure 1B. Is there evidence that the samples come from populations with different locations?*
Solution: *The sample size and distributional assumptions appear to satisfy the requirements of the second boxplot slippage test. The medians of both boxes fall outside of the other samples's boxes so we have to conclude that there is a significant difference in location between the two populations.*

Theoretical Basis

The boxplot slippage tests are closely related to the nonparametric two-sample Smirnov test. The test statistic of the Smirnov test is the magnitude of the difference (i.e. the slippage) between the two empirical cumulative distribution functions. The boxplot slippage tests uses a similar measure of slippage determined from the relative positions of features in the two boxplots being compared. The Smirnov test actually requires larger sample sizes than the $n \geq 5$ and $n \geq 30$ conditions stated here for the first and second boxplot slippage tests, respectively, however the Smirnov test makes no assumptions about the shape or homoscedasticity of the distributions being studied so its sample sizes are generally larger. The sample sizes specified for the boxplot slippage tests were determined by Monte Carlo simulation

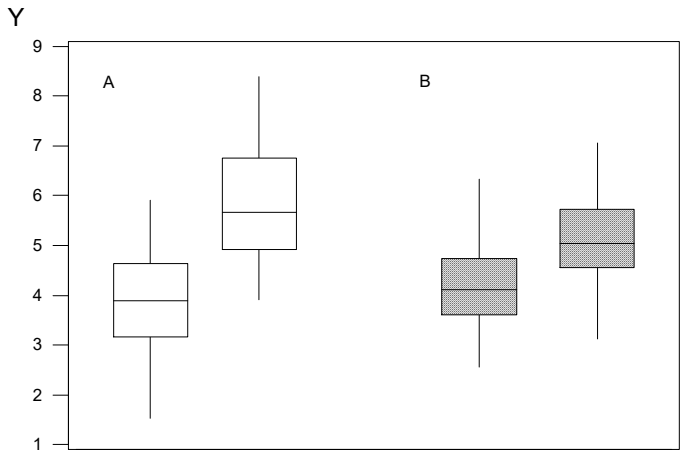

Figure 1: Data Sets To Be Tested for Differences In Location

using normal homoscedastic populations. If these requirements aren't met then the sample sizes should be made larger, per the Smirnov test, to validate the use of the boxplot slippage tests.

OC Curves

The performance of the boxplot slippage tests was evaluated by comparing them to each other, to Tukey's quick test, and to the two-sample t test using Monte Carlo simulation. The simulation used random samples drawn from normal homoscedastic populations with specified differences $\Delta z = \Delta\mu/\sigma$ between the two population means. Samples of size $n = 5, 12, 30, 80$ were considered for differences between the population means of $\Delta z = 0 : 3/0.5$. Each sample size and difference combination was simulated 10000 times. Plots of the OC curves for each sample size are shown in Figure 2 where $P_a(H_0)$ is the probability of accepting $H_0 : \mu_1 = \mu_2$. The OC curves corresponding to the first and second boxplot slippage tests, Tukey's quick test, and the two-sample t test have the labels "Box", "Median", "Tukey", and "t", respectively. The OC curves for Tukey's test were determined using $T \geq 7$ as the rejection criterion for H_0 and the t test was evaluated using a two-tailed test with $\alpha = 0.05$.

Good hypothesis tests must have reasonably low type 1 and type 2 error rates. The numerical value of the type 1 error rate in the Figure corresponds to the complement of $P_a(H_0)$ when $\Delta z = 0$. Tolerable values for the type 1 error rate are typically $\alpha \leq 0.05$ which corresponds to $P_a(H_0) \geq 0.95$. Low type 2 error rates correspond to small values of $P_a(H_0)$ when $\Delta z > 0$. Low type 2 error rates correspond to high power because power (P) is the complement of the type 2 error rate β. Ideally the OC curve for an hypothesis test should start off at $P_a(H_0) \geq 0.95$ when $\Delta z = 0$ and fall off quickly as Δz increases. Test methods that have OC curves with lower values of $P_a(H_0)$ when $\Delta z > 0$ are considered better methods than those that have OC curves with higher values of $P_a(H_0)$. The two-sample t test serves as the benchmark for "good" tests.

The OC curves for $n = 5$ show that the second (i.e. "Median") boxplot slippage test has a very high type 1 error rate so it is not an appropriate method of analysis for such small sample sizes. The first boxplot slippage test and Tukey's quick test give almost identical performance to each other. The two-sample t test has a slightly high type 1 error rate than they do but it also has slightly higher power when $\Delta z > 0$.

The OC curves for $n = 12$ still show that the second boxplot slippage test is still not safe to use because its type 1 error rate is too high. The first boxplot slippage test (Box) has a much lower low type 1 error rate than the other methods. It is not as powerful as the other tests but it is still comparable in performance to Tukey's quick test. One thing is certain - if two boxplots have slipped boxes when

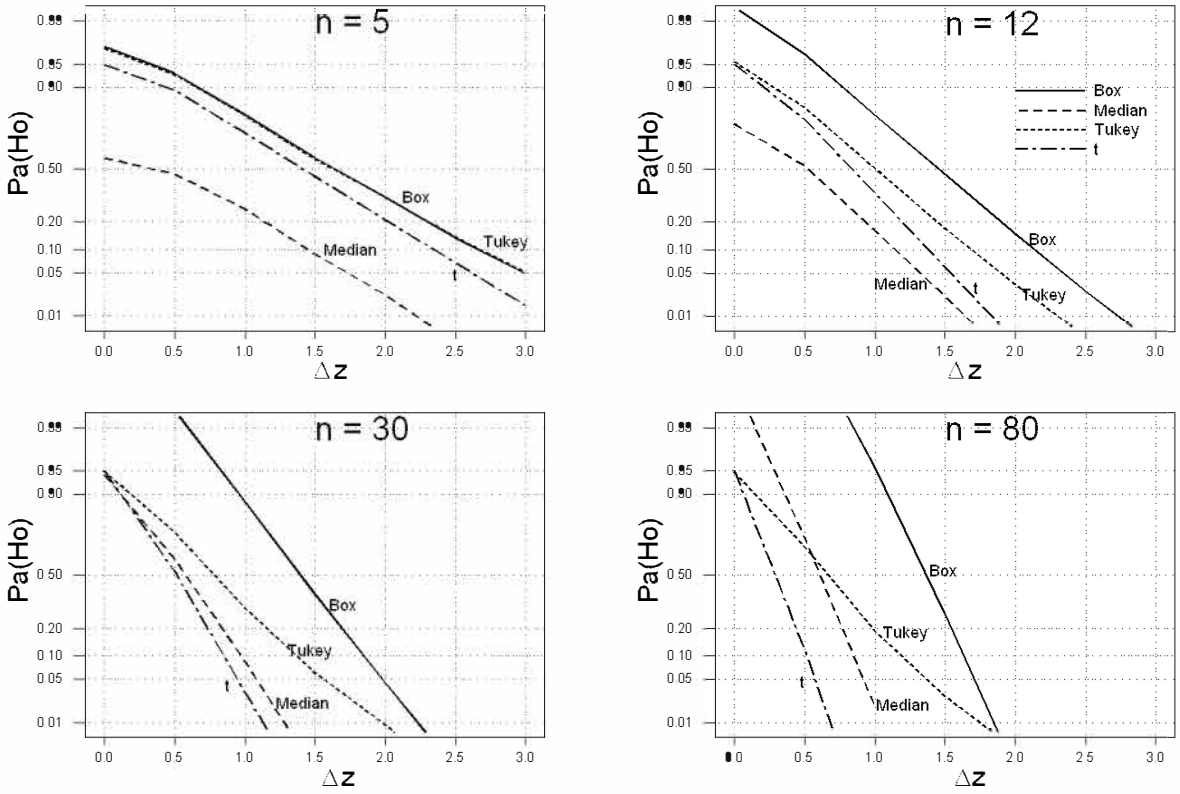

Figure 2: OC curves for four two-sample tests versus sample size.

$n \geq 12$ there is most certainly a statistically significant difference between the two treatments.

The OC curves for $n = 30$ show that the first boxplot slippage test (i.e. "Box") has considerably worse power than the other methods. However, the second boxplot slippage test (Median) has an acceptably low type 1 error rate, has power comparable to the t test, and is always more powerful than Tukey's quick test.

The OC curves for $n = 80$ show that the first boxplot slippage test's power is much worse than the other tests. The other three tests all have comparable type 1 error rates. For larger differences between the means, such as $\Delta z \geq 0.5$, the second boxplot slippage test has better power than does Tukey's quick test. However, the two-sample t test is by far the most powerful test available. One thing is for certain - if one or both of the medians are slipped from other samples's boxes when $n \geq 80$ then there is most certainly a statistically significant difference between the two treatments.

Despite the comparatively poor power of the first boxplot slippage test for large sample sizes, it does have a very low type 1 error rate which makes it safer to use than the other methods for making multiple pairwise comparisons. The first boxplot slippage test is also exceptionally easy to use for this application - if a single straight line can be drawn across all of the boxplots that passes through all of the boxes then there are no slipped pairs of boxes so no pairs of treatments are different from each other. The more treatments there are to be compared the greater the number of observations in each treatment should be to provide good protection against type 1 errors.

Example: *An experiment was performed to test for differences in location between five different treatments. Random samples of size $n = 12$ were drawn from each treatment and then boxplots were constructed for the response. The boxplots are shown in Figure 3. Is there evidence of any differences between treatments?*

Solution: *All five samples were of size $n = 12$ so the first boxplot slippage test is appropriate. By*

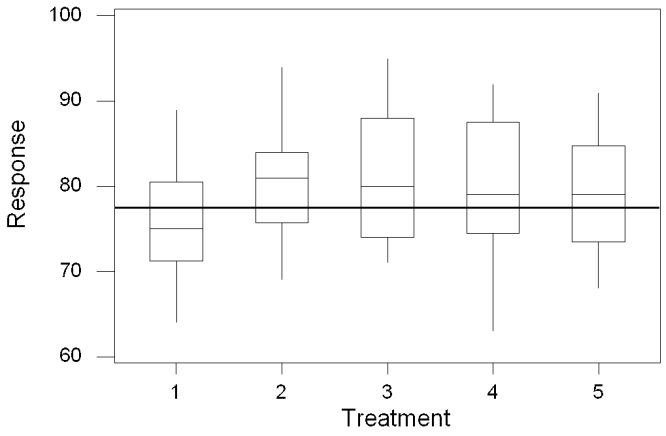

Figure 3: Multiple Comparisons Using Boxplot Slippage Tests

inspection of Figure 3, the boxplots all appear to be reasonably symmetric and of comparable size so the assumptions required to validate the use of the first boxplot slippage test appear to be satisfied.

Since there are five treatments there are $\binom{5}{2} = 10$ *pairs of comparisons to be made. Those comparisons are 12, 13, 14, 15, 23, 24, 25, 34, 35 where 12 indicates a comparison between the first and second treatment. With so many tests to perform the chances that one or more of them will result in a type 1 error (an erroneous claim that there is a difference between two treatments) is inflated by about a factor of ten compared to the risk of a single test. Figure 2 for* $n = 12$ *shows that the type 1 error rate for the first boxplot slippage test is very low - in fact, much less than* $\alpha = 0.01$ *- so the overall risk of performing ten box slippage tests is probably acceptably low.*

A single line was drawn across the boxplots in Figure 3 that passes through all five boxes. This indicates that all pairs of boxes are overlapped so, at least according to the first boxplot slippage test, there is no evidence for any location differences between the treatments. This simple observation, that the line passes through all of the boxes, corresponds to the execution of all ten possible pairs of boxplot slippage tests! This is a quick and easy way to perform multiple comparisons but the method is only valid because of the very low type 1 error rate of the first boxplot slippage test with this sample size.

References

Conover, W.J., Practical Nonparametric Statistics, 2nd Ed., John Wiley and Sons, 1980.

Tukey, J.W., "*A Quick, Compact, Two-Sample Test to Duckworth's Specifications,*" Technometrics, Vol. 1, No. 1, February, 1959, p. 31.