# The Median Test with Zero-Inflated Distributions

William D. Johnson*, Jeffrey H. Burton, Robbie A. Beyl, and Jacob E. Romer
Department of Biostatistics, Pennington Biomedical Research Center, Louisiana State University, 6400 Perkins Road, Baton Rouge, LA, 70808

## Abstract
The median test is often used to compare two or more distributions that have similar but asymmetric shapes and is typically used as an alternative, nonparametric approach to other tests of location such as the t-test. In the case of zero-inflated distributions, it is of interest to compare the distributions with respect to their proportions of observed zero values coupled with the comparison of the medians for their observed non-zero values. We present an application of the median test to simultaneously test the equality of the proportions of zeros as well as the medians among several groups. Results of simulation studies are reported to summarize some characteristics of the test.

**Keywords**: Asymptotic chi-square test, Large sample test, Median test, Nonparametric test, Zero-inflated distributions

## 1. Introduction

Suppose we are interested in comparing two or more independent samples from distributions that are asymmetric or have some unusual form. A useful comparison is often among the medians as this a nonparametric estimate of the location of the distribution. In this case, Mood's median test [1] is a common choice and, as a special case of Pearson's chi-square test, it is easy to explain and set-up. However, if the data have an excess of zeros (i.e., a zero-inflated distribution), the standard median test may not be the most appropriate tool to compare the distributions because it may ignore differences in the proportions of zeros that are less than the combined sample median. Instead of testing the equality of medians of the entire sample, we would like to split the hypothesis into two tests: one for the equality of the proportions of zeros and one for the equality of medians of the non-zero values. Typically, observations with zero-values are significant is some way (such as observations below a detection limit) and differences in the proportions of zeros between groups may be a distinguishing characteristic. Although the standard median test (not taking into account the proportion of zeros) may be more powerful in certain situations due to larger cell sizes, there is no component to test the equality of proportions of zeros, just the overall sample median.

Zero-inflated distributions can be modelled as a mixture of a point distribution (not necessarily at zero) and some other continuous or discrete distribution. There are many situations in which zero-inflated distributions may occur. One common cause is lower detection limits in measurement devices or laboratory procedures. If a value is less than the detectable limit, the observation is given a set value, thereby creating a "mass" of some particular value. Variables such as age can be collapsed or combined for similar observations as this may simplify analyses. For instance, all ages less (or greater) than a certain value may be given the same value. Count data modelled with zero-inflated Poisson models are perhaps most familiar given their prevalence. Although fairly common, especially in biomedical applications, there are very few tests for homogeneity of zero-inflated distributions, particularly nonparametric tests.

The primary focus of zero-inflated data has been for count data with excess zeros to be modelled as either zero-inflated Poisson [2, 3] or negative binomial distributions [4]. Some tests for homogeneity of zero-

inflated Poisson data include those described by Tse et al. [5] and Bedrick et al. [6]. Lachenbruch [7, 8] proposed two-part tests – one for the equality of the proportion of zeros as well as equality of a continuous distribution with a *z*-test, Kolmogorov-Smirnov test, or Wilcoxon test. The median test can be easily modified to test for equality of medians of zero-inflated distributions for either continuous or discrete data, and will be the focus of this paper.

## 2. Test Procedure

### 2.1 The Median Test

Suppose random samples are available from each of $K$ populations and we wish to test the null hypothesis that some arbitrary quantile, $Q$, of the populations are equal; that is, the null hypothesis $H_0$: $Q_1 = Q_2 = \cdots = Q_K$ is tested against the alternative, $H_1$: one or more inequalities exist such that $Q_i \neq Q_j$ for at least one $i \neq j$ where $i, j = 1, 2, \ldots, K$. We allow $Q$ to be any percentile, although the 50th (the median) is most commonly used; we use the term "median test" to include any percentile. The following procedure is a generalization of the median test, which is a special, single- percentile case of the general percentile test described by Johnson et al. [9]:

(1) Combine the $K$ samples and calculate the combined sample percentile estimate of $Q$. Denote the combined sample percentile as $q$.
(2) For each of the $K$ samples, sort the respective samples' observations into two bins or columns. Observations that are less than or equal to $q$ are placed in the first bin; otherwise, the second bin.
(3) Construct a $K \times 2$ contingency table where each row contains the sorted observations for one of the $K$ populations, as in Step 2.
(4) Perform the chi-square test on the contingency table with degrees of freedom equal to $(K - 1)$.

### 2.2 Median Test with Zero-Inflated Distributions

A zero-inflated distribution can be written as a mixture of two distributions: $D = \pi g + (1 - \pi) f$, where $\pi$ is the probability of an observation being in the point distribution, $g$, and $f$ is the non-constant distribution. Denote the distribution of the $i$th population as $D_i = \pi_i g + (1 - \pi_i) f_i$, $i = 1, \ldots, K$ and $H_0$: $D_1 = D_2 = \cdots = D_K$. It is assumed that $g$ is the same fixed constant in all populations because it is a parameter of a specific process that produces $D$, such as limits on a measurement device. In order for the null hypothesis (all $D_i$ are identical; $i = 1, 2, \ldots, K$) to be false, at least one inequality exists such that $D_i \neq D_j$ for at least one $i \neq j$ where $i, j = 1, 2, \ldots, K$. Thus, at least one of $\pi_i$ or $f_i$ must be unequal. Note that $D$ could also contain two point distributions, one at the minimum value and one at the maximum in the domain. In this case, $D_i = \pi_{i1} g_l + \pi_{i2} g_u + (1 - \pi_{i1} - \pi_{i2}) f_i$, where $g_l$ is the point distribution at the minimum value and $g_u$ is the point distribution at the maximum value, with probability $\pi_{i1}$ and $\pi_{i2}$, respectively.

The median test outlined in the previous section can be modified to simultaneously test the equality of proportions of zeros in all populations and the equality of the medians (or some other percentile) in all populations. Let $\pi_i$ ($i = 1, \ldots, K$) be the proportion of zeros in the $i$th population for a random variable $X$. For each sample, $X_i$, $P[X_i \leq 0] = \pi_i$ and the value of any percentile less than $\pi_i$ is 0. Note that the median is simply an estimate of a percentile and 0 could be considered just another percentile estimate – the estimate of the $\pi_i$th percentile. Since any percentile less than $\pi_i$ is equal to 0, we can select an arbitrary number such that each population's estimate is 0. However, because we use the combined sample estimates to create bins for the contingency table, we must select a percentile less than the combined sample proportion of zeros, denoted as $\bar{\pi}$.

Essentially, the purpose is to create an additional column in the contingency table where all values equal to 0 are placed. First, calculate the combined sample estimate of $q$ as in the standard median test. Then, further sort those values that are less than or equal to $q$ into another column. This creates a $K \times 3$

contingency table where the first column is for observations that are equal to 0 (the $\bar{\pi}$th combined sample percentile estimate), the second column for observations greater than 0 and less than or equal to $q$ (the combined sample estimate of $Q$th percentile), and the third column for observations greater than $q$. As with the median test, we perform a chi-square test on the contingency table with degrees of freedom equal to $2(K-1)$. Table 1 is an example of the resulting contingency table.

**Table 1:** Contingency table for testing equality of the medians with bin for zeros.

| Sample | Bin 1 ($X = 0$) | Bin 2 ($0 < X \leq q$) | Bin 3 ($X > q$) | Total |
|---|---|---|---|---|
| 1 | 14 | 12 | 78 | 104 |
| 2 | 29 | 21 | 41 | 91 |
| 3 | 41 | 28 | 26 | 95 |
| Total | 84 | 61 | 145 | 290 |

For the data in Table 1, we calculate a $\chi^2$ value of 46.5 with four degrees of freedom, which indicates that at least one of the proportions of zeros or the medians are unequal among at least two of the three groups. We can also see from Table 1 that, depending on the proportions of zeros in the samples, the median may result in small cell counts in the contingency table. If $\bar{\pi}$ is very close to 0.5, the median may not be a wise choice for $Q$ because the power of the chi-square test depends on the size of the particular cells. Care should be taken when choosing $Q$ to ensure a minimum expected value for each cell within each column.

We could perform a similar procedure for data with a point distribution at the maximum value as well. Instead of using the $\bar{\pi}$th combined sample percentile estimate for sorting observations, we use the $(1-\bar{\pi})$th combined sample estimate to sort observations that are greater than $q$. Let $u$ denote the value of the point distribution at the maximum of the domain which is equal to the $(1-\bar{\pi})$th combined sample percentile estimate. We would sort the data into three columns (each group by row): (1) $X < q$, (2) $q \leq X < u$, and (3) $X \geq u$. We can extend this further for data with point distributions at both the minimum and the maximum of the domain. Let $l$ ($u$) denote the $\bar{\pi}_l$th ($\bar{\pi}_u$th) combined sample percentile estimate corresponding to the minimum (maximum) value of the domain. We sort the data into four columns (each group by row): (1) $X \leq l$, (2) $l < X \leq q$, (3) $q < X < u$, and (4) $X \geq u$. Then, perform the chi-square test with $3(K-1)$ degrees of freedom.

## 3. Simulation Studies

The asymptotic properties of the test with zero-inflated distributions were investigated for skewed continuous data generated from gamma distributions. We used gamma distributions as a convenient method to generate skewed data, although the procedure is nonparametric. The proportion of zeros, sample size and parameters of the non-zero gamma distribution were varied and results are presented in Table 2. The subscripts for the parameters in the tables refer to the respective samples with sample '2' having constant values for the non-zero distribution. For the continuous case, sample '2' has shape parameter that is held at $\alpha_2 = 2$, and the scale parameter held at $\beta_2 = 2$. All simulations were conducted with R 3.1.2 using 10,000 replicate samples.

The results in Table 2 give some insight into the behavior of the test in various situations. Empirical type I error (first column of power results, boldface results) is adequate in samples of size 50. The power of the

test is determined by the interaction between the difference in the true ratio between rows in the contingency table (determined by the difference in the distributions and probability of zeros) and the relative proportion of each column, as well as sample size. In general, power is maximized when the relative proportion of column corresponds with the difference in the row profiles.

**Table 2:** Power simulations for testing the equality of medians with zero-inflated gamma distributions (empirical estimates of type 1 error in boldface).

| Sample Size ($n = m$) | $\pi_1$ | $\pi_2$ | Gamma Distribution Parameters | | | |
|---|---|---|---|---|---|---|
| | | | $\alpha_1 = 2$ $\beta_1 = 2$ | $\alpha_1 = 2.2$ $\beta_1 = 2.2$ | $\alpha_1 = 2.3$ $\beta_1 = 2.3$ | $\alpha_1 = 2.4$ $\beta_1 = 2.4$ |
| 50 | 0.1 | 0.1 | **0.0533** | 0.1363 | 0.2667 | 0.421 |
| | | 0.2 | 0.2122 | 0.3301 | 0.4276 | 0.5684 |
| | | 0.3 | 0.6119 | 0.6812 | 0.7432 | 0.8211 |
| 100 | 0.1 | 0.1 | **0.0532** | 0.2499 | 0.4937 | 0.7418 |
| | | 0.2 | 0.4129 | 0.5893 | 0.7520 | 0.8824 |
| | | 0.3 | 0.9051 | 0.9487 | 0.9685 | 0.9867 |
| 200 | 0.1 | 0.1 | **0.0502** | 0.4453 | 0.8041 | 0.9620 |
| | | 0.2 | 0.7152 | 0.8877 | 0.9671 | 0.9943 |
| | | 0.3 | 0.9977 | 0.9996 | 1.0000 | 1.000 |
| 50 | 0.2 | 0.2 | **0.0489** | 0.1233 | 0.2176 | 0.3522 |
| | | 0.3 | 0.1529 | 0.2398 | 0.3240 | 0.4528 |
| | | 0.4 | 0.4896 | 0.5472 | 0.6217 | 0.6887 |
| 100 | 0.2 | 0.2 | **0.0496** | 0.2096 | 0.4254 | 0.6581 |
| | | 0.3 | 0.2876 | 0.4514 | 0.6157 | 0.7738 |
| | | 0.4 | 0.8037 | 0.8608 | 0.9135 | 0.9516 |
| 200 | 0.2 | 0.2 | **0.0488** | 0.3910 | 0.7181 | 0.9283 |
| | | 0.3 | 0.5357 | 0.7500 | 0.9037 | 0.9757 |
| | | 0.4 | 0.9865 | 0.9917 | 0.9978 | 0.9997 |

For example, for any sample size and $\pi_1 = \pi_2$, $\alpha_1 \neq \alpha_2$, and $\beta_1 \neq \beta_2$, power is greater when the overall proportion of zeros is lower. In these cases, the first bin which contains all the zero observations has equal probability for both samples but the medians of the remaining observations are unequal. Thus scenarios with greater counts in the second and third columns (corresponding to unequal medians) have the greatest power. For example, simulations where $\pi_1 = \pi_2 = 0.1$ always have greater power than $\pi_1 = \pi_2 = 0.2$. Similarly, for any sample size and $\pi_1 = \pi_2/2$, $\alpha_1 = \alpha_2$, and $\beta_1 = \beta_2$ (the ratio of zeros is constant as well as non-zero distribution), power is greater when the overall proportion of zeros is greater. In cases where $\pi_1 \neq \pi_2$, $\alpha_1 \neq \alpha_2$, and $\beta_1 \neq \beta_2$, the relationship remains but is obscured by differences in the medians of the non-zero distributions. If we examine situations where the ratio of $\pi_1$ and $\pi_2$ are equal ($\pi_1 = 1/2\ \pi_2$) with unequal non-zero medians, we still observe greater power with greater $\bar{\pi}$ for the parameters in Table 2.

There are situations where the probability of zero is greater than 0.5, making the modified median test (as defined in Section 2) obsolete. One solution is to sort only the observations greater than zero (the median test on non-zero values) and append this $K \times 2$ table with the counts of the number of zeros, making a $K \times 3$ table. (This approach could be used for data with $\bar{\pi} < 0.5$; however, the results are similar to the procedure in Section 2 and neither approach is uniformly better in all scenarios). An alternative solution is to select $Q > \bar{\pi}$. Table 3 shows results of comparing gamma distributions with a high proportion of zeros with $Q = 0.85$.

**Table 3:** Power simulations for testing the equality of $Q = 0.85$ with zero-inflated gamma distributions (empirical estimates of type I error in boldface).

| Sample Size ($n = m$) | $\pi_1$ | $\pi_2$ | Gamma Distribution Parameters | | | |
|---|---|---|---|---|---|---|
| | | | $\alpha_1 = 2$ $\beta_1 = 2$ | $\alpha_1 = 2.2$ $\beta_1 = 2.2$ | $\alpha_1 = 2.3$ $\beta_1 = 2.3$ | $\alpha_1 = 2.4$ $\beta_1 = 2.4$ |
| 50 | 0.5 | 0.5 | **0.0515** | 0.1014 | 0.1652 | 0.2547 |
| | | 0.6 | 0.1323 | 0.1835 | 0.2656 | 0.3331 |
| | | 0.7 | 0.4286 | 0.4826 | 0.5260 | 0.5891 |
| 100 | 0.5 | 0.5 | **0.0507** | 0.1450 | 0.2716 | 0.4412 |
| | | 0.6 | 0.2283 | 0.3181 | 0.4298 | 0.5706 |
| | | 0.7 | 0.7492 | 0.7882 | 0.8378 | 0.8784 |
| 200 | 0.5 | 0.5 | **0.0502** | 0.2614 | 0.5219 | 0.7706 |
| | | 0.6 | 0.4225 | 0.6007 | 0.7537 | 0.8894 |
| | | 0.7 | 0.9685 | 0.9818 | 0.9885 | 0.9960 |

## 4. Concluding Remarks

We present a novel extension of the median test to compare zero-inflated distributions. By treating the proportions of zeros as another percentile estimate along with the median, we can simultaneously test the equality of the proportion of zeros as well as the equality of medians (or another percentile) between two or more groups. Zero-inflated distributions are very common yet there are few nonparametric procedures to test for homogeneity between populations of interest. The proposed test addresses this problem and provides a simple, easy to use procedure for testing any continuous or discrete set of data for any number of samples. Also, the test allows for comparisons of point distributions of any value at the minimum and/or maximum value in the domain. Thus it is not limited to point distributions at zero or non-negative values.

The chi-square test in itself could be considered a limitation as it is generally outperformed by other tests that do not require such large sample sizes. The choice of percentiles within the profile is limited by the mechanics of the chi-square test (such as minimum expected values) although percentiles other than the median may be chosen to avoid this. Power of the chi-square test is a function of the relative sizes of the cells and certain combinations of percentiles may reduce the power based on this alone. Further research could be done to investigate the severity of limitations of large-sample restrictions, as well as the intricate relationship between the underlying distributions, the choice of the percentile, and the power of the test. The exact power of the test could be explored in further research.

## Acknowledgments

This research was supported by 1 U54 GM104940 from the National Institute of General Medical Sciences of the NIH, which funds the Louisiana Clinical and Translational Science Center. William Johnson also received support from the National Center For Complementary & Integrative Health and the Office of Dietary Supplements of the National Institutes of Health under Award Number 3 P50AT002776 which funds the Botanical Research Center of Pennington Biomedical Research Center and the Department of Plant Biology and Pathology in the School of Environmental and Biological Sciences (SEBS) of Rutgers University. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## References

[1] Mood AM (1954) On the asymptotic efficiency of certain nonparametric two-sample tests. *The Annals of Mathematical Statistics* 25: 514−522.

[2] Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34: 1−14.

[3] Hall DB (2000) Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* 56: 1030−1039.

[4] Hausman J, Hall BH, Griliches Z (1984) Econometric models for count data with an application to the patents−R&D relationship. *Econometrica* 52: 909−938.

[5] Tse SK, Chow SC, Lu Q, Cosmatos D (2009) Testing Homogeneity of zero-inflated Poisson populations. *Biometrical Journal* 51: 159 – 170.

[6] Bedrick EJ, Hossain A (2013) Conditional tests for homogeneity of zero-inflated Poisson and Poisson-hurdle distributions. *Computational Statistics and Data Analysis* 61: 99 −106.

[7] Lachenbruch PA (1976) Analysis of data with clumping at zero. *Biometrische Zeitschrift* 18: 351−356.

[8] Lachenbruch PA (2001) Comparison of two-part models with competitors. *Statistics in Medicine* 20: 1217−1236.

[9] Johnson WD, Beyl RA, Burton JH, Johnson CM, Romer JE, Zhang L (2015) Use of Pearson's Chi-square for testing equality of percentile profiles across multiple populations. *Open Journal of Statistics* 5: 412−420.