Design of the Fuzzy Rank Tests Package

Charles J. Geyer

December 14, 2005

1 Sign Test

Suppose we observe data $X_1, \ldots, X_n$ that are IID from some distribution with median μ. The conventional sign test of a hypothesized value of μ is based on the test statistic W, which is the number of X_i strictly greater than μ.

If the distribution of the X_i is continuous, so no ties are possible, then under the following null hypothesis the distribution of W is $\text{Bin}(n, 1/2)$.

Hypothesis 1 *The hypothesized value of μ is the median of the distribution of the X_i.*

If the distribution of the X_i is not continuous, so ties are possible, we break the ties by "infinitesimal jittering" of the data. If l, t, and u data points are below, tied with, and above μ, respectively, after infinitesimal jittering we have $W = u + T$ of the jittered data points strictly greater than μ, where $T \sim \text{Bin}(t, 1/2)$. Again we have $W \sim \text{Bin}(n, 1/2)$, although, strictly speaking, this requires that we change the null hypothesis to the following.

Hypothesis 2 *The hypothesized value of μ is the median of the distribution of the infinitesimally jittered X_i.*

Since we cannot in practice distinguish infinitesimally separated points, we consider these two hypotheses the same in practice.

We call W the *latent test statistic* and base the fuzzy test on it. Table 1 in Geyer (submitted) shows how a fuzzy test based on such a test statistic works. The rest of that paper sketches the tie breaking by infinitesimal jittering we explore in detail here.

1.1 One-Tailed Tests

The latent fuzzy P-value for latent test statistic $u + T$ is $\text{Unif}(a, b)$ where

$$
\begin{aligned}
a &= \Pr\{W > u + T\} \\
b &= \Pr\{W \geq u + T\}
\end{aligned}
$$

This means the (actual, non-latent) fuzzy P-value has continuous, piecewise linear CDF with knots $\Pr\{W > u+t-k\}$ and corresponding values $\Pr\{T < k\}$, where $k = 0, \ldots, t+1$.

For a lower-tailed test, swap l and u and proceed as above.

1.2 Two-Tailed Tests

For a two-tailed test, we use same distributions of T and W as in the one-tailed test, but now the latent test statistic is

$$g(T) = \max(u + T, l + t - T).$$

The latent fuzzy P-value for latent test statistic $g(T)$ is $\text{Unif}(a, b)$ where

$$\begin{aligned} a &= \Pr\{|W| > g(T)\} \\ b &= \Pr\{|W| \geq g(T)\} \end{aligned}$$

Because of the symmetry of the distribution of W we always have

$$\begin{aligned} a &= 2\Pr\{W > g(T)\} \\ b &= \min\bigl(1, 2\Pr\{W \geq g(T)\}\bigr) \end{aligned}$$

Thus we can get the fuzzy P-value for a two-tailed test by simply calculating the fuzzy P-value for the one-tailed test for the tail the data favor and multiplying the knots by two *if* none of the resulting knots exceeds one.

If some of the resulting knots exceed one, then we are in an uninteresting case from a practical standpoint (the data give essentially no evidence against the null hypothesis), but we should do the calculation correctly anyway.

One way to think of the relation between the one-tailed and two-tailed fuzzy P-values is that if P is the one-tailed fuzzy P-value, then $\min(2P, 1-2P)$ is the two-tailed fuzzy P-value. One easy case discussed above is P is concentrated below $1/2$ so the two-tailed fuzzy P value is $2P$. The other easy case discussed above is P is concentrated above $1/2$ so the two-tailed fuzzy P value is $1-2P$, which is twice the one-tailed fuzzy P-value for the other-tailed test.

The complications arise when the support of the distribution of P contains $1/2$, in which case the upper bound of the support of the two-tailed fuzzy P-value will be one. Suppose we have $l \leq u$ and P is the fuzzy P-value for the upper tailed test (if $l > u$, swap). Then P is the mixture of $\text{Unif}(a_k, b_k)$ random variables, where

$$\begin{aligned} a_k &= \Pr\{W < l + k\} \\ b_k &= \Pr\{W \leq l + k\} \end{aligned} \tag{1}$$

for $k = 0, \ldots, t$, and the mixing probabilities are $\Pr\{T = k\}$. When $l+k \geq n/2$, the endpoints (1), when doubled, are wrong for the two-tailed P-value (because one or both exceeds one). There are two cases two consider. When $l+k = n/2$, which means n must be even, we then have

$$1 - 2a_k = 2b_k - 1$$

and the entire mixing probability $\Pr\{T = k\}$ should be assigned to the interval $(2a_k, 1)$. When $l + k > n/2$, which can happen whether n is odd or even, we have

$$1 < 2a_k < 2b_k$$

but also

$$\begin{aligned} 1 - 2a_k &= 2b_{k^*} \\ 1 - 2b_k &= 2a_{k^*} \end{aligned}$$

for some $k^* < k$. In fact, we have

$$l + k^* = n - (l + k)$$

hence

$$k^* = n - 2l - k$$

and now the mixing probability assigned to the interval $(2a_{k^*}, 2b_{k^*})$ should be $\Pr\{T = k^*\} + \Pr\{T = k\}$.

1.3 Two-Sided Confidence Intervals

In principle we get a two-tailed fuzzy confidence interval by "inverting" the two-tailed fuzzy test. So there is nothing left to do. In practice, we want to apply some additional theory.

Let $X_{(i)}$, $i = 0, \ldots, n + 1$ be the order statistics, where $X_{(0)} = -\infty$ and $X_{(n+1)} = +\infty$. The result of the sign test does not change as μ changes within an interval between order statistics. Thus we only need calculate for each order statistic and for each interval between order statistics. Moreover, we can save some work using the following theorems.

Theorem 1 *If*

$$2 \Pr\{W < m\} < \alpha,$$

where $W \sim \mathrm{Bin}(n, 1/2)$*, then the membership function of the fuzzy confidence interval with coverage* $1 - \alpha$ *is zero for*

$$\mu < X_{(m)} \quad \text{or} \quad X_{(n-m+1)} < \mu.$$

If the only m satisfying the condition is $m = 0$, then the assertion of the theorem is vacuously true.

For $\mu < X_{(m)}$ we have at least $n - m + 1$ data values above μ. Since $m \leq n/2$ the latent test statistic is at least $n - m + 1$. Hence the fuzzy P-value has support bounded above by $2 \Pr(W \geq n - m + 1) = 2 \Pr(W < m) < \alpha$. Hence we accept this μ with probability zero, and the membership function of the fuzzy confidence interval is zero at this μ. The case $\mu > X_{(n-m+1)}$ follows by symmetry.

Theorem 2 *If*

$$\alpha \leq 2 \Pr\{W \leq m\},$$

where $W \sim \text{Bin}(n, 1/2)$*, then the membership function of the fuzzy confidence interval with coverage* $1 - \alpha$ *is one for*

$$X_{(m+1)} < \mu < X_{(n-m)}. \tag{2}$$

If the only m satisfying the condition have $m + 1 \geq n - m$, then the assertion of the theorem is vacuously true.

If (2) holds, then we have at least $m + 1$ data values below μ and at least $m + 1$ above. Hence we also have at most $n - m - 1$ data values below μ and at most $n - m - 1$ above, and the latent test statistic is at most $n - m - 1$. Hence the fuzzy P-value has support bounded below by $2\Pr(W > n-m-1) = 2\Pr(W \leq m) \geq \alpha$. Hence we accept this μ with probability one, and the fuzzy confidence interval is one at this μ.

Thus we see that if we chose m to satisfy

$$2 \Pr\{W < m\} < \alpha \leq 2 \Pr\{W \leq m\}, \tag{3}$$

where $W \sim \text{Bin}(n, 1/2)$, then we only need to calculate the membership function of the fuzzy confidence interval at the points $X_{(m)}$, $X_{(m+1)}$, $X_{(n-m)}$, and $X_{(n-m+1)}$, which need not be distinct, and on the intervals $(X_{(m)}, X_{(m+1)})$ and $(X_{(n-m)}, X_{(n-m+1)})$, if nonempty, on which the membership function is constant. Thus there are at most 6 numbers to calculate. Theorems 1 and 2 give the membership function everywhere else.

These six numbers are calculated by carrying out the fuzzy two-tailed test for the relevant hypothesized value of μ and then calculating $\Pr\{P > \alpha\}$, where P is the fuzzy P-value.

1.3.1 No Ties

Conventional theory says, when the distribution of the X_i is continuous, that the interval $(X_{(k)}, X_{(n+1-k)})$ is a $1 - 2\Pr\{W < k\}$ confidence interval for the true unknown population median, where $W \sim \text{Bin}(n, 1/2)$.

Suppose we want coverage $1 - \alpha$. Then either one of the intervals already discussed has coverage $1 - \alpha$ or there is a unique $\alpha/2$ quantile of the $\text{Bin}(n, 1/2)$ distribution. Call it m. Then

$$1 - 2\Pr\{W < m\} > 1 - \alpha > 1 - 2\Pr\{W \leq m\}$$

and the left hand side is the coverage of $(X_{(m)}, X_{(n+1-m)})$ and the right hand side is the coverage of $(X_{(m+1)}, X_{(n-m)})$. Thus a mixture of these two confidence intervals is a randomized confidence interval with coverage $1 - \alpha$. The corresponding fuzzy confidence interval has membership function that is 1.0 on $(X_{(m+1)}, X_{(n-m)})$, is γ on the part of $(X_{(m)}, X_{(n+1-m)})$ not in the narrower interval, and zero elsewhere, where γ is determined as follows. The coverage of this interval is

$$\gamma\bigl[1 - 2\Pr\{W < m\}\bigr] + (1 - \gamma)\bigl[1 - 2\Pr\{W \leq m\}\bigr]$$

Setting this equal to $1 - \alpha$ and solving for γ gives

$$\gamma = \frac{2 \Pr\{W \le m\} - \alpha}{2 \Pr\{W = m\}} \tag{4}$$

and the condition that m is a unique $\alpha/2$ quantile of W guarantees $0 < \gamma < 1$.

Note that this discussion agrees with Theorems 1 and 2. We have simply obtained more information. Now we know that the membership function of the fuzzy confidence interval is γ on the intervals $(X_{(m)}, X_{(m+1)})$ and $(X_{(n-m)}, X_{(n-m+1)})$. Under the assumption that the distribution of the X_i is continuous, the values at points does not matter because points occur with probability zero.

Caveat The above discussion is predicated on $m + 1 < n - m$, which is the same as $m < (n-1)/2$. This can only fail for ridiculously small coverage probabilities. If $m \ge (n-1)/2$, then either $m = (n-1)/2$ when n is odd or $m = n/2$ when n is even, and

$$1 - \alpha < 1 - 2 \Pr\{W < m\} = \begin{cases} 2 \Pr\{W = (n-1)/2\}, & n \text{ odd} \\ \Pr\{W = n/2\}, & n \text{ even} \end{cases}$$

and either $1 - \alpha$ is very small or n is very small or both.

In this case our fuzzy confidence interval is zero outside the closure of the interval $(X_{(m)}, X_{(n-m+1)})$ and is γ on this interval, where the coverage is now

$$1 - \alpha = \gamma\bigl[1 - 2 \Pr\{W < m\}\bigr]$$

so γ is now given by

$$\gamma = \frac{1 - \alpha}{1 - 2 \Pr\{W < m\}} \tag{5}$$

and the "core" of the fuzzy confidence interval (the set on which its membership function is one) is empty.

There is a somewhat more complicated form that manages to be either (4) or (5), whichever is valid

$$\gamma = \frac{\Pr\{W \le m \text{ or } W \ge n - m\} - \alpha}{\Pr\{W = m \text{ or } W = n - m\}} \tag{6}$$

1.3.2 With Ties

We claim that the same solution works with ties, except that when ties are possible we must be careful about how the membership function of the fuzzy interval is defined at jumps. We claim the fuzzy interval still has the same form with membership function

$$I(\mu) = \begin{cases} 0 & \mu < X_{(m)} \\ \beta_1 & \mu = X_{(m)} \\ \gamma & X_{(m)} < \mu < X_{(m+1)} \\ \beta_2 & \mu = X_{(m+1)} \\ 1 & X_{(m+1)} < \mu < X_{(n-m)} \\ \beta_3 & \mu = X_{(n-m)} \\ \gamma & X_{(n-m)} < \mu < X_{(n-m+1)} \\ \beta_4 & \mu = X_{(n-m+1)} \\ 0 & \mu > X_{(n-m+1)} \end{cases} \tag{7}$$

where m is chosen so that (3) holds.

Any or all of the intervals on which $I(\mu)$ is nonzero may be empty either because of ties or, as mentioned in the "caveat" in the preceding section because $m+1 \geq n-m$. Any or all of the betas may also be forced to be equal because the order statistics they go with are tied. We may have $X_{(m)} = -\infty$ and $X_{(n-m+1)} = +\infty$, in which case the corresponding betas are irrelevant.

Theorem 3 *When* (7) *is as described above,* γ *in* (7) *is given by* (6).

If both of the intervals $(X_{(m)}, X_{(m+1)})$ and $(X_{(n-m)}, X_{(n-m+1)})$ are empty, then the theorem is vacuous.

For $X_{(m)} < \mu < X_{(m+1)}$ we have exactly m data values below μ and exactly $n-m$ above, and the latent test statistic is $n-m$. The fuzzy P-value is uniform on the interval with endpoints $2\Pr\{W < m\}$ and $2\Pr\{W \leq m\}$ in case $m+1 < n-m$ and otherwise uniform on the interval with endpoints $2\Pr\{W < m\}$ and one. The lower endpoint is the same in both cases, and we can write the upper endpoint as $\Pr\{W \leq m \text{ or } W \geq n-m\}$ in both cases. Hence the probability the fuzzy P-value is greater than α is given by (6), and this is the value of $I(\mu)$ for this μ. The case $X_{(n-m)} < \mu < X_{(n-m+1)}$ follows by symmetry.

Now we know every thing about the fuzzy confidence interval except for the betas, which can be determined by inverting the fuzzy test (as can the value at all points), so now we are down to inverting the test at at most four points $X_{(m)}$, $X_{(m+1)}$, $X_{(n-m)}$, and $X_{(n-m+1)}$, any or all of which may be tied. When they are tied, there is no simple formula for the corresponding beta (the simplest description is the one just given: invert the fuzzy test, the value is one minus the fuzzy decision). Thus we make no attempt at providing such a formula for the general case.

However, we can say a few things about the betas.

Theorem 4 *The fuzzy confidence interval given by* (7) *is convex.*

Convexity of a fuzzy set with membership function I is the property that all of the sets $\{\, \mu : I(u) \geq u \,\}$ are convex.

First consider the case when $X_{(m+1)} < X_{(n-m)}$ so there are points μ where $I(\mu) = 1$. Then, since all of the betas are probabilities, convexity only requires $\beta_1 \le \gamma \le \beta_2$ if $X_{(m)} < X_{(m+1)}$ and $\beta_4 \le \gamma \le \beta_3$ if $X_{(n-m)} < X_{(n-m+1)}$, and the latter follows from the former by symmetry.

So consider $X_{(m)} = \mu < X_{(m+1)}$. Say we have l, t, and u data values below, tied with, and above μ, respectively. Then we have $l+t = m$ and $u = n-m$ and $t \ge 1$. The latent test statistic is $n - m + T$, where $T \sim \text{Bin}(t, 1/2)$. The CDF of the fuzzy P-value has knots $2\Pr\{W > n - m + t - k\}$ and values $\Pr\{T < k\}$, where $k = 0, \ldots, t+1$. We can rewrite the knots $2\Pr\{W < m-t+k\}$. The two largest are $2\Pr\{W < m\}$ and $2\Pr\{W \le m\}$, which bracket α. The probability of accepting μ is thus the probability that a uniform on this interval is greater than α, which is γ given by (4) times $\Pr\{T = t\}$. Hence we do have $\beta_1 \le \gamma$.

Then consider $X_{(m)} < \mu = X_{(m+1)}$. With l, t, and u as above, we have $l = m$ and $u + t = n - m$ and $t \ge 1$. Now the two smallest knots are $2\Pr\{W < m\}$ and $2\Pr\{W \le m\}$. The probability of accepting μ is now

$$\beta_2 = \Pr\{T > 0\} + \gamma \cdot \Pr\{T = 0\}$$

so $\beta_2 > \gamma$. That finishes the case $X_{(m+1)} < X_{(n-m)}$.

We turn now to the case $X_{(m+1)} = X_{(n-m)}$. We conjecture (still to be proved) that $\beta_2 = \beta_3$ is the maximum of the membership function, in which case convexity requires $\beta_1 \le \gamma \le \beta_2$ if $X_{(m)} < X_{(m+1)}$ and $\beta_4 \le \gamma \le \beta_3$ if $X_{(n-m)} < X_{(n-m+1)}$. But we have already proved these, because the proofs above did not assume anything about whether $X_{(m+1)}$ was equal or unequal to $X_{(n-m)}$. Thus the only issue is whether $\beta_2 = \beta_3$ is the maximum.

If $X_{(m)} < X_{(m+1)}$, then we have proved $\beta_1 \le \gamma \le \beta_2$ which implies that the maximum does not occur to the left of $\beta_2 = \beta_3$. If $X_{(m)} = X_{(m+1)}$, then we have $\beta_1 = \beta_2 = \beta_3$, which also implies that the maximum does not occur to the left of $\beta_2 = \beta_3$. By symmetry, the maximum also does not occur to the right of $\beta_2 = \beta_3$. That finishes the case $X_{(m+1)} = X_{(n-m)}$.

We turn now to the case $X_{(m+1)} > X_{(n-m)}$, in which case we must have $m = n - m$, $\beta_1 = \beta_3$, and $\beta_2 = \beta_4$. There are two cases to consider. If $X_{(m)} < X_{(m+1)}$, then we conjecture that γ is the maximum of the membership function, in which case convexity requires $\beta_1 \le \gamma$ and $\beta_4 \le \gamma$, but these have already been proved. If $X_{(m)} = X_{(m+1)}$, then $\beta_1 = beta_2 = \beta_3 = \beta_4$ is the only nonzero value if $I(\mu)$ and convexity holds trivially. This finishes the proof of Theorem 4.

We can refine the calculations above in the most common case.

Theorem 5 *When there are no ties among the data, the membership function* (7) *is equal to the average of its left and right limits.*

This means

$$\beta_1 = \beta_4 = \gamma/2 \tag{8a}$$

$$\beta_2 = \beta_3 = \gamma/2 + 1/2 \tag{8b}$$

in the case $m + 1 < n - m$. When $m + 1 = n - m$, we still have (8a), but (8b) is replaced by $\beta_2 = \beta_3 = \gamma$. When $m = n - m$, we still have (8a), but (8b) is replaced by $\beta_1 = \beta_2 = \beta_3 = \beta_4$.

When there are no ties and $X_{(m)} = \mu$, then we have $m - 1$ data values below, one tied with, and $n - m$ above μ, respectively. The latent test statistic is $n - m + T$, where $T \sim \text{Bin}(1, 1/2)$. The fuzzy P-value is uniform on the interval with endpoints $2 \Pr\{W < m - 1\}$ and $2 \Pr\{W \le m\}$, and the probability of accepting μ, which is $\gamma \cdot \Pr\{T = 1\}$, is $\gamma/2$, because T is Bernoulli.

When there are no ties and $X_{(m+1)} = \mu$ and $m + 1 < n - m$, the fuzzy P-value is uniform on the interval with endpoints $2 \Pr\{W < m\}$ and $2 \Pr\{W \le m + 1\}$, and the probability of accepting μ, which is $\Pr\{T > 0\} + \gamma \cdot \Pr\{T = 0\}$ is $\gamma/2 + 1/2$.

When there are no ties and $X_{(m+1)} = \mu$ and $m + 1 = n - m$, the fuzzy P-value is uniform on the interval with endpoints $2 \Pr\{W < m\}$ and one, and the probability of accepting μ is is $(1 - \alpha)/[1 - 2 \Pr\{W < m\}]$, which is (5).

When there are no ties and $X_{(m+1)} = \mu$ and $m = n - m$, there is nothing to prove because $\beta_1 = \beta_3$ and $\beta_2 = \beta_4$ follow from $m = n - m$. This finishes the proof of Theorem 5.

1.4 One-Sided Confidence Intervals

From the preceding, one-sided intervals should now be obvious. We merely record a few specific details. A lower bound interval has the form

$$I(\mu) = \begin{cases} 0 & \mu < X_{(m)} \\ \beta_1 & \mu = X_{(m)} \\ \gamma & X_{(m)} < \mu < X_{(m+1)} \\ \beta_2 & \mu = X_{(m+1)} \\ 1 & X_{(m+1)} < \mu < \infty \end{cases} \tag{9}$$

where m is chosen so that

$$\Pr\{W < m\} < \alpha \le \Pr\{W \le m\} \tag{10}$$

and

$$\gamma = \frac{\Pr\{W \le m\} - \alpha}{\Pr\{W = m\}} \tag{11}$$

Theorem 4 and Theorem 5 still apply and the betas can still be determined by inverting the fuzzy test.

2 The Rank Sum Test

Now we have two samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$. The test is based on the differences

$$Z_{ij} = X_i - Y_j, \qquad i = 1, \ldots, m, \ j = 1, \ldots, n.$$

Let $Z_{(i)}$, $i = 1, \ldots, mn$ be the order statistics of the Z_{ij}. If we assume (1) the two samples are independent, (2) each sample is IID, (3) the distribution of the X_i is continuous, and (4) the distribution of the Y_j is the same as the distribution of the X_i except for a shift, then (i) the marginal distribution of the Z_{ij} is symmetric about the shift μ and (ii) the number of Z_{ij} less than μ has the distribution of the Mann-Whitney form of the test statistic for the Wilcoxon rank sum test.

Thus all of the theory is nearly the same as in the preceding section but with the following differences.

- The $Z_{(i)}$ now play the role formerly played by the $X_{(i)}$.
- The Mann-Whitney distribution now plays the role formerly played by the binomial distribution.
- We have Z_{ij} tied with μ when X_i is tied with $Y_j + \mu$. Infinitesimal jittering only breaks ties within one class of tied X_i and $Y_j + \mu$ values. The number of $Z_{ij} < \mu$ coming from such a class with m_k of the X_i tied with n_k of the $Y_j + \mu$ has the Mann-Whitney distribution for sample sizes m_k and n_k. The total number is the sum over each tied class.

2.1 One-Tailed Tests

Let $\text{MannWhit}(m, n)$ denote the Mann-Whitney distribution, the distribution of the number of negative Z_{ij} when there are no ties. The range of this distribution is zero to $N = mn$. It is symmetric, with center of symmetry $N/2$. Let $W \sim \text{MannWhit}(m, n)$.

If there are no ties and the observed value of the Mann-Whitney statistic is w, then the fuzzy P-value for a upper-tailed test is uniformly distributed on the interval with endpoints $\Pr\{W < w\}$ and $\Pr\{W \leq w\}$.

If there are ties, then let m_k and n_k be the number of X_i and the number of $Y_j + \mu$ tied at the k-th tie value, let $T_k \sim \text{MannWhit}(m_k, n_k)$, and let

$$T = T_1 + \ldots + T_K,$$

where K is the number of distinct tie values. There is no "brand name" for the distribution of T. The Mann-Whitney distribution obviously does not have an "addition rule" like the binomial: $\text{MannWhit}(m_1 + \ldots + m_K, n_1 + \cdots + n_K)$ is *clearly not* the distribution of T. We know the distribution of the T_k. We must simply calculate the distribution of T by brute force and ignorance (BFI) using the convolution of probability vectors. T ranges from zero to

$$t = \sum_{i=1}^{K} m_k n_k$$

Denote its distribution function by G.

Now let l be the number of Z_{ij} less than μ. The latent test statistic is $l + T$. The fuzzy P-value has a continuous, piecewise linear CDF with knots at $\Pr\{W < l + k\}$ and values $G(k)$ for $k = 0, \ldots, t + 1$.

The lower-tailed test is the same (merely swap X's and Y's or use the symmetry of the Mann-Whitney distribution and replace l by $N - l - t$).

2.2 Two-Tailed Tests

The problem with two-tailed tests is much the same as we saw with the sign test. If P is the one-tailed fuzzy P-value, then the two-tailed fuzzy P-value is $\min(2P, 1 - 2P)$.

2.3 Confidence Intervals

Everything is the same as with the sign test. Merely replace the $X_{(i)}$ there with the $Z_{(i)}$ here and replace the binomial distribution with the Mann-Whitney distribution.

References

Geyer, C. J. (submitted). Fuzzy P-values and ties in nonparametric tests. `http://www.stat.umn.edu/geyer/fuzz/ties.pdf`