Fuzzy Rank Tests and Confidence Intervals

Charles J. Geyer

May 9, 2026

Abstract

How to do exact-exact (rather than only conservative-exact) sign, signrank, and ranksum hypothesis tests, whether or not there are tied ranks. Also how to do the corresponding confidence intervals.

Exact-exact procedures must be either randomized or fuzzy. This package provides the latter.

1 License

2 R

- The version of R used to make this document is 4.6.0.
- The version of the `knitr` package used to make this document is 1.51.
- The version of the `fuzzyRankTests` package used to make this document is 0.5.

```
library(fuzzyRankTests)
```

3 Introduction

3.1 What This is About

We deal with three tests of statistical hypotheses:

- the sign test,
- Wilcoxon's signed rank test, and
- Wilcoxon's rank sum test (also called Mann-Whitney).

And we deal with two issues with these.

- Like all tests with discrete test statistics, exact tests are impossible unless the test is randomized.
- Tied data and tied ranks complicate the situation.

Assumptions:

- One Sample or Paired Comparison
  - Sign test: no assumptions.
  - Signed rank test: symmetric population distribution.
  - t test: normal population distribution.
- Two Independent Samples
  - Rank sum test: one population distribution is the other shifted.
  - t test: both population distributions normal with same variance.

This package does not do t tests, see R function `t.test` in core R for that. We only include them to show that the assumptions get more restrictive as one goes down the list.

For non-fuzzy tests the assumptions above need an additional assumption that the population distribution is continuous so there are no tied data or tied ranks. As will be seen, fuzzy tests and confidence intervals do not need this assumption.

3.2 Fuzzy Tests and Confidence Intervals

Despite being the official theory of testing statistical hypotheses since it was invented by Neyman and Pearson in the 1930's (Lehmann and Romano, 2022, Chapters 3 and 4) and despite being taught to all PhD statistics students, the theory of randomized hypothesis tests gets little application (I have never seen it used) because of the arbitrariness of the artificial randomization. Two statisticians can analyze exactly the same data using exactly the same hypothesis test and come to opposite decisions due to the artificial randomization.

Geyer and Meeden (2005) proposed a simple fix for this issue: "unrandomize" randomized tests in the sense that one reports not a decision or a P-value or a confidence interval that purports to be a realization of some random process (the artificial randomness in the hypothesis test) but rather report (a description of) the probability distribution of that random quantity. That is we report *abstract* randomness rather than *realized* randomness.

In more detail, a randomized hypothesis test rejects the null hypothesis with probability $\phi(X)$ when test statistic X is observed. This function ϕ is called the *critical function* of the test. Geyer and Meeden (2005) point out that the critical function also depends on the significance level α and the value of the parameter hypothesized under the null hypothesis (for one-tailed tests, the boundary point of the composite null hypothesis). So they write the critical function $\phi(x, \alpha, \theta)$. And they say the result of the test is to report this critical function, not some realization of some random variable related to it.

Geyer and Meeden (2005) go on to point out three different interpretations of the critical function.

- The function $\phi(\,\cdot\,, \alpha, \theta)$ is the critical function of the randomized test, as considered classically.
- The function $\phi(x, \,\cdot\,, \theta)$ is the (distribution function of) the abstract randomized (also called *fuzzy*) P-value of the randomized test.
- The function $1 - \phi(x, \alpha, \theta)$ is the (membership function of) the *fuzzy confidence interval*) that is dual to the randomized test.

There is no difference between $\phi(x)$ used classically and $\phi(x, \alpha, \theta)$ used by Geyer and Meeden (2005) when considered as a function of x for fixed α and θ. It is the same function of x either way. Geyer and Meeden (2005) say what one should report is the number $\phi(x, \alpha, \theta)$ rather than a decision (accept or reject the null hypothesis that purportedly has this number as its probability of rejection).

In order for the function $\phi(x, \,\cdot\,, \theta)$ to be a distribution function, the hypothesis test need only have nested critical regions (Geyer and Meeden, 2005, equation (1.4) and the surrounding discussion) and be continuous (which property our applications have). If we were to generate a random variable P having this distribution function, then rejecting the null hypothesis when $P < \alpha$ is the classical randomized test. Hence this is the P-value of that test. Geyer and Meeden (2005) are only saying that rather than

simulating such a P and reporting that number, one should report its distribution as described by the distribution function $\phi(x, \cdot, \theta)$ or perhaps by the probability density function of that distribution function.

The function $1-\phi(x, \alpha, \cdot)$ takes value between zero and one, including (if the test is actually randomized) values strictly between zero and one. Geyer and Meeden (2005) suggest we interpret this as the membership function of a fuzzy set, as in fuzzy set theory (Klir, St. Clair, and Yuan, 1997). One interprets the membership function as saying to what degree the point is in the fuzzy set. Geyer and Meeden (2005) say one should interpret it like partial credit on a test question. After all, that is what probability does. The coverage probability of the interval is

$$E_\theta\{1 - \phi(X, \alpha, \theta)\} = 1 - \alpha$$

and this means point x is being given "partial credit" $1 - \phi(x, \alpha, \theta)$ when θ is the true unknown parameter value.

3.3 Tied Data or Tied Ranks

Tied data (data points tied with the hypothesized value under the null hypothesis) or tied ranks (for the signed rank test or for the rank sum test) bring more issues. We deal with these using the methods of Thompson and Geyer (2007).

Now our model has data in two parts: the observable part x and the unobservable part y (also called missing data, latent variables, random effects, or hidden layer). So we write the critical function of our randomized test $\psi(x, y, \alpha, \theta)$. Then

$$\phi(x, \alpha, \theta) = E_\theta\{\psi(x, Y, \alpha, \theta)\} \tag{1}$$

is the critical function for the test based on the observed data x.

3.4 What this Package Does

For all three hypothesis tests this package does, the null distribution of the test statistic is discrete and symmetric. Let T be the test statistic for an upper tailed test and τ be the center of symmetry of its null distribution. Then $-T$ is the test statistic for the lower tailed test, and $|T - \tau|$ is the test statistic for the two-tailed test.

In all three cases, the fuzzy P-value is uniformly distributed on the interval with endpoints $\text{pr}_\theta(W > w)$ and $\text{pr}_\theta(W \geq w)$, where W is the test statistic considered as a random variable and w is its observed value.

Hence the critical function of the test is

$$\phi(w, \alpha, \theta) = \begin{cases} 0, & \alpha \le \text{pr}_\theta(W > w) \\ \frac{\alpha - \text{pr}_\theta(W > w)}{\text{pr}_\theta(W = w)}, & \text{pr}_\theta(W > w) < \alpha < \text{pr}_\theta(W \ge w) \\ 1, & \text{pr}_\theta(W \ge w) \le \alpha \end{cases}$$

when there are no ties in the data or the ranks.

When there are ties in the data or the ranks, we assume the data have been measured with inadequate precision. If more precise measurement had been used there would be no ties in the data or the ranks. We assume that all orderings of the hypothetical precise data consistent with the observed (imprecise) data are equiprobable (since there is no data favoring any such ordering).

Thus the critical function when there are ties is just the average of the critical functions (1) for the precise data (with no ties) consistent with the observed imprecise data.

3.5 Ordered Categorical Data

We do not recommend the procedures in this package as competitors for procedures for ordered categorical data (Agresti, 2013, Sections 8.2 and 8.3). If one has ordered categorical response data, then one should probably use statistical models and procedures designed specifically for that.

But if the ordered categories have arisen from imprecise measurement, then one could also justify using the fuzzy procedures this package provides for such data.

3.6 Other Procedures for Tied Data or Tied Ranks

We take Hollander, Wolfe, and Chicken (2014) to be authoritative about existing practice.

3.6.1 Sign Test

For the sign test, their recommended procedure is to report the usual P-value for a discrete test: $\text{pr}_\theta(W \ge w)$ when there are no ties (data values equal to the value hypothesized by the null hypothesis).

When there are ties, Hollander, et al. (2014, Subsection Ties of Section 3.4) say one should eliminate the ties from the data and then proceed as above.

We say this is unacceptable. It is cherry-picking data that favor the alternative hypothesis (suppressing data that favor the null hypothesis). This correction for ties, although widely used, can never be justified.

To be fair to Hollander, et al. (2014) they say (Comment 34 of Section 3.4) that one should not do their recommended procedure when the number of ties "represent a sizable percentage of the total." So they already recognize the wrongness. They also give two other procedures.

- A randomized procedure that is what we "unrandomize" turning it into a fuzzy P-value. They do not like randomized procedures and hence do not recommend them. But we do not either. Hence the unrandomization, which escapes their criticism.

- A conservative procedure that counts all ties in favor of the null hypothesis. Our procedure also calculates this: its P-value is the upper endpoint of the support of the distribution of our fuzzy P-value. So we take that into account (including exactly how conservative it is).

3.6.2 Signed Rank Test

This section is much like the preceding one *mutatis mutandis*. The issues surrounding exactness and ties are much the same. Ranks bring in a few technical details, which we do not need to emphasize because the computer does all the work dealing with them.

For the signed rank test, the recommended procedure of Hollander, et al. (2014, Section 3.1) is to report the usual P-value for a discrete test: $\text{pr}_\theta(W \geq w)$ when there are no ties (either data values equal to the value hypothesized by the null hypothesis or tied ranks).

When there are ties, Hollander, et al. (2014, Subsection Ties of Section 3.1) say one should (i) eliminate data values equal to the value hypothesized by the null hypothesis and (ii) use average ranks when there are tied ranks. Using average ranks changes the null distribution of the test statistic to something not easily understood, so one uses the asymptotic normal distribution of the test statistic under the null hypothesis, which has its asymptotic variance corrected for ties.

We say (i) is unacceptable. It is cherry-picking data that favor the alternative hypothesis (suppressing data that favor the null hypothesis). Although widely used, it can never be justified.

We also do not need (ii) because we use unrandomized randomized tests (Section 3.4 above) instead.

To be fair to Hollander, et al. (2014) they say (Comments 9 and 10 of Section 3.1) that one should not do their recommended procedure unless the "zero values are a very small percentage" of the total. So they already recognize the wrongness. They also give two other procedures.

- A randomized procedure that is what we "unrandomize" turning it into a fuzzy P-value. They do not like randomized procedures and hence do not recommend them. But we do not either. Hence the unrandomization, which escapes their criticism.

- A conservative procedure that counts all ties in favor of the null hypothesis. Our procedure also calculates this: its P-value is the upper endpoint of the support of the distribution of our fuzzy P-value. So we take that into account (including exactly how conservative it is).

They also discuss (Comment 11 of Section 3.1) another procedure that keeps the tied ranks but uses intensive computation to calculate the exact permutation distribution conditioning on the pattern of ties. Since we have an alternative, we are not interested in this either.

3.6.3 Rank Sum Test

For some reason, the discussion in Hollander, et al. (2014) of this test is not parallel to the other two. They do not discuss randomized versions of this test, although they obviously exist and work just as well as for the other two. Hence this package does the fuzzy hypothesis tests and confidence intervals that are justified in the same way as for the other two procedures.

4 Examples

4.1 Sign Test

4.1.1 No Zero Values

For an example with no zero values, we do Example 3.5 in Hollander, et al. (2014)

```
z <- c(-0.8, 7.5, 46.9, 17.6, -4.6, 54.0, 48.3, 3.9, 16.7,
    19.7, -8.5, 7.1, 40.7, 23.8, 14.8, 20.6, 25.0, 24.7,
    -1.8, 21.9, 4.7, 24.7, 52.8, 8.5, 1.9)
fuzzy.sign.test(z, alternative = "greater")
```

```
##
##  sign test
##
## data:  z
## alternative hypothesis: true mu is greater than 0
##
## fuzzy P-value has continuous, piecewise linear CDF with knots and
## values
##
##      knots values
##  7.826e-05      0
##  4.553e-04      1
```

Since (the support of the distribution of) the fuzzy P-value is far below common criteria of statistical significance, this is strong evidence against the null hypothesis. Note that the upper endpoint of the support of (the distribution of) the fuzzy P-value is the conventional P-value given by Hollander, et al. (2014).

A 95% fuzzy confidence interval for the median difference is given by

```
fuzzy.sign.ci(z) |> plot()
```

Figure 1 shows (the membership function of) this fuzzy confidence interval. Although we say this example has no ties, that means it has no ties at the hypothesized value under the null hypothesis, which in this case is zero. It does have ties at the upper endpoint of the support of the fuzzy confidence interval, which affects the value at that point.

4.1.2 With Zero Values

For an example with zero values, we make up some data.

```
z <- c(-1.3, -0.4, 0.0, 0.0, 0.3, 0.5, 0.9, 1.1, 1.1, 1.1, 2.3,
    2.5, 3.1, 4.5, 5.5)
fuzzy.sign.test(z)

##
##  sign test
##
## data:  z
```

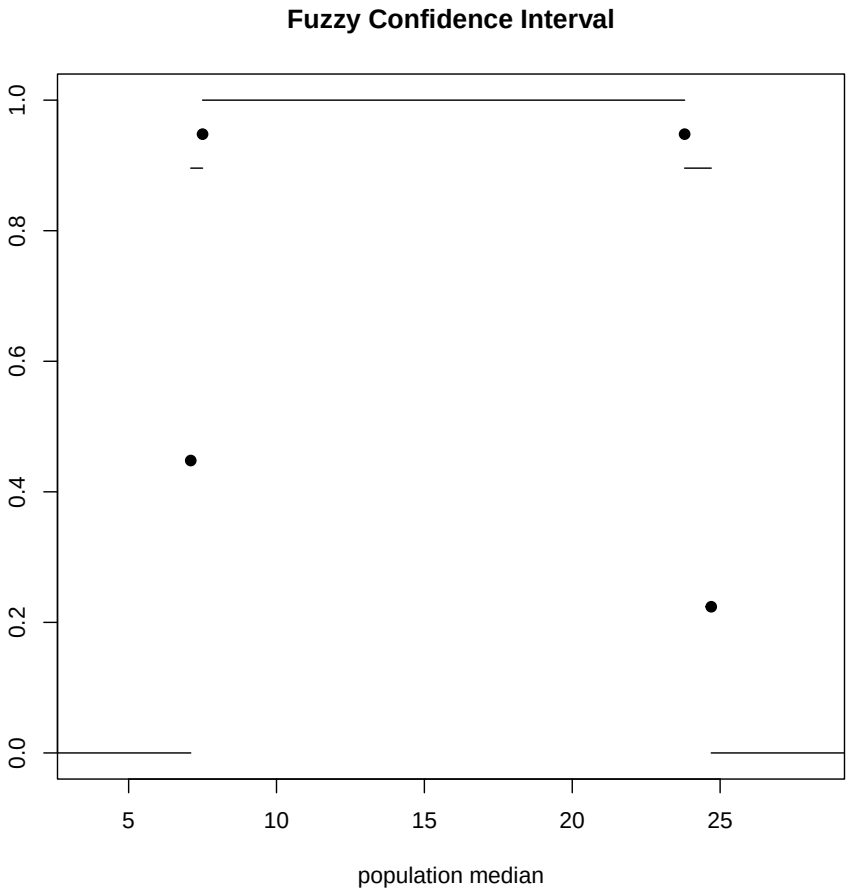

Figure 1: 95% fuzzy confidence interval for Example 3.5 of Hollander et al. (2014). Interval dual to sign test.

```
## alternative hypothesis: true mu is not equal to 0
##
## fuzzy P-value has continuous, piecewise linear CDF with knots and
## values
##
##      knots values
##  0.0009766   0.00
##  0.0073853   0.25
##  0.0351563   0.75
##  0.1184692   1.00
```

This might be called borderline statistically significant. It is equivocal.

We can plot the probability density function (Figure 2).

```
fuzzy.sign.test(z) |> plot()
```

Or we can plot the cumulative distribution function (Figure 3).

```
fuzzy.sign.test(z) |> plot(type = "cdf")
```

It is left as an exercise for the reader, if he or she is interested, to remove the zeroes from the data and redo, and then try to defend those results. (We do not think any defense can be valid.)

The interpretation of the PDF (Figure 2) is that the area under the curve to the left of α is the probability the null hypothesis is rejected at level α.

The interpretation of the CDF (Figure 3) is that the height of the curve at α is the probability the null hypothesis is rejected at level α.

The 95% fuzzy confidence interval is Figure 4.

```
fuzzy.sign.ci(z) |> plot()
```

4.2 Signed Rank Test

Again, to illustrate the issues with ties, we just make up some data. Figure 5 is the PDF of the fuzzy P-value.

```
z <- c(-2.2, -1.3, -0.3, 0.0, 0.0, 0.3, 0.5, 0.9, 1.1, 1.3,
    1.3, 2.3, 2.5, 3.1, 4.5, 5.5)
fuzzy.signrank.test(z) |> plot()
```

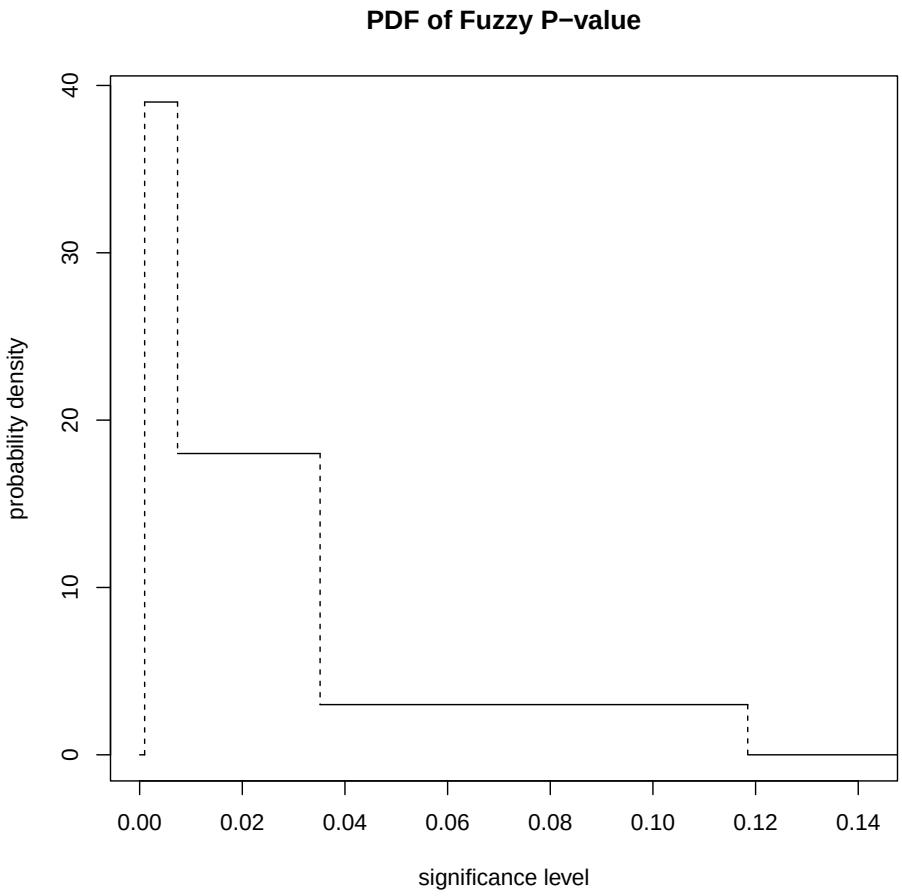

Figure 2: PDF of Fuzzy P-value.

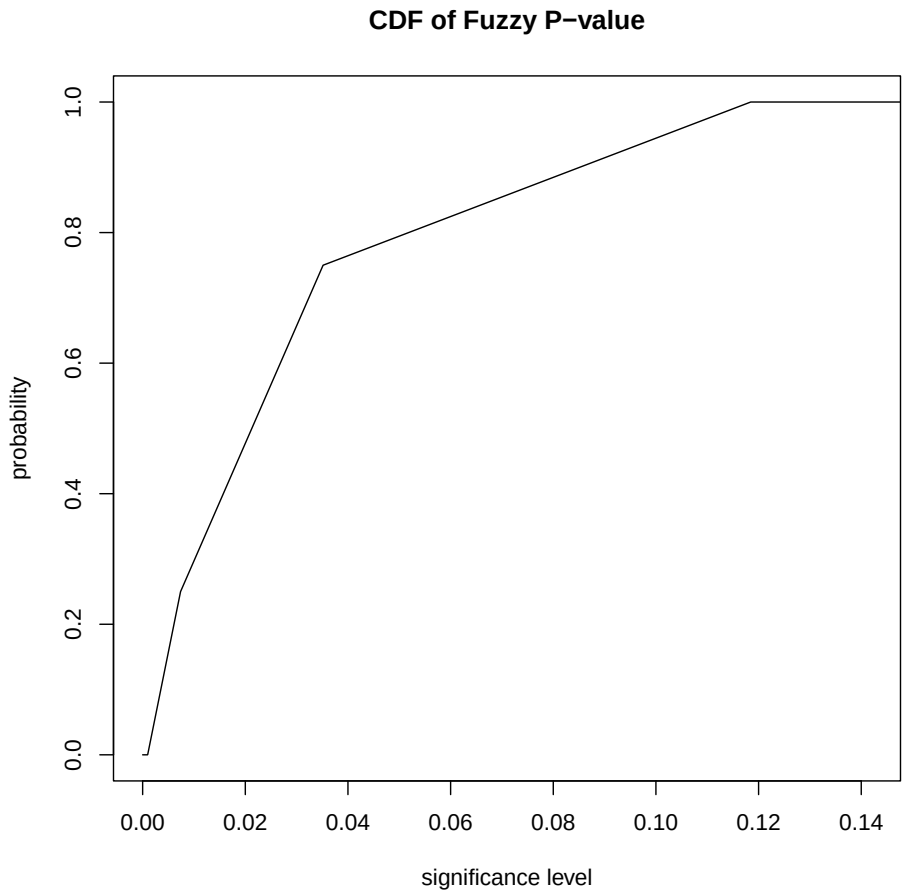

Figure 3: CDF of Fuzzy P-value.

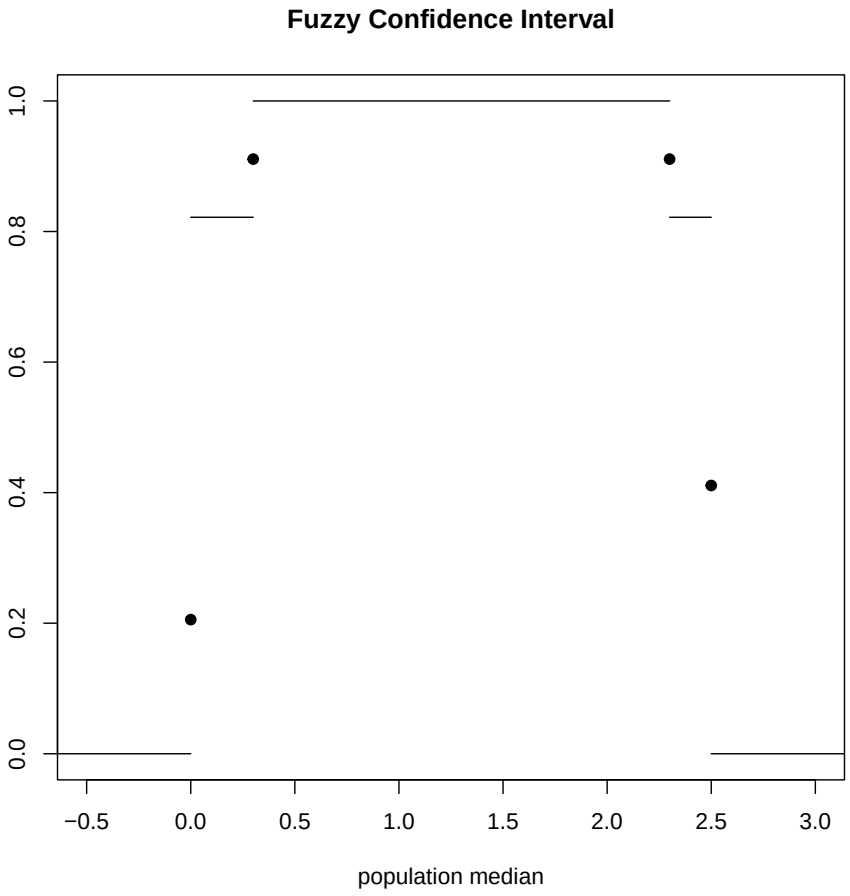

Figure 4: 95% fuzzy confidence interval for made-up data with ties. Interval dual to sign test.

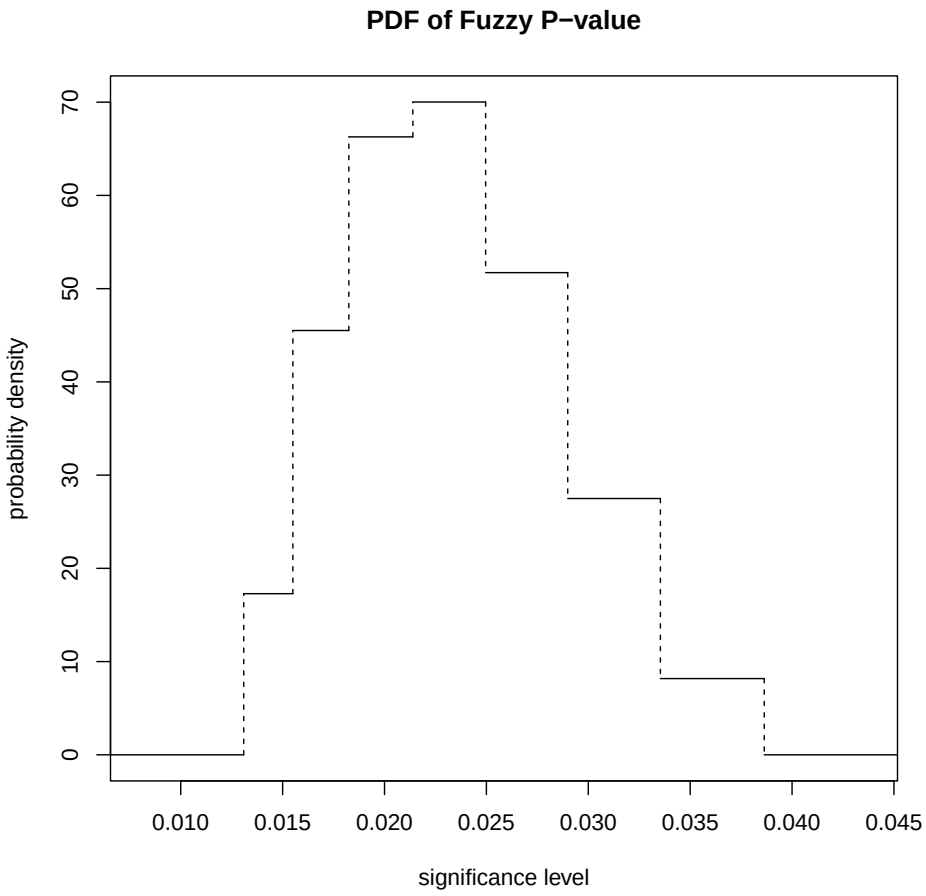

Figure 5: Signed rank test for made-up data.

And Figure 6 is the CDF of the fuzzy P-value.

```
fuzzy.signrank.test(z) |> plot(type = "cdf")
```

And Figure 7 is (the membership function of) the 95% fuzzy confidence interval.

```
fuzzy.signrank.ci(z) |> plot()
```

4.3 Rank Sum Test

Again, to illustrate the issues with ties, we just make up some data. Figure 8 is the PDF of the fuzzy P-value.

```
x <- c(1, 2, 3, 4, 4, 4, 5, 6, 7)
y <- c(4, 5, 7, 7, 8, 9, 10, 11)
fuzzy.ranksum.test(x, y) |> plot()
```

And Figure 9 is (the membership function of) the 95% fuzzy confidence interval.

```
fuzzy.ranksum.ci(x, y) |> plot()
```

References

Agresti, A. (2013). *Categorical Data Analysis*, third edition. John Wiley & Sons, Hoboken, NJ.

Geyer, C. J. and Meeden, G. D. (2005). Fuzzy and randomized confidence intervals and P-values (with discussion). *Statistical Science*, **20**, 358–387. doi:10.1214/088342305000000340.

Hollander, M., Wolfe, D. A., and Chicken, E. (2014). *Nonparametric Statistical Methods*, third edition. John Wiley & Sons, Hoboken, NJ.

Klir, G. J., St. Clair, U. H., and Yuan, B. (1997). *Fuzzy Set Theory: Foundations and Applications*. Prentice Hall, Upper Saddle River, NJ.

Lehmann, E. L., and Romano, J. P. (2022). *Testing Statistical Hypotheses*, fourth edition. Springer, Cham.

Thompson, E. A. and Geyer, C. J. (2007). Fuzzy P-values in latent variable problems. *Biometrika*, **94**, 49–60. doi:10.1093/biomet/asm001.

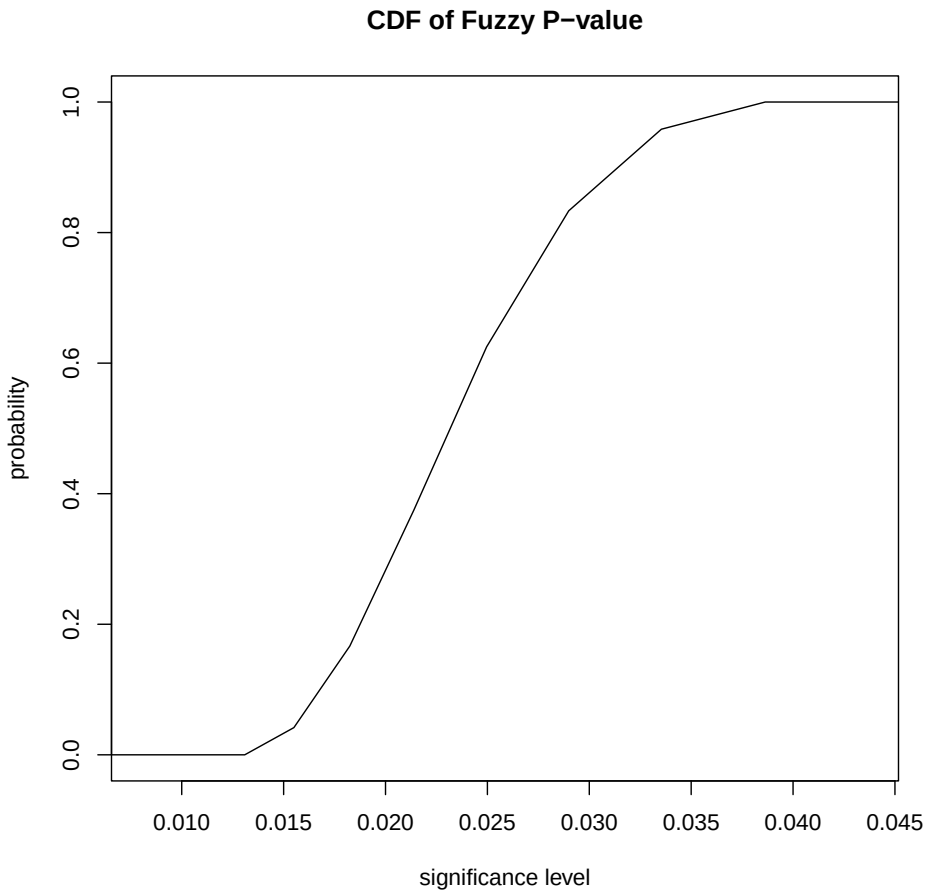

Figure 6: Signed rank test for made-up data.

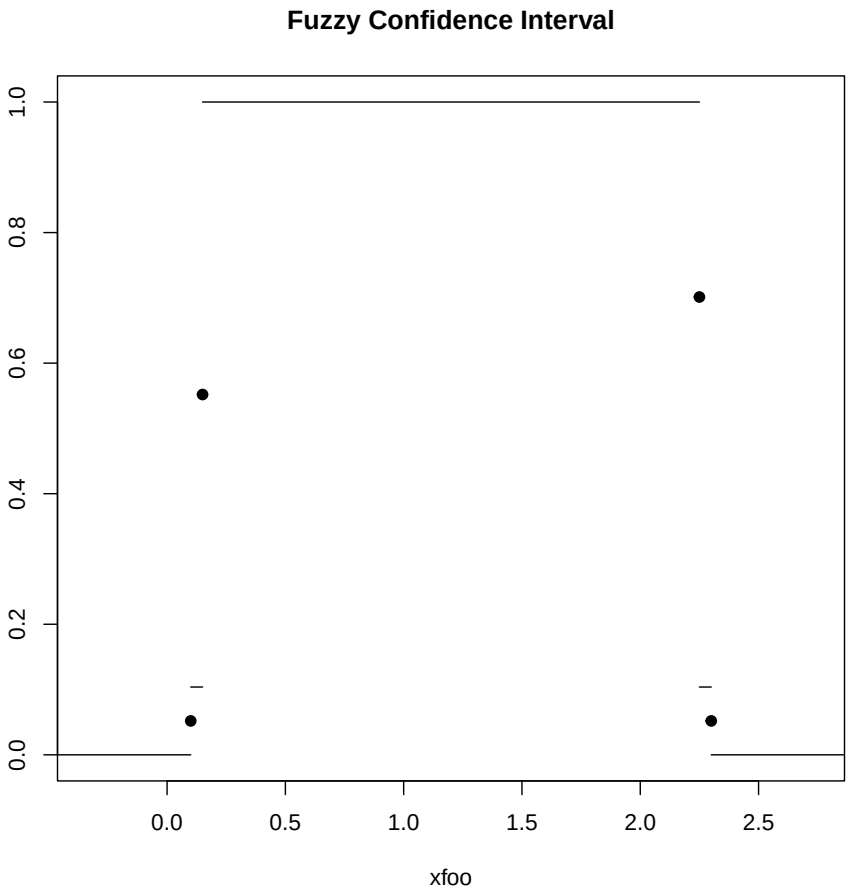

Figure 7: 95% Signed rank confidence interval for made-up data.

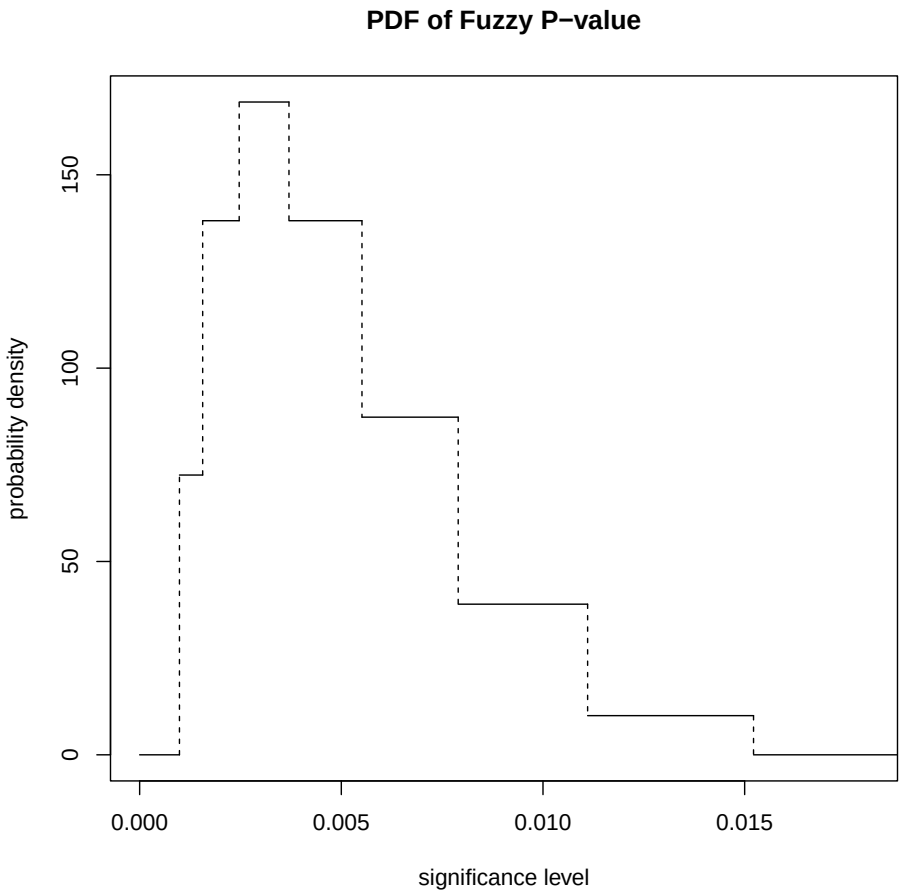

Figure 8: Rank sum test for made-up data.

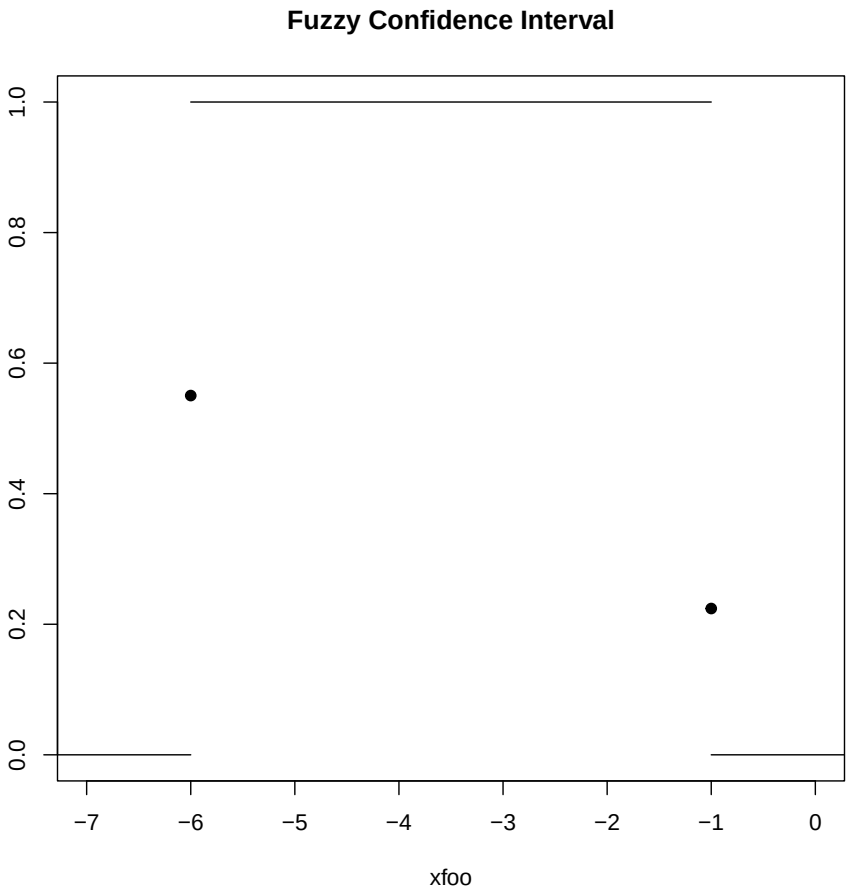

Figure 9: 95% rank sum confidence interval for made-up data.