Exact Distribution of a Two-Sample Nonparametric Test for Comparing Hazard Rates

William Breslin

March 2018

Abstract

A two-sample nonparametric test for comparing the hazard rate functions of two populations was proposed by Kochar in 1979. The exact null distribution of the test statistic does not have a known closed-form distribution function. In this paper, the exact null distribution of this test statistic is obtained using the statistical software R for small sample sizes. The normal approximation is discussed for large sample sizes ($m \geqslant 10$ and $n \geqslant 10$) as an alternative to finding the exact null distribution due to being computationally expensive to obtain. Additionally, the power of this test is compared via simulation to the rank-sum test and the log-rank test. The results from the simulation suggest that the hazard rate test is slightly more powerful than the rank-sum test, but less powerful than the log-rank test under the optimal condition for the log-rank test.

Introduction

Let X and Y be two non-negative independent random variables of continuous type with probability density functions f_X and f_Y, cumulative distribution functions F_X and F_Y, and survival functions $\bar{F}_X = 1 - F_X$ and $\bar{F}_Y = 1 - F_Y$ respectively. Let $r_X = f_X/\bar{F}_X$ and $r_Y = f_Y/\bar{F}_Y$ be their hazard rates.

Based on independent random samples $X_1, \ldots, X_m$ on X and $Y_1, \ldots, Y_n$ on Y, Kochar[1] (1979) considered the problem of testing the null hypothesis

$$H_0 : r_X(t) = r_Y(t), \quad \forall t \geqslant 0 \tag{1}$$

against the alternative hypothesis

$$H_A : r_X(t) \leqslant r_Y(t), \quad \forall t \geqslant 0 \tag{2}$$

with a strict inequality on a set of non-zero probability. The kernel of the proposed test statistic is

$$\phi(X_{i1}, X_{i2}, Y_{j1}, Y_{j2}) = \begin{cases} 1 & \text{if } yyxx \text{ or } xyyx \\ 0 & \text{if } xyxy \text{ or } yxyx \\ -1 & \text{if } xxyy \text{ or } yxxy \end{cases} \tag{3}$$

and is computed by choosing two values from the X-sample and two values from the Y-sample. These four values are then ranked, and depending on the sequencing of the ranks either a 1, 0, or -1 is assigned to ϕ. For example, the sequence $yyxx$ represents the case when both of the y values are smaller than both of the x values. In this case, ϕ is equal to 1. The test proposed in Kochar (1979) is to reject the null hypothesis H_0 against the alternative H_A for large values of the U-statistic,

$$W = \left\{ \binom{n}{2} \binom{m}{2} \right\}^{-1} \sum \phi(X_{i1}, X_{i2}, Y_{j1}, Y_{j2}) \tag{4}$$

The distribution of this test statistic does not have a known closed-form distribution; however, the procedure for computing this test statistic is relatively simple. The primary objective of this paper is to provide the means for computing the null distribution of this test statistic, along with the exact null distributions for small sample sizes.

Theory and Motivation

First we give a few definitions. A random variable X is said to be greater than the random variable Y in *stochastic ordering* if and only if the survival function of X is greater than the survival function of Y.

$$X \geqslant_{st} Y \iff \bar{F}_X(t) \geqslant \bar{F}_Y(t), \quad \forall t \geqslant 0 \tag{5}$$

Similarly, a random variable X is said to be greater than the random variable Y in *hazard rate ordering* if the hazard rate function of X is smaller than the hazard rate function of Y.

$$X \geqslant_{hr} Y \iff r_X(t) \leqslant r_Y(t) \quad \forall t \geqslant 0 \tag{6}$$

[1]This project was done under the direction of Prof. Subhash Kochar, whom I want to thank for his guidance and assistance.

For both of these definitions, the inequality must be strict on a set of non-zero probability. Note that the distribution of a random variable is uniquely determined by its hazard rate function, and vice-versa. That means that the null hypothesis in equation (1) is equivalent to:

$$H_0 : F_X(t) = F_Y(t), \quad \forall t \geqslant 0 \tag{7}$$

This is the null hypothesis for several nonparametric tests, such as the Wilcoxon rank-sum test and the log-rank (Savage) test. Even though this test uses hazard rate functions to compare the two samples, it is still testing whether or not the two samples came from populations with the same distribution. The main difference between the proposed hazard rate test and the rank-sum and log-rank tests is that the alternative hypotheses are not the same. The alternative hypothesis for the hazard rate test is that one of the random variables is greater than the other in hazard rate ordering, whereas the alternative for the rank-sum and log-rank tests is that one of the random variables is greater than the other in stochastic ordering.

$$H_A : X \geqslant_{hr} Y \tag{8}$$

An important fact that will be shown is that hazard rate ordering implies stochastic ordering, but the converse is not necessarily true. The following lemma will be useful for proving this.

Lemma: $X \geqslant_{hr} Y$ if and only if $\bar{F}_X(t)/\bar{F}_Y(t)$ is a non-decreasing function of t.

The proof below involves writing the hazard rate function in terms of the log of the survival function, and using the fact that the log function is monotonic.

Proof of Lemma:

$$\begin{aligned}
X \geqslant_{hr} Y &\iff r_X(t) \leqslant r_Y(t) \\
&\iff -\frac{d}{dt}\log\left[\bar{F}_X(t)\right] \leqslant -\frac{d}{dt}\log\left[\bar{F}_Y(t)\right] \\
&\iff \frac{d}{dt}\log\left[\bar{F}_X(t)/\bar{F}_Y(t)\right] \geqslant 0 \\
&\iff \frac{d}{dt}\bar{F}_X(t)/\bar{F}_Y(t) \geqslant 0 \\
&\iff \bar{F}_X(t)/\bar{F}_Y(t) \text{ is a non-decreasing in } t
\end{aligned}$$

This lemma is used to prove that hazard rate ordering implies stochastic ordering.

Theorem: If $X \geqslant_{hr} Y$, then $X \geqslant_{st} Y$.

This proof is fairly straightforward. By the above lemma, it can easily be seen that $\bar{F}_X(t)/\bar{F}_Y(t) \geqslant 1$, since $\bar{F}_X(0)/\bar{F}_Y(0) = 1$ and since $\bar{F}_X(t)/\bar{F}_Y(t)$ is non-decreasing. This fact will be used to prove the above theorem, and is the point of failure for the converse of this theorem.

Proof of Theorem:

$$\begin{aligned}
X \geqslant_{hr} Y &\iff \bar{F}_X(t)/\bar{F}_Y(t) \text{ is non-decreasing} \\
&\implies \bar{F}_X(t)/\bar{F}_Y(t) \geqslant 1 \text{ for every } t > 0 \\
&\iff \bar{F}_X(t) \geqslant \bar{F}_Y(t) \text{ for every } t > 0 \\
&\iff X \geqslant_{st} Y
\end{aligned}$$

This shows that hazard rate ordering is a stronger condition than stochastic ordering. This is beneficial because if sufficient evidence is found to reject the null hypothesis for the alternative of hazard rate ordering, then sufficient evidence has also been found to reject the null hypothesis for the alternative of stochastic ordering. Because of this, it should be more likely that we correctly reject the null hypothesis for this test, thus theoretically increasing the power of this test relative to the stochastic ordering alternative. This is the motivation for using hazard rate functions to test for equality of distribution functions.

Now for the theory behind the test statistic given in Kochar (1979). Note that the alternative hypothesis for this test, that $X \geqslant_{hr} Y$, is true if and only if for every $s \geqslant t \geqslant 0$

$$\delta(s,t) \stackrel{\text{def}}{=} \bar{F}_X(s)\bar{F}_Y(t) - \bar{F}_X(t)\bar{F}_Y(s) \geqslant 0 \tag{9}$$

Equation (9) will be the basis from which the test statistic will be constructed. Let $s = \max\{x, y\}$ and $t = \min\{x, y\}$. For this choice of s and t, it will always be true that $s \geqslant t \geqslant 0$, assuming that the variables are non-negative. Now define $\eta(F_X, F_Y)$, as shown in equation (10) below.

$$\eta(F_X, F_Y) \stackrel{\text{def}}{=} \mathbb{E}\Big[\delta(\max\{X,Y\}, \min\{X,Y\})\Big] \tag{10}$$

$$= \iint\limits_{0 \leqslant x \leqslant y} \delta(s,t)\Big[dF_X(x)dF_Y(y) + dF_X(y)dF_Y(x)\Big] \tag{11}$$

Substituting equation (9) into equation (11), expanding the integrand, and evaluating the integral results in the following probability expression.

$$\begin{aligned}
&P(Y_1 \leqslant Y_2 \leqslant X_1 \leqslant X_2) + P(X_1 \leqslant Y_1 \leqslant Y_2 \leqslant X_2) \\
&-P(X_1 \leqslant X_2 \leqslant Y_1 \leqslant Y_2) - P(Y_1 \leqslant Y_2 \leqslant X_1 \leqslant X_2)
\end{aligned} \tag{12}$$

The kernel ϕ as defined in equation (3) will lead to an unbiased estimator of $\eta(F_X, F_Y)$. The X and Y sequences in equation (3) are obtained from the probability expressions in equation (12) above. The first

two probability expressions suggest that $X \geqslant_{hr} Y$ whereas the second two suggest $X \leqslant_{hr} Y$. The test statistic for this test is simply counting the number of sequences that suggest a hazard rate ordering in the direction of the alternative hypothesis, and subtracting from that the number of sequences suggesting a hazard rate ordering in the direction opposite of the alternative hypothesis.

Exact Null Distribution of W

Consider two data-sets X and Y of size m and n respectively. This test is a permutation test, so we consider all possible permutations of the combined data-set, and then separate the permutations back into two samples of size m and n. For every distinct permutation obtained this way, the test statistic W will be computed, and the frequency distribution of W is the null distribution for this test.

$$\#\text{ of distinct permutations} = \binom{n+m}{n} \tag{13}$$

To compute the test statistic for a given permutation, all possible combinations of two X and two Y must be compared to the sequences in equation (4).

$$\#\text{ of sequences} = \binom{n}{2}\binom{m}{2} \tag{14}$$

Now to find the null distribution computationally. The most difficult part of computing the test statistic is comparing the sequences of X and Y values. Table 1 shows expressions that are equivalent to the sequences described in equation (3). These expressions are not as concise as the original conditions; however, they are computationally much easier to implement.

Table 1: Equivalent Sequence Definitions

Sequence	Equivalent Condition
$yyxx$	$\min\{x_1, x_2\} > \max\{y_1, y_2\}$
$xyyx$	$\max\{x_1, x_2\} > \max\{y_1, y_2\}$ & $\min\{x_1, x_2\} < \min\{y_1, y_2\}$
$xxyy$	$\min\{y_1, y_2\} > \max\{x_1, x_2\}$
$yxxy$	$\max\{y_1, y_2\} > \max\{x_1, x_2\}$ & $\min\{y_1, y_2\} < \min\{x_1, x_2\}$

The code for finding the null distribution is broken up into two pieces. The first piece is a function that computes the kernel $\phi(\cdot)$ for a given permutation. The second piece finds the exact null distribution for given sample sizes m and n.

Function: $\phi(\mathbf{x}, \mathbf{y})$

```
kernel = function(x.pairs,y.pairs){
  w=0
  for (i in 1:nrow(x.pairs)){
    for (j in 1:nrow(y.pairs)){
      x1 = x.pairs[i,1]; x2 = x.pairs[i,2]
      y1 = y.pairs[j,1]; y2 = y.pairs[j,2]
      if (min(x1,x2) > max(y1,y2)){
        w = w+1}
      if (max(x1,x2) > max(y1,y2) & min(x1,x2) <
          min(y1,y2)){
        w = w+1}
      if (min(y1,y2) > max(x1,x2)){
        w = w-1}
      if (max(y1,y2) > max(x1,x2) & min(y1,y2) <
          min(x1,x2)){
        w = w-1}}}
  return(w)}
```

Exact Null Distribution

```
# Note: this script may take some time to run
    to completion for large samples.

nulldist = function(m,n){
  N = m+n
  Ranks = 1:N
  x.data = t(combn(Ranks,m)); y.data=c()
  for (i in 1:nrow(x.data)){
    if (i == 1){
      y.data =
          t(as.data.frame(setdiff(Ranks,x.data[1,])))}
    else{
      temp = setdiff(Ranks, x.data[i,])
      y.data = rbind(y.data,temp)}
    x.pairs = t(combn(x.data[i,],2))
    y.pairs = t(combn(y.data[i,],2))
    if (i == 1){
      w =
          kernel(x.pairs,y.pairs)/(choose(n,2)*choose(m,2))}
    else{
      temp =
          kernel(x.pairs,y.pairs)/(choose(n,2)*choose(m,2))
      w = c(w,temp)}}
  support = sort(unique(w))
  for (i in support){
    if (i == min(support)){
      freq = sum(w==i)
      prob = freq/length(w)}
    else{
      freq = c(freq, sum(w==i))
      prob = c(prob, sum(w==i)/length(w))}}
  table = cbind(support,freq,prob)
  colnames(table)=c("W","Occurrences","Probability")
  return(table)}
```

The R output below is the result of applying the nulldist(·) function to selected small sample sizes. The tables get large very quickly, so complete distribution tables will only be provided in the appendix for sample sizes below $m = 6$ and $n = 6$; If exact distributions for larger sample sizes are needed, the code on the previous page can be used to find them. For larger sample sizes the computation time needed for the code to run to completion increases non-linearly. It is recommended that the nulldist(·) function is only used up to $n = 10$ and $m = 10$.

Exact Distribution (m=3, n=3)

```
# Input
> nulldist(3,3)

# Output
            W Occurrences Probability
[1,] -1.0000000         2        0.10
[2,] -0.5555556         3        0.15
[3,] -0.3333333         2        0.10
[4,] -0.1111111         3        0.15
[5,]  0.1111111         3        0.15
[6,]  0.3333333         2        0.10
[7,]  0.5555556         3        0.15
[8,]  1.0000000         2        0.10
```

Exact Distribution (m=3, n=4)

```
# Input
> nulldist(3,4)

# Output
             W Occurrences Probability
 [1,] -1.0000000         2  0.05714286
 [2,] -0.6666667         4  0.11428571
 [3,] -0.5555556         1  0.02857143
 [4,] -0.3333333         4  0.11428571
 [5,] -0.2222222         2  0.05714286
 [6,] -0.1111111         1  0.02857143
 [7,]  0.0000000         6  0.17142857
 [8,]  0.1111111         2  0.05714286
 [9,]  0.2222222         1  0.02857143
[10,]  0.3333333         5  0.14285714
[11,]  0.4444444         1  0.02857143
[12,]  0.5555556         1  0.02857143
[13,]  0.6666667         3  0.08571429
[14,]  1.0000000         2  0.05714286
```

The Normal Approximation

The null distribution is symmetric when $n = m$, and approximately symmetric otherwise. As a result, the expected value of W under H_0 is approximately 0. It was shown in Kochar (1979) that the null variance of W is given by the following expression, where $N = n + m$.

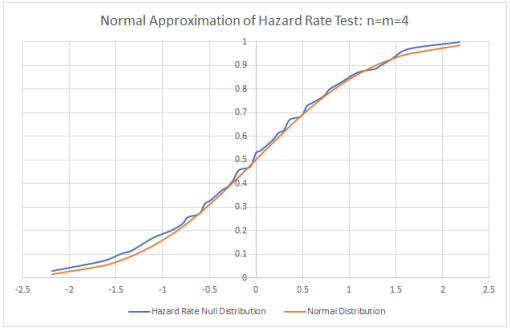
$$\frac{16mnN - (11m^2 + 11n^2 + 6mn) - 3N + 8}{210\binom{n}{2}\binom{m}{2}} \tag{15}$$

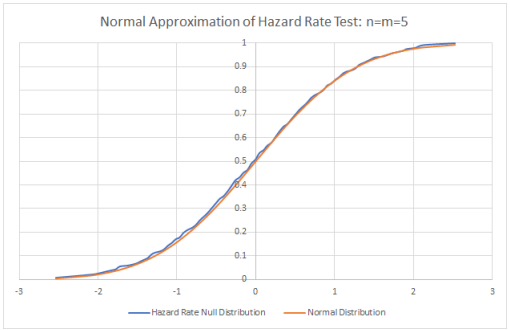
Also from Kochar (1979), it is known that as the sample sizes increase, the distribution of Z approaches a standard normal distribution, where Z is defined as:

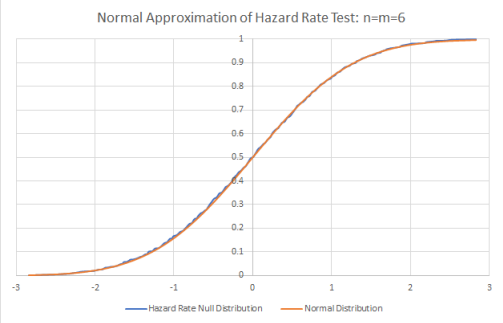
$$Z = \frac{W - \mathbb{E}_0(W)}{\mathrm{Sd}_0(W)} \tag{16}$$

When working with larger samples, it is convenient to compute the p-value using the normal approximation due to the computational expense required to find the exact null distribution of W. In this section, it will be shown that the normal approximation is sufficient even for relatively small values of m and n.

For now, consider the case where both samples are either of size 6 or of size 10. Below are normal quantile plots comparing how closely the exact null distribution follows a normal distribution.







From these plots, it appears that the samples of size 6 appear to deviate from normality out in the tails of the distribution. On the other hand, the samples of size 10 appear to follow a normal distribution very closely, even out in the tails of the distribution.

The question that needs to be answered is: at what sample size does the normal approximation sufficiently approximate the exact null distribution of W? From the normal quantile plots, it appears that samples of size 10 could be sufficiently approximated by a normal distribution. To get a more precise look at this approximation, Table 2 shows the exact critical values for a right-tailed test at various significance levels, as well as the critical values obtained from the normal approximation. These critical values were found for some commonly used significance levels; however, as the exact distribution of W is discrete, there is no guarantee that it is possible to obtain a specific significance level exactly. The actual significance level obtained from the exact distribution is also shown in the table below for reference.

Table 2: Critical Values for $n = m = 10$

α	Exact W	Actual α	Normal Approx. of W
0.10	.338	(0.0998)	.336
0.05	.433	(0.0498)	.431
0.025	.513	(0.0249)	.513
0.01	.603	(0.0099)	.609

As made evident by Table 2, the critical values obtained from the normal approximation are very close to those from the exact null distribution, so this shows that for samples of size $m = 10$ and $n = 10$ and above, the normal approximation is sufficient. For sample sizes between 6 and 10, the normal approximation is questionable, but is still fairly close starting at $m = 8$ and $n = 8$.

Power Comparisons

Part of the motivation behind the hazard rate test was that it would potentially be more powerful than other, similar tests. In this section, the power of this test is compared to the Wilcoxon rank-sum test and the (Savage) log-rank test. The power was compared through simulation using an exponential distribution, for which the log-rank test is the locally most powerful rank based test. The following steps describe the simulation method used to estimate the power of these tests.

1. Take a random sample of size m from an exponential distribution with mean $\mu = 1$.

2. Take a random sample of size n from an exponential distribution with mean $\mu > 1$, for select values of μ as shown in the tables below.

3. Compute the p-values for each test from these samples

4. Repeat this process 1000 times

5. Find the proportion of times that we reject the null hypothesis at $\alpha = 0.05$ for each test.

Table 3: Power Comparison: $n = m$

m, n	μ	HR	Wilcoxon	Log-Rank
	1.2	.059	.055	.086
$m = 10$	1.4	.102	.088	.143
$n = 10$	1.6	.152	.129	.196
	1.8	.209	.172	.258
	2	.262	.223	.327
	1.2	.067	.065	.077
$m = 15$	1.4	.114	.113	.153
$n = 15$	1.6	.196	.183	.243
	1.8	.291	.257	.328
	2	.373	.328	.431
	1.2	.073	.074	.090
$m = 20$	1.4	.134	.131	.177
$n = 20$	1.6	.231	.211	.278
	1.8	.335	.307	.401
	2	.446	.427	.534

Table 4: Power Comparison: $n \neq m$

m, n	μ	HR	Wilcoxon	Log-Rank
	1.2	.064	.060	.089
$m = 10$	1.4	.093	.102	.138
$n = 15$	1.6	.149	.156	.213
	1.8	.215	.209	.295
	2	.300	.274	.380
	1.2	.073	.073	.108
$m = 10$	1.4	.116	.128	.169
$n = 20$	1.6	.186	.178	.262
	1.8	.254	.244	.347
	2	.345	.324	.463
	1.2	.078	.067	.101
$m = 15$	1.4	.133	.116	.170
$n = 20$	1.6	.212	.193	.267
	1.8	.298	.275	.370
	2	.386	.354	.482

Tables 3 and 4 show the resulting power estimates from these simulations. It appears that the hazard rate test is more powerful than the rank-sum test in all cases, but is less powerful than the log-rank test under the optimal condition for applying the log-rank test.

Power Simulations

```
library(HazardRateTest)
library(survival)

set.seed(100)

Power.Simulation = function(alt, n, m){
  Hazard.Reject = Wilcox.Reject = Logrank.Reject
      = 0
  itr = 1000

  for (i in 1:itr){
    x = rexp(n=n, rate=1)
    y = rexp(n=m, rate=1/alt)
    H = hazard_test(x,y)
    Hazard.p = H$p.value
    if (Hazard.p <= 0.05){Hazard.Reject =
        Hazard.Reject+1}
    W = wilcox.test(x,y)
    Wilcox.p = W$p.value
    if (Wilcox.p <= 0.05){Wilcox.Reject =
        Wilcox.Reject+1}
    failure = c(x,y)
    N = length(x)+length(y)
    censor = rep(1,N)
    group = c(rep(1, length(x)),rep(2,
        length(y)))
    logrank = survdiff(Surv(failure, censor) ~
        group)
    chi.sq = logrank$chisq
    Logrank.p =
        1-pchisq(chi.sq,length(logrank$n)-1)
    if (Logrank.p <= 0.05){Logrank.Reject =
        Logrank.Reject+1}
  }

  Hazard.Power = Hazard.Reject/itr;
      print(Hazard.Power)
  Wilcox.Power = Wilcox.Reject/itr;
      print(Wilcox.Power)
  Logrank.Power = Logrank.Reject/itr;
      print(Logrank.Power)
}
```

Conclusion

The primary objective for this project was to find the null distribution of the test statistic based on the concept of hazard rate ordering, developed by Prof. Kochar in 1979. The code included within the paper is only sufficient for finding the exact distribution when the sample sizes are small due to the large amount of computing power needed for larger sample sizes. For sample sizes starting at $m = 10$ and $n = 10$, it is shown that the normal approximation gives satisfactory results that are very close to those obtained using the exact distribution. So finding the exact distribution for larger sample sizes is not necessary. The exact distributions for samples up to size $m = 10$ and $n = 10$ are included in an Excel file included with this submission.

References

[1] Kochar, Subhash. (1979). Distribution-Free Comparison of Two Probability Distributions with Reference to their Hazard Rates. *Biometrika*, **66**, 437-441. 10.1093/biomet/66.3.437.

[2] Kochar, Subhash. (1981). A New Distribution-Free Test for the Equality of Two Failure Rates. *Biometrika*, **68**, 423-426. 10.1093/biomet/68.2.423.