SOME NONPARAMETRIC TESTS

1. Introduction

In parametric statistics, we may consider particular parametric families, such as the normal distribution in testing for equality of variances via F tests, or for equality of means via t tests or analysis of variance. In regression, the assumption of i.i.d. $N(0, \sigma^2)$ errors is used in testing whether regression coefficients are significantly different from 0.

The Wilks test applies to more general families of distributions indexed by θ in a finite-dimensional parameter space Θ. Similarly, the χ^2 test of composite hypotheses applies to fairly general parametric subfamilies of multinomial distributions.

In nonparametric statistics, there actually still are parameters in a sense, such as the median m or other quantiles, but we don't have distributions determined uniquely by such a parameter. Instead there are more general restrictions on the distribution function F of the observations, such as that F is continuous. So the families of possible distributions are infinite-dimensional. Given n observations $X_1, ..., X_n$, we can always form their order statistics $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ and their empirical distribution function $F_n(x) := \frac{1}{n} \sum_{j=1}^{n} 1_{X_j \le x}$. Some tests can be based on these. This handout will consider two tests of whether two samples both came from the same (unknown) distribution, the Mann–Whitney–Wilcoxon rank-sum test and the Kolmogorov–Smirnov test. Also we will have the Wilcoxon "signed rank" test of whether paired variables (X_j, Y_j) have the same distribution as (Y_j, X_j).

2. Symmetry of random variables

Two real random variables X and Y are said to have the same distribution or to be *equal in distribution*, written $X =_d Y$, if for some F, $\Pr(X \le x) = \Pr(Y \le x) = F(x)$ for all x. If $X =_d Y$, then for any constant c, $X + c =_d Y + c$. But one cannot necessarily add the same random variable to both sides of an equality in distribution. Let X and Y be i.i.d. $N(0, 1)$. Then $X =_d Y$, but $X + X \neq_d X + Y$ because $X + X = 2X$ is $N(0, 4)$, but $X + Y$ is $N(0, 2)$.

A real random variable X is said to have a *symmetric* distribution (around 0) or to be *symmetric* if X and $-X$ have the same distribution.

Date: 18.650, Nov. 20, 2015.

Then, given a real number m, (the distribution of) X is said to be *symmetric around* m if $X - m$ is symmetric around 0, i.e. $X - m$ and $m - X$ have the same distribution. Equivalently, X and $2m - X$ have the same distribution.

Examples. Any $N(\mu, \sigma^2)$ is symmetric around μ. Any $t(d)$ distribution is symmetric around 0. A Beta(a, b) distribution is symmetric around m if and only if both $m = 1/2$ and $a = b$. If X has a density f with $f(x) > 0$ for all $x > 0$ and $f(x) = 0$ for all $x \le 0$, then X is not symmetric around any m. This applies, for example, to gamma distributions, including χ^2 and exponential distributions, and to F distributions.

Fact 1. *Suppose a random variable X is symmetric around some m.*
(a) Then m is a median of X.
(b) Actually m is **the** *median of X, in the sense that if X has a non-degenerate interval of medians, m must be the midpoint of that interval.*
(c) If $E|X| < +\infty$ then also $EX = m$.

Proof. (a) Let X be symmetric around m. Then from the definitions,

$$\Pr(X \ge m) = \Pr(X - m \ge 0) = \Pr(m - X \ge 0) = \Pr(X \le m).$$

Then $2\Pr(X \le m) = \Pr(X \le m) + \Pr(X \ge m) = 1 + \Pr(X = m) \ge 1$ and so $\Pr(X \le m) = \Pr(X \ge m) \ge 1/2$ and m is indeed a median of X.
(b) Suppose for some $c \ne 0$, $m + c$ is a median of X. Then c is a median of $X - m$, so it is also a median of $m - X$. It follows that $-c$ is a median of $X - m$, so $m - c$ is a median of X. So the interval of medians of X is symmetric around m and m is the midpoint of it.
(c) $E(X - m) = EX - m = E(m - X) = m - EX$, so $2EX = 2m$ and $EX = m$. □

Example. Consider the binomial$(5, 1/2)$ distribution. Its distribution function F satisfies $F(x) = 1/2$, $2 \le x < 3$. So it has an interval $[2, 3]$ of medians. The distribution is symmetric around 5/2=2.5, *the* median as the midpoint of the interval of medians, and also the mean.

3. The Mann–Whitney–Wilcoxon rank-sum test

This is a test of whether two samples come from the same distribution, against the alternative that members of one sample tend to be larger than those of the other sample (a location or shift alternative). No parametric form of the distributions is assumed. They can be quite general, as long as the distribution functions are continuous. One might want to use such a test, called a nonparametric test, if, for example,

the data have outliers and so appear not to be normally distributed. Rice considers this test in subsection 11.2.3 pp. 435–443.

The general assumption for the test is that real random variables $X_1, \ldots, X_m$ are i.i.d. with a distribution function F, and independent of $Y_1, \ldots, Y_n$ which are i.i.d. with another distribution function G, with both F and G continuous. The hypothesis to be tested is H_0: $F = G$. The test works as follows: let $N = m + n$ and combine the samples of X's and Y's into a total sample $Z_1, \ldots, Z_N$. Arrange the Z_k in order (take their order statistics) to get $Z_{(1)} < Z_{(2)} < \cdots < Z_{(N)}$. With probability 1, no two of the order statistics are equal because F and G are continuous. Let $\text{rank}(V) = k$ if $V = Z_{(k)}$ for $V = X_i$ or Y_j. Let $T_X = \sum_{i=1}^{m} \text{rank}(X_i)$. Then T_X will be the test statistic. H_0 will be rejected if either T_X is too small, indicating that the X's tend to be less than the Y's, or if T_X is too large, indicating that the Y's tend to be less than the X's. To determine quantitatively what values are too small or too large, we need to look into the distribution of T_X under H_0.

If H_0 holds then $Z_1, \ldots, Z_N$ are i.i.d. with distribution $F = G$. Let E_0 be expectation, and Var_0 the variance, when the hypothesis H_0 is true. Let R_i be the rank of X_i. Then R_i has the discrete uniform distribution on $\{1, 2, \ldots, N\}$, $\Pr(R_i = k) = 1/N$ for $k = 1, ..., N$. This distribution has mean $E_0 R_i = \frac{1}{N}\frac{N(N+1)}{2} = \frac{N+1}{2}$. A variable U with this distribution has

$$E(U^2) = \frac{1}{N}\sum_{k=1}^{N} k^2 = \frac{1}{N}\frac{N(N+1)(2N+1)}{6} = \frac{(N+1)(2N+1)}{6}. \tag{1}$$

It follows that the variance $\text{Var}_0(R_i)$ of the distribution is

$$\frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4} = \frac{4N^2 + 6N + 2 - 3N^2 - 6N - 3}{12} = \frac{N^2 - 1}{12}. \tag{2}$$

Recalling that the continuous $U[a, b]$ distribution has a variance $(b - a)^2/12$, the 12 in the denominator is to be expected. Moreover, let X have the discrete uniform distribution on $\{1, ..., N\}$, as each R_i does. Let V have a $U[-1/2, 1/2]$ distribution and be independent of X. Then $X + V$ is easily seen to have a $U[1/2, N + \frac{1}{2}]$ distribution, so $\text{Var}(X+V) = N^2/12$, while by independence, $\text{Var}(X+V) = \text{Var}(X) + \text{Var}(V) = \text{Var}(X) + 1/12$, so $\text{Var}(X) = (N^2 - 1)/12$, giving another proof of (2).

We know the mean and variance of each rank R_i for $i = 1, \ldots, m$. To find the mean and variance of the sum $T_X = \sum_{i=1}^{m} R_i$, the mean is easy, namely $E_0 T_X = m(N + 1)/2$. For the variance, we need to find

the covariances of ranks R_i and R_j for $i \neq j$, all of which equal the covariance of R_1 and R_2. These two ranks are not independent because they cannot have the same value. First we find $E_0(R_1R_2)$. To make this easier we can express it as $E_0(R_1E_0(R_2|R_1))$. (Using the conditional expectation breaks the calculation into two easier ones.) We have

$$E_0(R_2|R_1) = \frac{1}{N-1}\left[\frac{N(N+1)}{2} - R_1\right]$$

because, given R_1, R_2 can have any of the $N-1$ values in $\{1, 2, \ldots, N\}$ other than R_1, each with probability $1/(N-1)$. It follows by (1) that

$$E_0(R_1R_2) = E_0(R_1E_0(R_2|R_1)) = \frac{1}{N-1}\left[\frac{N(N+1)}{2}\cdot\frac{N+1}{2} - \frac{2N^2+3N+1}{6}\right]$$

$$= \frac{1}{N-1}\left[\frac{3N^3+6N^2+3N-(4N^2+6N+2)}{12}\right]$$

$$= \frac{1}{N-1}\left[\frac{3N^3+2N^2-3N-2}{12}\right] = \frac{(3N+2)(N^2-1)}{12(N-1)} = \frac{(3N+2)(N+1)}{12}.$$

Thus under H_0 the covariance of R_1 and R_2 is

$$E_0(R_1R_2) - (E_0R_1)^2 = \frac{3N^2+5N+2}{12} - \frac{N^2+2N+1}{4}$$

$$= \frac{3N^2+5N+2-3N^2-6N-3}{12} = -\frac{N+1}{12},$$

and so for $1 \leq i < k \leq m$

$$\mathrm{Cov}_0(R_i, R_k) = \mathrm{Cov}_0(R_1, R_2) = -\frac{N+1}{12}. \tag{3}$$

By the way from (2), the standard deviation of an individual rank is asymptotic to $N/\sqrt{12}$ as $N \to +\infty$ and so the correlation of two different ranks is asymptotic to $-1/N$. It makes sense that the covariance and correlation should be negative, because if one rank is large, another will tend to be smaller. It also makes sense that the correlation should approach 0 for N large, as the influence of one rank on another becomes smaller.

By the formula for the variance of a sum of (dependent) variables, (2), and (3),

$$\begin{aligned}
\mathrm{Var}_0(T_X) &= m\mathrm{Var}_0(R_1) + m(m-1)\mathrm{Cov}_0(R_1, R_2) \\
&= m\left(\frac{N^2-1}{12}\right) - m(m-1)\left(\frac{N+1}{12}\right) \\
&= \frac{mN^2 - m - m^2N + mN - m^2 + m}{12} \\
&= \frac{mN(N-m) + m(N-m)}{12} \\
&= \frac{(N+1)m(N-m)}{12} = \frac{(N+1)mn}{12},
\end{aligned}$$

which agrees with the formula given in Rice, Third Ed., p. 438 Theorem A.

Under H_0, T_X has a distribution symmetric around its mean $m(N+1)/2$, because $r \mapsto N+1-r$ is a one-to-one transformation of the set $\{1, 2, ..., N\}$ onto itself; if each rank R_i is replaced by $N+1-R_i$, then T_X is changed to $m(N+1) - T_X$. When m and n are both large, T_X becomes approximately normal with its given mean and variance. Rice (p. 441) says that the normal approximation works well when m and n are both larger than 10. He also gives tables for $\max(m, n) \leq 20$ (covering 3 pages). For $m < 10 < 20 < n$ one can use R.

Rice considers the statistic $T_Y = \sum_{j=1}^{n} \mathrm{rank}(Y_j)$. Its mean and variance under H_0 equal those of T_X with m and n interchanged. This does not change the variance. One can see that because $T_X + T_Y \equiv N(N+1)/2$, the sum of all N ranks, so $T_Y \equiv N(N+1)/2 - T_X$, so T_Y and T_X have the same variance.

It is arbitrary which variables are called X's and which are called Y's. For Rice's Table 8, he instead uses $n_1 = \min(m, n)$, the size of the "smaller sample," and $n_2 = \max(m, n)$. If $m = n$, an arbitrary choice is made of which sample is called smaller, say it's the X's. Let R be the sum of the ranks of the elements of the smaller sample. Let $R' = n_1(N+1) - R$ and $R^* := \min(R, R')$. Then H_0 is rejected in a two-sided test at level α if R^* is less or equal to the critical value given for n_1, n_2, and α for the two-sided test. For a one-sided test against the alternative that the elements of the smaller sample tend to be smaller than the elements of the other sample, use R instead of R^*, with α for a one-sided test. For a one-sided test against the alternative that the elements of the smaller sample tend to be larger than those of the other sample, use R' instead of R.

Under H_0, if the X_i are the smaller sample, T_X has a distribution symmetric around $n_1(N+1)/2$, and so T_X has the same distribution as R'.

3.1. **The Mann–Whitney form.** The statistics T_X or T_Y are as originally defined by F. Wilcoxon. "Mann–Whitney" refers to another test statistic defined as follows. (Rice also treats both forms, but his Table 8 is for the Wilcoxon rank-sum form.) Assume, as we have been, that the X_i and Y_j are independent and from continuous distributions, so that there are no ties. Let $M_X := \sum_{i=1}^{m}\sum_{j=1}^{n} 1_{Y_j<X_i}$, called the Mann–Whitney statistic. Symmetrically let $M_Y := \sum_{i=1}^{m}\sum_{j=1}^{n} 1_{X_i<Y_j}$. If there are no ties then $M_Y \equiv mn - M_X$. We can write the rank-sum (Wilcoxon) statistic $T_X = \sum_{i=1}^{m} \mathrm{rank}(X_i)$ equivalently as $\sum_{i=1}^{m}\mathrm{rank}(X_{(i)})$, recalling that ranks are in the combined sample of X's and Y's. For each i we can write $\mathrm{rank}(X_{(i)}) = i + \sum_{j=1}^{n} 1_{Y_j<X_i}$, because i is the rank of $X_{(i)}$ among the X's only, which is increased by 1 in the combined sample for each $Y_j < X_{(i)}$. So

$$T_X \equiv M_X + \frac{m(m+1)}{2}.$$

Since $P_0(Y_j < X_i) = 1/2$ for each i and j, we have $E_0 M_X = \frac{mn}{2}$. This also fits with $E_0 T_X = m(N+1)/2$. The smallest possible value of M_X is 0 and the largest is mn. Under H_0, the distribution of M_X is symmetric around its mean $mn/2$, by the symmetry of T_X around $m(N+1)/2$. Since T_X and M_X differ only by a constant, they have the same variance under H_0 but different means. Tests of H_0 based on M_X or T_X are equivalent; Rice uses T_X and R uses M_X.

3.2. **The Mann–Whitney–Wilcoxon test in R.** In R, let x = c($X_1, ..., X_m$) be one sample and y = c($Y_1, ..., Y_n$) the other. Then wilcox.test(x,y) performs a two-sided rank-sum test of H_0, giving a p-value. The statistic "W" R computes is the Mann–Whitney statistic M_X. The p-value is the probability under H_0 that $\min(W, mn - W)$ is less than or equal to its observed value. For given x and y, wilcox.test(x,y) and wilcox.test(y,x) will give the statistics M_X and M_Y respectively, but the same p-value. If there are no ties, R computes M_X exactly but uses a normal approximation to the p-value if $\max(m, n) \geq 50$. The option wilcox.test(x,y,exact=TRUE) will try to compute exact p-values. Systems may run out of memory in trying this, for m and n of the order of several hundred or of "a few thousand" depending on the system. It seems one should use "exact = TRUE" if, for example, m is a single-digit number and $n \geq 50$.

3.3. **The exact distribution of T_X.** Under H_0, with F continuous, the ranks of the X_i in the combined sample are a random subset of m members of $\{1, 2, ..., M = m+n\}$. In other words, each such subset has probability $1/\binom{N}{m}$ of being chosen. Thus for each possible value k of T_X, $\Pr_0(T_X = k)$ is $\binom{N}{m}^{-1}$ times the number of subsets $J \subset \{1, ..., N\}$ with m elements such that $\sum_{j\in J} j = k$. This can be used when m is small, e.g. $m = 1$ or 2, so that a normal approximation is not accurate, yet n is too large for Rice's table to apply.

3.4. **Insensitivity to outliers.** The Wilcoxon–Mann–Whitney test, like other nonparametric methods, is not sensitive to outliers. If one observation is the largest, it will have rank N, and if it is made larger by an arbitrary amount, the rank and so the test statistic will not change. Whereas, the two-sample t-test is sensitive to outliers, beside depending on a normality assumption in which the two normal distributions have the same variance although they may have different means.

4. The Kolmogorov–Smirnov test

Here, as in the rank-sum test, we are given $X_1, ..., X_m$ assumed i.i.d. (F) and independent of $Y_1, ..., Y_n$ i.i.d. (G), and again want to test the hypothesis H_0 that $F = G$. The test statistic will have a well-defined distribution not depending on F under H_0 as long as F is continuous, so we will assume that. Unlike the Mann–Whitney–Wilcoxon test which aims to detect location (shift) alternatives, the Kolmogorov–Smirnov test works against arbitrary alternatives $F \neq G$. The Kolmogorov–Smirnov test will tend to be less powerful against location alternatives.

The test works as follows. Let $F_m(x) := \frac{1}{m}\sum_{i=1}^{m} 1_{X_i\le x}$ and $G_n(x) := \frac{1}{n}\sum_{j=1}^{n} 1_{Y_j\le x}$ be the empirical distribution functions of the two samples. The basic test statistic is $D_{m,n} := \sup_x |(F_m - G_n)(x)|$. The hypothesis H_0 is rejected for large enough values of $D_{m,n}$, how large depending on m and n. By the Glivenko–Cantelli theorem (proved in graduate probability, e.g. 18.175; Dudley, 2002, Theorem 11.4.2 p. 400), with probability 1, $\sup_x |(F_m - F)(x)| \to 0$ as $m \to \infty$ and $\sup_x |(G_n - G)(x)| \to 0$ as $n \to \infty$. If the hypothesis $H_0 : F = G$ is true, then $D_{m,n}$ will approach 0 with probability 1 as m and n both go to infinity. Whereas, if $F(x) \neq G(x)$ for at least one value of x, then $D_{m,n}$ will not approach 0. Thus, the test can detect any departure from H_0.

Tabulation of critical values of $D_{m,n}$, say for a few values of α such as 0.05, 0.01, and 0.001, is space-consuming because of the two variables m, n. Interpolation and extrapolation from such tables is not feasible in general, as adjoining values of m for a given n may give quite different

behavior of $D_{m,n}$. For example if $m = n = 20$, then $D_{m,n}$ has 21 possible values $0.05j$ for $j = 0, 1, ..., 20$ (where $j = 0$ could only occur in case of [many!] ties, which should not happen for continuous F) whereas for $m = 19$ and $n = 20$, $D_{m,n}$ has many more possible values, although not all numbers $j/380$ for $j = 1, ..., 380$ are actually possible.

To get a limiting distribution as m and n both go to infinity, the normalized test statistic is defined as

$$KS_{m,n} := \sqrt{\frac{mn}{m+n}} D_{m,n}. \tag{4}$$

The limiting distribution is then given by:

Theorem 1 (Smirnov, 1939)**.** *If H_0 holds and $F = G$ is continuous, then for any $M > 0$,*

$$\begin{aligned} \lim_{m,n\to\infty} \Pr(KS_{m,n} \geq M) &= 2\sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2M^2) \\ &= 2\left(e^{-2M^2} - e^{-8M^2} + e^{-18M^2} - \cdots\right) < 2e^{-2M^2}. \end{aligned} \tag{5}$$

The sum in Theorem 1 converges quite fast unless M is small, so a few terms or even the first term can give a good approximation. The convergence to the limit as $m, n \to \infty$, however, is slow and irregular, because of the effects of whether the least common multiple of m and n is large or not in relation to m and n.

4.1. **The Kolmogorov–Smirnov test in R.** If in R, x = c($X_1, ..., X_m$) and y = c($Y_1, ..., Y_n$), then ks.test(x,y) performs the test and gives a p-value, with a warning in case there are any ties among the values of X_i and Y_j. According to R documentation given on two websites I found, R finds exact p-values if $mn \leq 10^4$ and there are no ties. For $mn > 10^4$ it uses the asymptotic distribution given on the right side of (5) for $M = KS_{mn}$. By calling ks.test(x,y,exact = TRUE) one can request exact p-values for any m and n. In one example, I set x = runif(101) and y = runif(102,0.14,1.14). Then since $mn > 10^4$, the default ks.test gives the asymptotic approximation to the p-value which was 0.05152, whereas ks.test(x,y,exact=TRUE) gave 0.04605. So according to the exact p-value, the hypothesis $F = G$ should be rejected at level $\alpha = 0.05$, which the asymptotic approximation does not give. For other x and y generated the same way, ks.test(x,y) gave a p-value 0.1193, so $F = G$ would not be rejected at level $\alpha = 0.1$ using the asymptotic approximation, whereas ks.test(x,y,exact=TRUE) gave a p-value $0.0981 < 0.1$, so $F = G$ should be rejected at $\alpha = 0.1$. For the same x and y, wilcox.test(x,y) gave a p-value 0.02065, so $F = G$

would be rejected for the more usual $\alpha = 0.05$. This illustrates that for location (shift) alternatives, which this is ($U[0,1]$ vs. a shift of itself to the right by 0.14), the Mann–Whitney–Wilcoxon test is more powerful than the Kolmogorov–Smirnov test.

Setting the boundary at $mn = 10^4$ may be excessively cautious. On the system I use, an exact p-value for $m = 2000$ and $n = 2001$ was done without any noticeable waiting. But for m and n much larger, computation and memory use might be excessive.

4.2. **The one-sample Kolmogorov test and inequalities.** Before N. V. Smirnov (1939) proposed the two-sample test, Kolmogorov (1933) gave a test for the simple hypothesis H_F that observed $X_1, ..., X_n$ are i.i.d. with a distribution function F, with the test statistic $D_n := \sup_x |(F_n - F)(x)|$, where F_n is the empirical distribution function based on $X_1, ..., X_n$. Kolmogorov found that under H_F, if F is continuous, for $K_n := \sqrt{n} \sup_x |(F_n - F)(x)|$, and any $M > 0$, $\Pr(K_n > M)$ converges as $n \to \infty$ to the right side of (5). Dvoretzky, Kiefer and Wolfowitz (1956) had shown that for some constant $C < +\infty$,

$$\Pr(K_n > M) \le C \exp(-2M^2) \tag{6}$$

for all n and for all $M > 0$. Moreover, Massart (1990) proved that one can take $C = 2$, which is the best possible constant. An expanded proof is given in Dudley (2014, §1.5). Note that $2\exp(-2M^2)$ is the leading term on the right side of (5). A question then is, under $H_0 : \ F = G$, for what C if any can one replace K_n by $KS_{m,n}$. Simple examples with $1 \le m \le n \le 3$ show that

$$P_{m,n,M} := \Pr(KS_{m,n} > M) \le C \exp(-2M^2) \tag{7}$$

does not hold for $C = 2$. If the bound (7) holds, say with $C = 2$, and if for $M = KS_{m,n}$, $2\exp(-2M^2) \le \alpha$, then in a conservative test, we can reject $H_0 : F = G$ at level α.

4.3. **The relation of the one-sample and two-sample normalizations.** We have

$$\sqrt{\frac{mn}{m+n}}(F_m - G_n) = \sqrt{\frac{mn}{m+n}}((F_m - F) - (G_n - F))$$

$$= \sqrt{\frac{n}{m+n}}\sqrt{m}(F_m - F) - \sqrt{\frac{m}{m+n}}\sqrt{n}(G_n - F).$$

Under $H_0 : \ F = G$, the two one-sample processes $\sqrt{m}(F_m - F)$ and $\sqrt{n}(G_n - F)$ are independent of each other, have expectation 0 at all x, and the suprema of their absolute values each have limiting

distributions as on the right side of (5). For any fixed x, the limit distribution of each is $N(0, F(x)(1-F(x))$. Since

$$\left(\sqrt{\frac{n}{m+n}}\right)^2 + \left(\sqrt{\frac{m}{m+n}}\right)^2 \equiv 1,$$

for a fixed x the limit distribution of $\sqrt{\frac{mn}{m+n}}(F_m - G_n)(x)$ is also $N(0, F(x)(1-F(x)))$. One can also show that covariances of $H(x_1)$ and $H(x_2)$ for $-\infty < x_1 < x_2 < +\infty$ are the same for $H = \sqrt{\frac{mn}{m+n}}(F_m - G_n)$ as for $H = \sqrt{m}(F_m - F)$ and for $H = \sqrt{n}(G_n - G)$, not depending on m or n. This suggests, without proving, why (5) should hold provided that it holds in the one-sample case for $K_n = \sqrt{n}D_n$.

5. The Wilcoxon signed-rank test

Suppose we've observed pairs (X_j, Y_j) which are independent between pairs, not necessarily identically distributed. The hypothesis H_0 to be tested is that for each j, the distribution of (X_j, Y_j) is the same as that of (Y_j, X_j). Take the differences $D_j := X_j - Y_j$. Then H_0 implies that for each j, D_j has a distribution symmetric around 0. Find $|D_j|$ for $j = 1, ..., n$ and their order statistics, which will be called

$$0 \le |D|_{(1)} \le |D|_{(2)} \le \cdots \le |D|_{(n)}.$$

Suppose that each D_j has a continuous distribution, so that there are no ties,

$$0 < |D|_{(1)} < |D|_{(2)} < \cdots < |D|_{(n)}.$$

Define ranks by $R_j := k$ if $|D_j| = |D|_{(k)}$. Let $W_+ := \sum_{j=1}^n R_j 1_{D_j>0}$, the sum of the ranks of $|D_j|$ for those j with $D_j > 0$. Then W_+ is called the Wilcoxon signed-rank statistic. Under H_0, $1_{D_j>0} = 1$ or 0 with probability 1/2 each, independently of each other and the R_j, in other words these are Bernoulli (1/2) variables. Since the ranks are the integers $1, 2, ..., n$ in some order, by the independence, W_+ is equal in distribution to $\sum_{j=1}^n jB_j$ where B_j are i.i.d. Bernoulli (1/2) variables. It follows that under H_0,

$$E_0 W_+ = \sum_{j=1}^n j/2 = n(n+1)/4,$$

and its variance is

$$\mathrm{Var}_0(W_+) = \sum_{j=1}^n \mathrm{Var}(jB_j) = \sum_{j=1}^n \frac{j^2}{4} = \frac{n(n+1)(2n+1)}{24}.$$

It's easily seen that under H_0 the distribution of W_+ is symmetric around its mean $m = n(n+1)/4$, because $1 - B_j$ are also i.i.d. Bernoulli

(1/2) and independent of the R_j, and $\sum_{j=1}^n j(1-B_j) = n(n+1)/2 - \sum_{j=1}^n jB_j = 2m - \sum_{j=1}^n jB_j$.

For n large enough, the distribution of W_+ under H_0 is approximately normal with the given mean and variance. Rice, p. 449, says $n \geq 20$ is sufficient for the normal approximation.

The test statistic W_+ is used in a two-sided way: H_0 is rejected if W_+ is either too large or too small, relative to its distribution under H_0.

5.1. **The signed-rank test in R.** R does the signed rank test, setting x = c($X_1, ..., X_n$) and y = c($Y_1, ..., Y_n$) (which must of course be two vectors of the same length) via wilcox.test(x,y,paired=TRUE). (The default wilcox.test(x,y) with paired = FALSE does the Mann–Whitney–Wilcoxon test.) The normal approximation is used for $n \geq 50$.

5.2. **The exact distribution of W_+ under H_0.** For each set $J \subset \{1, ..., n\}$, the probability under H_0 that $B_j = 1$ if and only if $j \in J$ is $1/2^n$. For any $k = 0, 1, ..., n(n+1)/2$, the probability that $W_+ = k$ is $1/2^n$ times the number of subsets J with $\sum_{j \in J} j = k$. For example if $n \geq 4$, so $n(n+1)/2 \geq 10$, there are two sets J over which the sum is $W_+ = 4$, namely $\{4\}$ and $\{1, 3\}$, so $P_0(W_+ = 4) = 2/2^n$. There is one set, the empty set, over which the sum is 0, so $P_0(W_+ = 0) = 1/2^n$.

For $n \leq 25$, Rice, Table 9 p. A24 gives a table (the last of the tables) for the distribution of W_+. He gives $W_\alpha = k$ such that $P_0(W_+ \leq k)$ is "closest to" $\alpha/2$ (which is not exactly what it is in all cases). One would then reject H_0, according to Rice, if $W := \min\left(W_+, \frac{n(n+1)}{2} - W_+\right) \leq W_\alpha$. The actual size of the test might then be larger than α. For an example (mentioned in the header of the table), if $n = 8$, the table gives $W_{0.05} = 4$, where $P_0(W_+ \leq 4) = 7/256 \doteq 0.0273$ which is larger than but closer to 0.025 than is $P_0(W_+ \leq 3) = 5/256 \doteq 0.01953$. On the other hand, if $n = 6$, the table gives $W_{0.05} = 0$, where $P_0(W_+ \leq 1) = 1/32 \doteq 0.03125$ which is larger than but closer to 0.025 than is $P_0(W_+ = 0) = 1/64 \doteq 0.015625$. Evidently Rice felt that in the latter case the p-value for $W_+ \leq 1$ is too much larger than $\alpha/2$.

Actual sizes larger than a nominal value α can also occur when p-values are computed only by some approximation. When, as here, they are computed exactly, it seems to me unusual to declare an outcome significant at level α when one knows it is not.

5.3. **Applications of the signed rank test.** In one kind of study, there are pairs of individuals who have been selected to be alike in some

respects such as age, gender, and health status with respect to a condition for which there is an experimental treatment. One individual from each pair, chosen at random, would get the treatment, and a measurement for the treated individual would be Y_j for the jth pair. The other individual in the pair would not get the treatment (if it's a medication they might get a placebo) and X_j would be a measurement on the untreated member of the pair. In another similar kind of study, there would be just n individuals, and X_j and Y_j would be measurements on the jth individual before and after getting the treatment. In either case H_0 would be equivalent to the hypothesis that the treatment made no difference. In such studies one may be interested in one-sided alternatives: W_+ significantly larger than $n(n+1)/4$ would indicate that the X_j tend to be larger than the Y_j, and W_+ significantly less than $n(n+1)/4$ would indicate that the X_j tend to be smaller than the Y_j, which could indicate either unsafety or effectiveness of the treatment depending on which the measurement shows.

6. Normal approximations corrected for continuity

Let X be an integer-valued random variable whose distribution is approximately normal, such as a value of a rank-sum or signed rank statistic under H_0 for large enough m, n or n respectively, or a Poisson distribution with large λ. Let X have mean μ and standard deviation σ, so that $(X-\mu)/\sigma$ has approximately a $N(0,1)$ distribution. Let k be an integer such that $\Pr(X = k) > 0$. The simple normal approximation to $\Pr(X \le k)$ would be $\Phi((k-\mu)/\sigma)$. But since X is integer-valued, $\Pr(X \le k) = \Pr(X \le u)$ for $k \le u < k+1$. What the correction for continuity does is to take u as the midpoint of the given interval, i.e., $u = k + \frac{1}{2}$, and so to approximate $\Pr(X \le k)$ by

$$\Pr(X \le k) \sim \Phi\left(\frac{k + \frac{1}{2} - \mu}{\sigma}\right).$$

Corrections for continuity are widely adopted, as it's believed that they usually improve the approximation, although they don't in all cases. In particular, R uses corrections for continuity for normal approximations to p-values for wilcox.test, in both the unpaired, rank-sum, and paired, signed-rank, cases, for m and/or n above some values. When R does so, it tells you so in the output; "with continuity correction" also alerts you that a normal approximation is being used.

7. Ties

Although we've been assuming that all observations are distinct, ties often occur in real data, e.g. in large data sets because of rounding. Then "tied ranks" can be assigned as follows. Let the order statistics be

$$X_{(1)} \le \cdots \le X_{(i-1)} < X_{(i)} = X_{(i+1)} = \cdots = X_{(j)} < X_{(j+1)} \le \cdots \le X_{(n)}.$$

Then the observations X_k which gave $X_{(i)}, X_{(i+1)}, ..., X_{(j)}$ are all assigned the tied rank which is the average of $i, i+1, ..., j$, namely $\frac{i+j}{2}$. When there are ties, the distributions of all the test statistics change. R will give warning messages saying that exact p-values cannot be computed in case of ties. If m and n are large and there are only a few ties, the distributions should not be too seriously affected.

In the signed rank test, a tie $X_j = Y_j$ within a pair is problematic because an arbitrarily small change in X_j or Y_j affects whether $D_j = X_j - Y_j > 0$ or not. Similarly, in the Mann–Whitney–Wilcoxon rank-sum test, a tie $X_i = Y_j$ is problematic because arbitrarily small changes in either can switch the Mann–Whitney statistic term $1_{Y_j<X_i}$ between 0 and 1. A tie $X_i = X_k$ or $Y_i = Y_k$ doesn't affect such terms, but it will cause R to use a normal approximation even for small m and n, where it may be inaccurate. One can break such ties, adding small numbers, as long as the terms $1_{Y_j<X_i}$ are unchanged.

8. Two-sample Dvoretzky–Kiefer–Wolfowitz inequalities

This is an appendix, not part of the course material. Wei and Dudley (2011, 2012) recently showed that for $m = n$, (7) with $C = 2$ fails for $n \le 457$ but it holds for all $n \ge 458$. For $n \ge 4$, the smallest n for which H_0 can be rejected, (7) holds for $C = 2.16863$. They also showed for $m \ne n$ that (7) holds with $C = 2$ for $n \ge 4$ and $1 \le m < n \le 200$. Here $2\exp(-2M^2)$ is not necessarily a good approximation to $P_{m,n,M}$: Wei and Dudley show that for $100 < m < n \le 200$, the ratio $2\exp(-2M^2)/P_{m,n,M}$ is always at least 1.05, and for certain values of m it is at least 1.09. So the error in the approximation is *at least* 5 to 10%, as it also was in the example with $m = 101$ and $n = 102$ mentioned under "The Kolmogorov–Smirnov test in R."

9. Historical Notes

Wilcoxon (1945) first defined the rank-sum test, but only for $m = n$ and without developing it very much. Mann and Whitney (1947) defined a test using their form of statistic. They noted that it was equivalent to Wilcoxon's rank-sum statistic up to adding a constant.

They evaluated the mean and variance of their statistic under the null hypothesis H_0 and actually proved asymptotic normality for m and n both large, by way of even-order moments $E_0((M_X - E_0 M_X)^{2k})$ around the mean (by the symmetry, the odd-order moments are 0).

Kolmogorov in 1933 proposed the one-sample test of the hypothesis H_F that $X_1, ..., X_n$ are i.i.d. (F). Kolmogorov, a leading Russian probabilist, had earlier published works on probability in German. In 1933 there were few statistics journals in the world, and the editor of the Italian actuarial journal, Cantelli, was receptive to papers from Russia.

Wilcoxon (1945) first defined the signed-rank test, in the second half of the short paper where he defined the two-sample rank-sum test.

REFERENCES

Dudley, R. M. (2002), *Real Analysis and Probability*, 2d ed., Cambridge University Press.

Dudley, R. M. (2014), *Uniform Central Limit Theorems*, 2d ed., Cambridge University Press.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956), Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator, *Ann. Math. Statist.* **27**, 642–669.

Kolmogorov, A. N. (1933), Sulla determinazione empirica di un legge di distribuzione, *Giorn. Ist. Ital. Attuari* **4**, 83–91.

Mann, H. B., and Whitney, D. R. (1947), On a test of whether one of two random variables is stochastically larger than another, *Ann. Math. Statist.* **18**, 50–60.

Massart, P. (1990), The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality, *Ann. Probab.* **18**, 1269–1283.

Smirnov, N. V. (1939), An estimate of divergence between empirical curves of a distribution in two independent samples, *Bull. Mosk. Gos. Univ.* **2**, 3–14 (in Russian).

Wei, Fan, and Dudley, R. M. (2011), Dvoretzky–Kiefer–Wolfowitz inequalities for the two-sample case, arXiv:1107.5356v2 [Math.St] 11 Aug. 2011.

Wei, Fan, and Dudley, R. M. (2012), Two-sample Dvoretzky–Kiefer–Wolfowitz inequalities, *Statist. Probab. Letters* **82**, 636–644.

Wilcoxon, F. (1945), Individual comparisons by ranking methods, *Biometrics Bulletin* **1** no. 6, pp. 80–83.