13

Rank, Sign, and Permutation Statistics

Statistics that depend on the observations only through their ranks can be used to test hypotheses on departures from the null hypothesis that the observations are identically distributed. Such rank statistics are attractive, because they are distribution-free under the null hypothesis and need not be less efficient, asymptotically. In the case of a sample from a symmetric distribution, statistics based on the ranks of the absolute values and the signs of the observations have a similar property. Rank statistics are a special example of permutation statistics.

13.1 Rank Statistics

The *order statistics* $X_{N(1)} \leq X_{N(2)} \leq \cdots \leq X_{N(N)}$ of a set of real-valued observations $X_1, \ldots, X_N$ ith order statistic are the values of the observations positioned in increasing order. The *rank* R_{Ni} of X_i among $X_1, \ldots, X_N$ is its position number in the order statistics. More precisely, if $X_1, \ldots, X_N$ are all different, then R_{Ni} is defined by the equation

$$X_i = X_{N(R_{Ni})}.$$

If X_i is tied with some other observations, this definition is invalid. Then the rank R_{Ni} is defined as the average of all indices j such that $X_i = X_{N(j)}$ (sometimes called the *midrank*), or alternatively as $\sum_{j=1}^{N} 1\{X_j \leq X_i\}$ (which is something like an *uprank*).

In this section it is assumed that the random variables $X_1, \ldots, X_N$ have continuous distribution functions, so that ties in the observations occur with probability zero. We shall neglect the latter null set. The ranks and order statistics are written with double subscripts, because N varies and we shall consider order statistics of samples of different sizes. The vectors of order statistics and ranks are abbreviated to $X_{N(\,)}$ and R_N, respectively.

A *rank statistic* is any function of the ranks. A linear rank statistic is a rank statistic of the special form $\sum_{i=1}^{N} a_N(i, R_{Ni})$ for a given $(N \times N)$ matrix $\big(a_N(i, j)\big)$. In this chapter we are be concerned with the subclass of *simple linear rank statistics*, which take the form

$$\sum_{i=1}^{N} c_{Ni}\, a_{N,R_{Ni}}.$$

Here $(c_{N1}, \ldots, c_{NN})$ and $(a_{N1}, \ldots, a_{NN})$ are given vectors in $\mathbb{R}^N$ and are called the *coefficients* and *scores*, respectively. The class of simple linear rank statistics is sufficiently large

to contain interesting statistics for testing a variety of hypotheses. In particular, we shall see that it contains all "locally most powerful" rank statistics, which in another chapter are shown to be asymptotically efficient within the class of all tests.

Some elementary properties of ranks and order statistics are gathered in the following lemma.

13.1 ***Lemma.*** *Let $X_1, \ldots, X_N$ be a random sample from a continuous distribution function F with density f. Then*

(i) the vectors $X_{N(\cdot)}$ and R_N are independent;

(ii) the vector $X_{N(\cdot)}$ has density $N! \prod_{i=1}^{N} f(x_i)$ on the set $x_1 < \cdots < x_N$;

(iii) the variable $X_{N(i)}$ has density $N\binom{N-1}{i-1} F(x)^{i-1}\big(1 - F(x)\big)^{N-i} f(x)$; for F the uniform distribution on $[0, 1]$, it has mean $i/(N+1)$ and variance $i(N-i+1)/\big((N+1)^2(N+2)\big)$;

(iv) the vector R_N is uniformly distributed on the set of all $N!$ permutations of $1, 2, \ldots, N$;

(v) for any statistic T and permutation $r = (r_1, \ldots, r_N)$ of $1, 2, \ldots, N$,

$$\mathrm{E}\big(T(X_1, \ldots, X_N) \mid R_N = r\big) = \mathrm{E}T\big(X_{N(r_1)}, \ldots, X_{N(r_N)}\big);$$

(vi) for any simple linear rank statistic $T = \sum_{i=1}^{N} c_{Ni} a_{N,R_{Ni}}$,

$$\mathrm{E}T = N\bar{c}_N\bar{a}_N; \qquad \operatorname{var} T = \frac{1}{N-1}\sum_{i=1}^{N}(c_{Ni} - \bar{c}_N)^2 \sum_{i=1}^{N}(a_{Ni} - \bar{a}_N)^2.$$

Proof. Statements (i) through (iv) are well-known and elementary. For the proof of (v), it is helpful to write $T(X_1, \ldots, X_N)$ as a function of the ranks and the order statistics. Next, we apply (i). For the proof of statement (vi), we use that the distributions of the variables R_{Ni} and the vectors (R_{Ni}, R_{N_j}) for $i \neq j$ are uniform on the sets $I = \{1, \ldots, N\}$ and $\big\{(i, j) \in I^2 : i \neq j\big\}$, respectively. Furthermore, a double sum of the form $\sum_{i \neq j}(b_i - \bar{b})(b_j - \bar{b})$ is equal to $-\sum_i (b_i - \bar{b})^2$. ■

It follows that rank statistics are *distribution-free* over the set of all models in which the observations are independent and identically distributed. On the one hand, this makes them statistically useless in situations in which the observations are, indeed, a random sample from some distribution. On the other hand, it makes them of great interest to detect certain differences in distribution between the observations, such as in the two-sample problem. If the null hypothesis is taken to assert that the observations are identically distributed, then the critical values for a rank test can be chosen in such a way that the probability of an error of the first kind is equal to a given level α, for any probability distribution in the null hypothesis. Somewhat surprisingly, this gain is not necessarily counteracted by a loss in asymptotic efficiency, as we see in Chapter 14.

13.2 ***Example (Two-sample location problem).*** Suppose that the total set of observations consists of two independent random samples, inconsistently with the preceding notation written as $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$. Set $N = m + n$ and let R_N be the rank vector of the *pooled sample* $X_1, \ldots, X_m, Y_1, \ldots, Y_n$.

We are interested in testing the null hypothesis that the two samples are identically distributed (according to a continuous distribution) against the alternative that the distribution of the second sample is stochastically larger than the distribution of the first sample. Even without a more precise description of the alternative hypothesis, we can discuss a collection of useful rank statistics. If the Y_j are a sample from a stochastically larger distribution, then the ranks of the Y_j in the pooled sample should be relatively large. Thus, any measure of the size of the ranks $R_{N,m+1}, \ldots, R_{NN}$ can be used as a test statistic. It will be distribution-free under the null hypothesis.

The most popular choice in this problem is the *Wilcoxon statistic*

$$W = \sum_{i=m+1}^{N} R_{Ni}.$$

This is a simple linear rank statistic with coefficients $c = (0, \ldots, 0, 1, \ldots, 1)$, and scores $a = (1, \ldots, N)$. The null hypothesis is rejected for large values of the Wilcoxon statistic. (The Wilcoxon statistic is equivalent to the *Mann-Whitney statistic* $U = \sum_{i,j} 1\{X_i \leq Y_j\}$ in that $W = U + \frac{1}{2}n(n+1)$.)

There are many other reasonable choices of rank statistics, some of which are of special interest and have names. For instance, the *van der Waerden statistic* is defined as

$$\sum_{i=m+1}^{N} \Phi^{-1}(R_{Ni}).$$

Here Φ^{-1} is the standard normal quantile function. We shall see ahead that this statistic is particularly attractive if it is believed that the underlying distribution of the observations is approximately normal. A general method to generate useful rank statistics is discussed below. □

A critical value for a test based on a (distribution-free) rank statistic can be found by simply tabulating its null distribution. For a large number of observations this is a bit tedious. In most cases it is also unnecessary, because there exist accurate asymptotic approximations. The remainder of this section is concerned with proving asymptotic normality of simple linear rank statistics under the null hypothesis. Apart from being useful for finding critical values, the theorem is used subsequently to study the asymptotic efficiency of rank tests.

Consider a rank statistic of the form $T_N = \sum_{i=1}^{N} c_{Ni} a_{N,R_{Ni}}$. For a sequence of this type to be asymptotically normal, some restrictions on the coefficients c and scores a are necessary. In most cases of interest, the scores are "generated" through a given function $\phi : [0, 1] \mapsto \mathbb{R}$ in one of two ways. Either

$$a_{Ni} = \mathrm{E}\phi(U_{N(i)}), \tag{13.3}$$

where $U_{N(1)}, \ldots, U_{N(N)}$ are the order statistics of a sample of size N from the uniform distribution on [0, 1]; or

$$a_{Ni} = \phi\left(\frac{i}{N+1}\right). \tag{13.4}$$

For well-behaved functions ϕ, these definitions are closely related and almost identical, because $i/(N+1) = \mathrm{E}U_{N(i)}$. Scores of the first type correspond to the locally most

powerful rank tests that are discussed ahead; scores of the second type are attractive in view of their simplicity.

13.5 ***Theorem.*** *Let R_N be the rank vector of an i.i.d. sample $X_1, \ldots, X_N$ from the continuous distribution function F. Let the scores a_N be generated according to (13.3) for a measurable function ϕ that is not constant almost everywhere, and satisfies $\int_0^1 \phi^2(u)\,du < \infty$. Define the variables*

$$T_N = \sum_{i=1}^{N} c_{Ni} a_{N,R_{Ni}}, \qquad \tilde{T}_N = N\bar{c}_N\bar{a}_N + \sum_{i=1}^{N} (c_{Ni} - \bar{c}_N)\phi\big(F(X_i)\big).$$

Then the sequences T_N and $\tilde{T}_N$ are asymptotically equivalent in the sense that $\mathrm{E}T_N = \mathrm{E}\tilde{T}_N$ and $\operatorname{var}(T_N - \tilde{T}_N)/\operatorname{var} T_N \to 0$. The same is true if the scores are generated according to (13.4) for a function ϕ that is continuous almost everywhere, is nonconstant, and satisfies $N^{-1}\sum_{i=1}^{N}\phi^2\big(i/(N+1)\big) \to \int_0^1 \phi^2(u)\,du < \infty$.

Proof. Set $U_i = F(X_i)$, and view the rank vector R_N as the ranks of the first N elements of the infinite sequence $U_1, U_2, \ldots$. In view of statement (v) of the Lemma 13.1 the definition (13.3) is equivalent to

$$a_{N,R_{Ni}} = \mathrm{E}\big(\phi(U_i) \mid R_N\big).$$

This immediately yields that the projection of $\tilde{T}_N$ onto the set of all square-integrable functions of R_N is equal to $T_N = \mathrm{E}(\tilde{T}_N \mid R_N)$. It is straightforward to compute that

$$\frac{\operatorname{var} T_N}{\operatorname{var} \tilde{T}_N} = \frac{1/(N-1)\sum(c_{Ni} - \bar{c}_N)^2 \sum(a_{Ni} - \bar{a}_N)^2}{\sum(c_{Ni} - \bar{c}_N)^2 \operatorname{var}\phi(U_1)} = \frac{N}{N-1}\frac{\operatorname{var} a_{N,R_{N1}}}{\operatorname{var}\phi(U_1)}.$$

If it can be shown that the right side converges to 1, then the sequences T_N and $\tilde{T}_N$ are asymptotically equivalent by the projection theorem, Theorem 11.2, and the proof for the scores (13.3) is complete.

Using a martingale convergence theorem, we shall show the stronger statement

$$\mathrm{E}\big(a_{N,R_{N1}} - \phi(U_1)\big)^2 \to 0. \tag{13.6}$$

Because each rank vector R_{j-1} is a function of the next rank vector R_j (for one observation more), it follows that $a_{N,R_{N1}} = \mathrm{E}\big(\phi(U_1) \mid R_1, \ldots, R_N\big)$ almost surely. Because ϕ is square-integrable, a martingale convergence theorem (e.g., Theorem 10.5.4 in [42]) yields that the sequence $a_{N,R_{N1}}$ converges in second mean and almost surely to $\mathrm{E}\big(\phi(U_1) \mid R_1, R_2, \ldots\big)$. If $\phi(U_1)$ is measurable with respect to the σ-field generated by $R_1, R_2, \ldots$, then the conditional expectation reduces to $\phi(U_1)$ and (13.6) follows.

The projection of U_1 onto the set of measurable functions of R_{N1} equals the conditional expectation $\mathrm{E}(U_1 \mid R_{N1}) = R_{N1}/(N+1)$. By a straightforward calculation, the sequence $\operatorname{var}\big(R_{N1}/(N+1)\big)$ converges to $1/12 = \operatorname{var} U_1$. By the projection Theorem 11.2 it follows that $R_{N1}/(N+1) \to U_1$ in quadratic mean. Because R_{N1} is measurable in the σ-field generated by $R_1, R_2, \ldots$, for every N, so must be its limit U_1. This concludes the proof that $\phi(U_1)$ is measurable with respect to the σ-field generated by $R_1, R_2, \ldots$ and hence the proof of the theorem for the scores 13.3.

Next, consider the case that the scores are generated by (13.4). To avoid confusion, write these scores as $b_{Ni} = \phi\big(1/(N+1)\big)$, and let a_{Ni} be defined by (13.3) as before. We shall prove that the sequences of rank statistics S_N and T_N defined from the scores a_N and b_N, respectively, are asymptotically equivalent.

Because $R_{N1}/(N+1)$ converges in probability to U_1 and ϕ is continuous almost everywhere, it follows that $\phi\big(R_{N1}/(N+1)\big) \to \phi(U_1)$. The assumption on ϕ is exactly that $\mathrm{E}\phi^2\big(R_{N1}/(N+1)\big)$ converges to $\mathrm{E}\phi^2(U_1)$. By Proposition 2.29, we conclude that $\phi\big(R_{N1}/(N+1)\big) \to \phi(U_1)$ in second mean. Combining this with (13.6), we obtain that

$$\frac{1}{N}\sum_{i=1}^{N}(a_{Ni} - b_{Ni})^2 = \mathrm{E}\left(a_{N,R_{N1}} - \phi\left(\frac{R_{N1}}{N+1}\right)\right)^2 \to 0.$$

By the formula for the variance of a linear rank statistic, we obtain that

$$\frac{\operatorname{var}(S_N - T_N)}{\operatorname{var} T_N} = \frac{\sum_{i=1}^{N}\big(a_{Ni} - b_{Ni} - (\bar{a}_N - \bar{b}_N)\big)^2}{\sum_{i=1}^{N}(a_{Ni} - \bar{a}_N)^2} \to 0,$$

because $\operatorname{var} a_{N,R_{N1}} \to \operatorname{var}\phi(U_1) > 0$. This implies that $\operatorname{var} S_N/\operatorname{var} T_N \to 1$. The proof is complete. ■

Under the conditions of the preceding theorem, the sequence of rank statistics $\sum c_{Ni}a_{N,R_{Ni}}$ is asymptotically equivalent to a sum of independent variables. This sum is asymptotically normal under the Lindeberg-Feller condition, given in Proposition 2.27. In the present case, because the variables $\phi\big(F(X_i)\big)$ are independent and identically distributed, this is implied by

$$\frac{\max_{1\le i\le N}(c_{Ni} - \bar{c}_N)^2}{\sum_{i=1}^{N}(c_{Ni} - \bar{c}_N)^2} \to 0. \tag{13.7}$$

This is satisfied by the most important choices of vectors of coefficients.

13.8 ***Corollary.*** *If the vector of coefficients c_N satisfies* (13.7)*, and the scores are generated according to* (13.3) *for a measurable, nonconstant, square-integrable function ϕ, then the sequence of standardized rank statistics $(T_N - \mathrm{E}T_N)/\operatorname{sd} T_N$ converges weakly to an $N(0,1)$-distribution. The same is true if the scores are generated by* (13.4) *for a function ϕ that is continuous almost everywhere, is nonconstant, and satisfies $N^{-1}\sum_{i=1}^{N}\phi^2\big(i/(N+1)\big) \to \int_0^1 \phi^2(u)\,du$.*

13.9 ***Example (Monotone score generating functions).*** Any nondecreasing, nonconstant function ϕ satisfies the conditions imposed on score-generating functions of the type (13.4) in the preceding theorem and corollary. The same is true for every ϕ that is of bounded variation, because any such ϕ is a difference of two monotone functions.

To see this, we recall from the preceding proof that it is always true that $R_{N1}/(N+1) \to U_1$, almost surely. Furthermore,

$$\mathrm{E}\phi^2\left(\frac{R_{N1}}{N+1}\right) = \frac{1}{N}\sum_{i=1}^{N}\phi^2\left(\frac{i}{N+1}\right) \le \frac{N+1}{N}\sum_{i=1}^{N}\int_{i/(N+1)}^{(i+1)/(N+1)}\phi^2(u)\,du.$$

The right side converges to $\int\phi^2(u)\,du$. Because ϕ is continuous almost everywhere, it follows by Proposition 2.29 that $\phi\big(R_{N1}/(N+1)\big) \to \phi(U_1)$ in quadratic mean. □

13.10 ***Example (Two-sample problem).*** In a two-sample problem, in which the first m observations constitute the first sample and the remaining observations $n = N - m$ the second, the coefficients are usually chosen to be

$$c_{Ni} = \begin{cases} 0 & i = 1, \ldots, m \\ 1 & i = m+1, \ldots, m+n. \end{cases}$$

In this case $\bar{c}_N = n/N$ and $\sum_{i=1}^N (c_{Ni} - \bar{c}_N)^2 = mn/N$. The Lindeberg condition is satisfied provided both $m \to \infty$ and $n \to \infty$. □

13.11 ***Example (Wilcoxon test).*** The function $\phi(u) = u$ generates the scores $a_{Ni} = i/(N+1)$. Combined with "two-sample coefficients," it yields a multiple of the Wilcoxon statistic. According to the preceding theorem, the sequence of Wilcoxon statistics $W_N = \sum_{i=m+1}^N R_{Ni}/(N+1)$ is asymptotically equivalent to

$$\tilde{W}_N = -\frac{n}{N}\sum_{i=1}^m F(X_i) + \frac{m}{N}\sum_{j=1}^n F(Y_j) + N\frac{n}{N}\frac{1}{2}.$$

The expectations and variances of these statistics are given by $EW_N = E\tilde{W}_N = n/2$, $\operatorname{var} W_N = mn/\big(12(N+1)\big)$, and $\operatorname{var} \tilde{W}_N = mn/(12N)$. □

13.12 ***Example (Median test).*** The *median test* is a two-sample rank test with scores of the form $a_{Ni} = \phi\big(i/(N+1)\big)$ generated by the function $\phi(u) = 1_{(0,1/2]}(u)$. The corresponding test statistic is

$$\sum_{i=m+1}^N 1\left\{R_{Ni} \le \frac{N+1}{2}\right\}.$$

This counts the number of Y_j less than the median of the pooled sample. Large values of this test statistic indicate that the distribution of the second sample is stochastically smaller than the distribution of the first sample. □

The examples of rank statistics discussed so far have a direct intuitive meaning as statistics measuring a difference in location. It is not always obvious to find a rank statistic appropriate for testing certain hypotheses. Which rank statistics measure a difference in scale, for instance?

A general method of generating rank statistics for a specific situation is as follows. Suppose that it is required to test the null hypothesis that $X_1, \ldots, X_N$ are i.i.d. versus the alternative that $X_1, \ldots, X_N$ are independent with X_i having a distribution with density $f_{c_{Ni}\theta}$, for a given one-dimensional parametric model $\theta \mapsto f_\theta$. According to the Neyman-Pearson lemma, the most powerful rank test for testing $H_0 : \theta = 0$ against the simple alternative $H_1 : \theta = \theta$ rejects the null hypothesis for large values of the quotient

$$\frac{P_\theta(R_N = r)}{P_0(R_N = r)} = N!\, P_\theta(R_N = r).$$

Equivalently, the null hypothesis is rejected for large values of $P_\theta(R_N = r)$. This test depends on the alternative θ, but this dependence disappears if we restrict ourselves to

alternatives θ that are sufficiently close to 0. Indeed, under regularity conditions,

$$\begin{aligned}
&\mathrm{P}_\theta(R_N = r) - \mathrm{P}_0(R_N = r) \\
&= \int \cdots \int_{R_N = r} \left(\prod_{i=1}^N f_{c_{Ni}\theta}(x_i) - \prod_{i=1}^N f_0(x_i) \right) dx_1 \cdots dx_N \\
&= \theta \int \cdots \int_{R_N = r} \sum_{i=1}^N c_{Ni} \frac{\dot f_0}{f_0}(x_i) \prod_{i=1}^N f_0(x_i)\, dx_1 \cdots dx_N + o(\theta) \\
&= \theta \frac{1}{N!} \sum_{i=1}^N c_{Ni} \mathrm{E}_0 \left(\frac{\dot f_0}{f_0}(X_i) \mid R_N = r \right) + o(\theta).
\end{aligned}$$

Conclude that, for small $\theta > 0$, large values of $\mathrm{P}_\theta(R_N = r)$ correspond to large values of the simple linear rank statistic $T_N = \sum_{i=1}^N c_{Ni} a_{N,R_{Ni}}$, for the vector a_N of scores given by

$$a_{Ni} = \mathrm{E}_0 \frac{\dot f_0}{f_0}\left(X_{N_{(i)}}\right) = \mathrm{E} \frac{\dot f_0}{f_0}\left(F_0^{-1}\left(U_{N_{(i)}}\right)\right).$$

These scores are of the form (13.3), with score-generating function $\phi = (\dot f_0 / f_0) \circ F_0^{-1}$. Thus the corresponding rank statistics are asymptotically equivalent to the statistics $\sum_{i=1}^N c_{Ni} \dot f_0 / f_0(X_i)$.

Rank statistics with scores generated as in the preceding display yield *locally most powerful* rank tests. They are most powerful within the class of all rank tests, uniformly in a sufficiently small neighbourhood $(0, \varepsilon)$ of 0. (For a precise statement, see problem 13.1). Such a local optimality property may seem weak, but it is actually of considerable importance, particularly if the number of observations is large. In the latter situation, any reasonable test can discriminate well between the null hypothesis and "distant" alternatives. A good test proves itself by having high power in discriminating "close" alternatives.

13.13 ***Example (Two-sample scale).*** To generate a test statistic for the two-sample scale problem, let $f_\theta(x) = e^{-\theta} f(e^{-\theta} x)$ for a fixed density f. If X_i has density $f_{c_{Ni}\theta}$ and the vector c is chosen equal to the usual vector of two-sample coefficients, then the first m observations have density $f_0 = f$; the last $n = N - m$ observations have density f_θ. The alternative hypothesis that the second sample has larger scale corresponds to $\theta > 0$. The scores for the locally most powerful rank test are given by

$$a_{Ni} = -\mathrm{E}\left(1 + F^{-1}\left(U_{N_{(i)}}\right) \frac{f'}{f}\left(F^{-1}\left(U_{N_{(i)}}\right)\right)\right).$$

For instance, for f equal to the standard normal density this leads to the rank statistic $\sum_{i=m+1}^N a_{N,R_{Ni}}$ with scores

$$a_{Ni} = \mathrm{E}\Phi^{-1}\left(U_{N_{(i)}}\right)^2 - 1.$$

The same test is found for f equal to a normal density with a different mean or variance. This follows by direct calculation, or alternatively from the fact that rank statistics are location and scale invariant. The latter implies that the probabilities $\mathrm{P}_{\mu,\sigma,\theta}(R_N = r)$ of the rank vector R_N of a sample of independent variables $X_1, \ldots, X_N$ with X_i distributed according to $e^{-\theta} f\left(e^{-\theta}(x - \mu)/\sigma\right)/\sigma$ do not depend on (μ, σ). Thus the procedure to generate locally most powerful scores yields the same result for any (μ, σ). □

13.14 ***Example (Two-sample location).*** In order to find locally most powerful tests for location, we choose $f_\theta(x) = f(x-\theta)$ for a fixed density f and the coefficients c equal to the two-sample coefficients. Then the first m observations have density $f(x)$ and the last $n = N - m$ observations have density $f(x-\theta)$. The scores for a locally most powerful rank test are

$$a_{Ni} = -\mathrm{E}\left(\frac{f'}{f}\left(F^{-1}\left(U_{N(i)}\right)\right)\right).$$

For the standard normal density, this leads to a variation of the van der Waerden statistic. The Wilcoxon statistic corresponds to the logistic density. □

13.15 ***Example (Log rank test).*** The *cumulative hazard function* corresponding to a continuous distribution function F is the function $\Lambda = -\log(1-F)$. This is an important modeling tool in survival analysis. Suppose that we wish to test the null hypothesis that two samples with cumulative hazard functions Λ_X and Λ_Y are identically distributed against the alternative that they are not. The hypothesis of *proportional hazards* postulates that $\Lambda_Y = \theta\Lambda_X$ for a constant θ, meaning that the second sample is a factor θ more "at risk" at any time. If we wish to have large power against alternatives that satisfy this postulate, then it makes sense to use the locally most powerful scores corresponding to a family defined by $\Lambda_\theta = \theta\Lambda_1$. The corresponding family of cumulative distribution functions F_θ satisfies $1 - F_\theta = (1-F_1)^\theta$ and is known as the family of *Lehmann alternatives*. The locally most powerful scores for this family correspond to the generating function

$$\phi(u) = \frac{\partial}{\partial\theta}\log\frac{\partial}{\partial x}(1-F_\theta)(x)_{|\theta=1, x=F_1^{-1}(u)} = 1 - \log(1-u).$$

It is fortunate that the score-generating function does not depend on the baseline hazard function Λ_1. The resulting test is known as the *log rank test*. The test is related to the *Savage test*, which uses the scores

$$a_{N,i} = \sum_{j=N-i+1}^{N}\frac{1}{j} \approx -\log\left(1 - \frac{i}{N+1}\right).$$

The log rank test is a very popular test in survival analysis. Then usually it needs to be extended to the situation that the observations are censored. □

13.16 ***Example (More-sample problem).*** Suppose the problem is to test the hypothesis that k independent random samples $X_1, \ldots, X_{N_1}, X_{N_1+1}, \ldots, X_{N_2}, \ldots, X_{N_{k-1}+1}, \ldots, X_{N_k}$ are identical in distribution. Let $N = N_k$ be the total number of observations, and let R_N be the rank vector of the pooled sample $X_1, \ldots, X_N$. Given scores a_N inference can be based on the rank statistics

$$T_{N1} = \sum_{i=1}^{N_1} a_{N,R_{Ni}}, \quad T_{N2} = \sum_{i=N_1+1}^{N_2} a_{N,R_{Ni}}, \ldots, T_{Nk} = \sum_{i=N_{k-1}+1}^{N_k} a_{N,R_{Ni}}.$$

The testing procedure can consist of several two-sample tests, comparing pairs of (pooled) subsamples, or on an overall statistic. One possibility for an overall statistic is the chi-square

statistic. For $n_j = N_j - N_{j-1}$ equal to the number of observations in the jth sample, define

$$C_N^2 = \sum_{j=1}^{k} \frac{(T_{N_j} - n_j \bar{a}_N)^2}{n_j \operatorname{var} \phi(U_1)}.$$

If the scores are generated by (13.3) or (13.4) and all sample sizes n_j tend to infinity, then every sequence T_{N_j} is asymptotically normal under the null hypothesis, under the conditions of Theorem 13.5. In fact, because the approximations $\tilde{T}_{N_j}$ are jointly asymptotically normal by the multivariate central limit theorem, the vector $T_N = (T_{N1}, \ldots, T_{Nk})$ is asymptotically normal as well. By elementary calculations, if $n_i/N \to \lambda_i$,

$$\frac{T_N - \mathrm{E}T_N}{\sqrt{N}\, \mathrm{sd}\, \phi(U_1)} \rightsquigarrow N_k\left(0, \begin{pmatrix} \lambda_1(1-\lambda_1) & -\lambda_1\lambda_2 & \cdots & -\lambda_1\lambda_k \\ -\lambda_2\lambda_1 & \lambda_2(1-\lambda_2) & \cdots & -\lambda_2\lambda_k \\ \vdots & \vdots & & \vdots \\ -\lambda_k\lambda_1 & -\lambda_k\lambda_2 & \cdots & \lambda_k(1-\lambda_k) \end{pmatrix}\right).$$

This limit distribution is similar to the limit distribution of a sequence of multinomial vectors. Analogously to the situation in the case of Pearson's chi-square tests for a multinomial distribution (see Chapter 17), the sequence C_N^2 converges in distribution to a chi-square distribution with $k-1$ degrees of freedom.

There are many reasonable choices of scores. The most popular choice is based on $\phi(u) = u$ and leads to the *Kruskal-Wallis* test. Its test statistic is usually written in the form

$$\frac{12}{N(N-1)} \sum_{j=1}^{k} n_j \left(\bar{R}_{j\cdot} - \frac{N+1}{2}\right)^2, \qquad \bar{R}_{j\cdot} = \frac{\sum_{i=N_{j-1}+1}^{N_j} R_{Ni}}{n_j}.$$

This test statistic measures the distance of the average scores of the k samples to the average score $(N+1)/2$ of the pooled sample.

An alternative is to use locally asymptotically powerful scores for a family of distributions of interest. Also, choosing the same score generating function for all subsamples is convenient, but not necessary, provided the chi-square statistic is modified. □

13.2 Signed Rank Statistics

The *sign* of a number x, denoted $\operatorname{sign}(x)$, is defined to be -1, 0, or 1 if $x < 0$, $x = 0$ or $x > 0$, respectively. The *absolute rank* R_{Ni}^+ of an observation X_i in a sample $X_1, \ldots, X_N$ is defined as the rank of $|X_i|$ in the sample of absolute values $|X_1|, \ldots, |X_N|$. A simple linear *signed rank statistic* has the form

$$\sum_{i=1}^{N} a_{N, R_{Ni}^+} \operatorname{sign}(X_i).$$

The ordinary ranks of a sample can always be derived from the combined set of absolute ranks and signs. Thus, the vectors of absolute ranks and signs are together statistically more informative than the ordinary ranks. The difference is dramatic if testing the location of a symmetric density of a given form, in which case the class of signed rank statistics contains asymptotically efficient test statistics in great generality.

The main attraction of signed rank statistics is their simplicity, particularly their being distribution-free over the set of all symmetric distributions. Write $|X|$, R_N^+, and $\operatorname{sign}_N(X)$ for the vectors of absolute values, absolute ranks, and signs.

13.17 ***Lemma.*** *Let $X_1, \ldots, X_N$ be a random sample from a continuous distribution that is symmetric about zero. Then*

(i) the vectors $\big(|X|, R_N^+\big)$ and $\operatorname{sign}_N(X)$ are independent;

(ii) the vector R_N^+ is uniformly distributed over $\{1, \ldots, N\}$;

(iii) the vector $\operatorname{sign}_N(X)$ is uniformly distributed over $\{-1, 1\}^N$;

(iv) for any signed rank statistic, $\operatorname{var} \sum_{i=1}^N a_{N,R_{Ni}^+} \operatorname{sign}(X_i) = \sum_{i=1}^N a_{Ni}^2$.

Consequently, for testing the null hypothesis that a sample is i.i.d. from a continuous, symmetric distribution, the critical level of a signed rank statistic can be set without further knowledge of the "shape" of the underlying distribution.

The null hypothesis of symmetry arises naturally in the two-sample problem with paired observations. Suppose that, given independent observations $(X_1, Y_1), \ldots, (X_N, Y_N)$, it is desired to test the hypothesis that the distribution of $X_i - Y_i$ is "centered at zero." If the observations (X_i, Y_i) are exchangeable, that is, the pairs (X_i, Y_i) and (Y_i, X_i) are equal in distribution, then $X_i - Y_i$ is symmetrically distributed about zero. This is the case, for instance, if, given a third variable (usually called "factor"), the observations X_i and Y_i are conditionally independent and identically distributed. For the vector of absolute ranks to be uniformly distributed on the set of all permutations it is necessary to assume in addition that the differences are identically distributed.

For the signs alone to be distribution-free, it suffices, of course, that the pairs are independent and that $\mathrm{P}(X_i < Y_i) = \mathrm{P}(X_i > Y_i) = \frac{1}{2}$ for every i. Consequently, tests based on only the signs have a wider applicability than the more general signed rank tests. However, depending on the model they may be less efficient.

13.18 ***Theorem.*** *Let $X_1, \ldots, X_N$ be a random sample from a continuous distribution that is symmetric about zero. Let the scores a_N be generated according to (13.3) for a measurable function ϕ such that $\int_0^1 \phi^2(u)\,du < \infty$. For F^+ the distribution function of $|X_1|$, define*

$$T_N = \sum_{i=1}^N a_{N,R_{Ni}^+} \operatorname{sign}(X_i), \qquad \tilde{T}_N = \sum_{i=1}^N \phi\big(F^+(|X_i|)\big) \operatorname{sign}(X_i).$$

Then the sequences T_N and $\tilde{T}_N$ are asymptotically equivalent in the sense that $N^{-1}\operatorname{var}(T_N - \tilde{T}_N) \to 0$. Consequently, the sequence $N^{-1/2}T_N$ is asymptotically normal with mean zero and variance $\int_0^1 \phi^2(u)\,du$. The same is true if the scores are generated according to (13.4) for a function ϕ that is continuous almost everywhere and satisfies $N^{-1}\sum_{i=1}^N \phi^2\big(i/(N+1)\big) \to \int_0^1 \phi^2(u)\,du < \infty$.

Proof. Because the vectors $\operatorname{sign}_N(X)$ and $\big(|X|, R_N^+\big)$ are independent and $\mathrm{E}\operatorname{sign}_N(X) = 0$, the means of both T_N and $\tilde{T}_N$ are zero. Furthermore, by the independence and the orthogonality of the signs,

$$\mathrm{E}(\tilde{T}_N - T_N)^2 = N\mathrm{E}\Big(a_{N,R_{N1}^+} - \phi\big(F^+(|X_1|)\big)\Big)^2.$$

The expectation on the right side is exactly the expression in (13.6), evaluated for the special choice $U_1 = F^+\big(|X_1|\big)$. This can be shown to converge to zero as in the proof of Theorem 13.5. ■

13.19 ***Example (Wilcoxon signed rank statistic).*** The *Wilcoxon signed rank statistic* $W_N = \sum_{i=1}^N R_{Ni}^+ \operatorname{sign}(X_i)$ is obtained from the score-generating function $\phi(u) = u$. Large values of this statistic indicate that large absolute values $|X_i|$ tend to go together with positive X_i. Thus large values of the Wilcoxon statistic suggest that the location of the X_i is larger than zero. Under the null hypothesis that $X_1, \ldots, X_N$ are i.i.d. and symmetrically distributed about zero, the sequence $N^{-3/2}W_N$ is asymptotically normal $N(0, 1/3)$. The variance of W_N is equal to $N(2N+1)(N+1)/6$.

The signed rank statistic is asymptotically equivalent to the U-statistic with kernel $h(x_1, x_2) = 1\{x_1 + x_2 > 0\}$. (See problem 12.9.) This connection yields the limit distribution also under nonsymmetric distributions. □

Signed rank statistics that are locally most powerful can be obtained in a similar fashion as locally most powerful rank statistics were obtained in the previous section. Let f be a symmetric density, and let $X_1, \ldots, X_N$ be a random sample from the density $f(\cdot - \theta)$. Then, under regularity conditions,

$$\begin{aligned}
&\mathrm{P}_\theta\big(\operatorname{sign}_N(X) = s, R_N^+ = r\big) - \mathrm{P}_0\big(\operatorname{sign}_N(X) = s, R_N^+ = r\big) \\
&\quad = -\theta\, \mathrm{E}_0 \sum_{i=1}^N \operatorname{sign}(X_i) \frac{f'}{f}\big(|X_i|\big)\big\{\operatorname{sign}_N(x) = s, R_N^+ = r\big\} + o(\theta) \\
&\quad = -\theta \frac{1}{2^N N!} \sum_{i=1}^N s_i \mathrm{E}_0\left(\frac{f'}{f}\big(|X_i|\big) \mid R_{Ni}^+ = r_i\right) + o(\theta).
\end{aligned}$$

In the second equality it is used that $f'/f(x)$ is equal to $\operatorname{sign}(x) f'/f\big(|x|\big)$ by the skew symmetry of f'/f. It follows that *locally most powerful signed rank statistics* for testing f against $f(\cdot - \theta)$ are obtained from the scores

$$a_{Ni} = -\mathrm{E}\frac{f'}{f}\big((F^+)^{-1}(U_{N(i)})\big).$$

These scores are of the form (13.3) with score-generating function $\phi = -(f'/f) \circ (F^+)^{-1}$, whence locally most powerful rank statistics are asymptotically linear by Theorem 13.18. By the symmetry of F, we have $(F^+)^{-1}(u) = F^{-1}\big((u+1)/2\big)$.

13.20 ***Example.*** The Laplace density has score function $f'/f(x) = \operatorname{sign}(x) = 1$, for $x \geq 0$. This leads to the locally most powerful scores $a_{Ni} \equiv 1$. The corresponding test statistic is the *sign statistic* $T_N = \sum_{i=1}^N \operatorname{sign}(X_i)$. Is it surprising that this simple statistic possesses an optimality property? It is shown to be asymptotically optimal for testing $H_0 : \theta = 0$ in Chapter 15. □

13.21 ***Example.*** The locally most powerful score for the normal distribution are $a_{Ni} = \mathrm{E}\Phi^{-1}\big((U_{N_{(i)}} + 1)/2\big)$. These are appropriately known as the normal (absolute) scores. □

13.3 Rank Statistics for Independence

Let $(X_1, Y_1), \ldots, (X_N, Y_N)$ be independent, identically distributed bivariate vectors, with continuous marginal distributions. The problem is to determine whether, within each pair, X_i and Y_i are independent.

Let R_N and S_N be the rank vectors of the samples $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_N$, respectively. If X_i and Y_i are positively dependent, then we expect the vectors R_N and S_N to be roughly parallel. Therefore, rank statistics of the form

$$\sum_{i=1}^{N} a_{N,R_{Ni}} b_{N,S_{Ni}},$$

with a_N and b_N increasing vectors, are reasonable choices for testing independence.

Under the null hypothesis of independence of X_i and Y_i, the vectors R_N and S_N are independent and both uniformly distributed on the permutations of $\{1, \ldots, N\}$. Let R_N^o be the vector of ranks of $X_1, \ldots, X_N$ if first the pairs $(X_1, Y_1), \ldots, (X_N, Y_N)$ have been put in increasing order of $Y_1 < Y_2 < \cdots < Y_N$. The coordinates of R_N^o are called the *antiranks*. Under the null hypothesis, the antiranks are also uniformly distributed on the permutations of $\{1, \ldots, N\}$. By the definition of the antiranks,

$$\sum_{i=1}^{N} a_{N,R_{Ni}} b_{N,S_{Ni}} = \sum_{i=1}^{N} a_{N,R_{Ni}^o} b_{Ni}.$$

The right side is a simple linear rank statistic and can be shown to be asymptotically normal by Theorem 13.5.

13.22 ***Example (Spearman rank correlation).*** The simplest choice of scores corresponds to the generating function $\phi(u) = u$. This leads to the *rank correlation coefficient* ρ_N, which is the ordinary sample correlation coefficient of the rank vectors R_N and S_N. Indeed, because the rank vectors are permutations of the numbers $1, 2, \ldots, N$, their sample mean and variance are fixed, at $(N+1)/2$ and $N(N+1)/12$, respectively, and hence

$$\begin{aligned}\rho_N &= \frac{\sum_{i=1}^{N}(R_{Ni} - \bar{R}_N)(S_{Ni} - \bar{S}_N)}{\left(\sum_{i=1}^{N}(R_{Ni} - \bar{R}_N)^2 \sum_{i=1}^{N}(S_{Ni} - \bar{S}_N)^2\right)^{1/2}} \\ &= \frac{12}{N(N-1)(N+1)} \sum_{i=1}^{N} R_{Ni} S_{Ni} - 3\frac{N+1}{N-1}.\end{aligned}$$

Thus the tests based on the rank correlation coefficient ρ_N are equivalent to tests based on the signed rank statistic $\sum R_{Ni} S_{Ni}$.

It is straightforward to derive from Theorem 13.5 that the sequence $\sqrt{N}\rho_N$ is asymptotically standard normal under the null hypothesis of independence. □

*13.4 Rank Statistics under Alternatives

Let R_N be the rank vector of the independent random variables $X_1, \ldots, X_N$ with continuous distribution functions $F_1, \ldots, F_N$. Theorem 13.5 gives the asymptotic distribution of simple, linear rank statistics under very mild conditions on the score-generating function,

but under the strong assumption that the distribution functions F_i are all equal. This is sufficient for setting critical values of rank tests for the null hypothesis of identical distributions, but for studying their asymptotic efficiency we also need the asymptotic behavior under alternatives. For instance, in the two-sample problem we are interested in the asymptotic distributions under alternatives of the form $F, \ldots, F, G, \ldots, G$, where F and G are the distributions of the two samples.

For alternatives that converge to the null hypothesis "sufficiently fast," the best approach is to use Le Cam's third lemma. In particular, if the log likelihood ratios of the alternatives $F_n, \ldots, F_n, G_n, \ldots, G_n$ with respect to the null distributions $F, \ldots, F, F, \ldots, F$ allow an asymptotic approximation by a sum of the type $\sum \ell_i(X_i)$, then the joint asymptotic distribution of the rank statistics and the log likelihood ratios under the null hypothesis can be obtained from the multivariate central limit theorem and Slutsky's lemma, because Theorem 13.5 yields a similar approximation for the rank statistics. Next, we can apply Le Cam's third lemma, as in Example 6.7, to find the limit distribution of the rank statistics under the alternatives. This approach is relatively easy, and is sufficiently general for most of the questions of interest. See sections 7.5 and 14.1.1 for examples.

More general alternatives must be handled directly and appear to require stronger conditions on the score-generating function. One possibility is to write the rank statistic as a functional of the empirical distribution function $\mathbb{F}_N$, and the weighted empirical distribution $\mathbb{F}_N^c(x) = N^{-1}\sum_{i=1}^N c_{Ni} 1\{X_i \le x\}$ of the observations. Because $R_{Ni} = N\mathbb{F}_N(X_i)$, we have

$$\frac{1}{N}\sum_{i=1}^N c_{Ni} a_{N,R_{Ni}} = \int a_{N,N\mathbb{F}_N(x)}\, d\mathbb{F}_N^c(x).$$

Next, we can apply a von Mises analysis, using the convergence of the empirical distribution functions to Brownian bridges. This method is explained in general in Chapter 20.

In this section we illustrate another method, based on Hájek's projection lemma. To avoid technical complications, we restrict ourselves to smooth score-generating functions. Let $\overline{F}_N$ be the average of $F_1, \ldots, F_N$ and let $\overline{F}_N^c$ be the weighted sum $N^{-1}\sum_{i=1}^N c_{Ni} F_i$, and define

$$T_N = \sum_{i=1}^N c_{Ni}\phi\left(\frac{R_{Ni}}{N+1}\right),$$

$$\hat{T}_N = \sum_{i=1}^N \left[c_{Ni}\phi\big(\overline{F}_N(X_i)\big) + \int_{X_i}^\infty \phi'\big(\overline{F}_N(x)\big)\, d\overline{F}_N^c(x)\right].$$

We shall show that the variables $\hat{T}_N$ are the Hájek projections of approximations to the variables T_N, up to centering at mean zero. The Hájek projections of the variables T_N themselves give a better approximation but are more complicated.

13.23 ***Lemma.*** *If $\phi : [0,1] \mapsto \mathbb{R}$ is twice continuously differentiable, then there exists a universal constant K such that*

$$\operatorname{var}(T_N - \hat{T}_N) \le K\frac{1}{N}\sum_{i=1}^N (c_{Ni} - \bar{c}_N)^2 \left(\|\phi'\|_\infty^2 + \|\phi''\|_\infty^2\right).$$

Proof. Because the inequality is for every fixed N, we delete the index N in the proof. Furthermore, because the assertion concerns a variance and both T_N and $\hat{T}_N$ change by a

constant if the c_{Ni} are replaced by $c_{Ni} - \bar{c}_N$, it is not a loss of generality to assume that $\bar{c}_N = 0$. (Evaluate the integral defining $\hat{T}_N$ to see this.)

The rank of X_i can be written as $R_i = 1 + \sum_{k \neq i} 1\{X_k \leq X_i\}$. This representation and a little algebra show that

$$\left| \mathrm{E}\left(\frac{R_i}{N+1} \,\middle|\, X_i \right) - \overline{F}(X_i) \right| = \frac{1}{N+1} \left| 1 - \overline{F}(X_i) - F_i(X_i) \right| \leq \frac{1}{N}.$$

Furthermore, applying the Marcinkiewitz-Zygmund inequality (e.g., [23, p. 356]) conditionally on X_i, we obtain that

$$\begin{aligned} &\mathrm{E}\left(\frac{R_i}{N+1} - \overline{F}(X_i) \right)^4 \\ &= \frac{1}{(N+1)^4} \mathrm{E}\left(\sum_{k \neq i} \left(1\{X_k \leq X_i\} - F_k(X_i) \right) + 1 - \overline{F}(X_i) - F_i(X_i) \right)^4 \\ &\lesssim \frac{1}{N^2} \mathrm{E}\mathrm{E}\left(\frac{1}{N} \sum_{k \neq i} \left(1\{X_k \leq X_i\} - F_k(X_i) \right)^4 \,\middle|\, X_i \right) + \frac{1}{N^4} \lesssim \frac{1}{N^2}. \end{aligned}$$

Next, developing ϕ in a two-term Taylor expansion around $\overline{F}(X_i)$, for each term in the sum that defines T, we see that there exist random variables K_i that are bounded by $\|\phi''\|_\infty$ such that

$$\begin{aligned} T = &\sum_{i=1}^N c_i \phi\left(\overline{F}(X_i) \right) + \sum_{i=1}^N c_i \left(\frac{R_{Ni}}{N+1} - \overline{F}(X_i) \right) \phi'\left(\overline{F}(X_i) \right) \\ &+ \sum_{i=1}^N c_i \left(\frac{R_{Ni}}{N+1} - \overline{F}(X_i) \right)^2 K_i \qquad =: T_0 + T_1 + T_2. \end{aligned}$$

Using the Cauchy-Schwarz inequality and the fourth-moment bound obtained previously, we see that the quadratic term T_2 is bounded above in second mean as in the lemma. The leading term T_0 is a sum of functions of the single variables X_i, and is the first part of $\hat{T}$. We shall show that the linear term T_1 is asymptotically equivalent to its Hájek projection, which, moreover, is asymptotically equivalent to the second part of $\hat{T}$, up to a constant. The Hájek projection of T_1 is equal to, up to a constant,

$$\begin{aligned} &\sum_i c_i \sum_j \mathrm{E}\left[\frac{R_i}{N+1} \phi'\left(\overline{F}(X_i) \right) \,\middle|\, X_j \right] - \sum_i c_i \overline{F}(X_i) \phi'\left(\overline{F}(X_i) \right) \\ &= \sum_i c_i \left[\sum_{j \neq i} \mathrm{E}\left[\frac{R_i}{N+1} \phi'\left(\overline{F}(X_i) \right) \,\middle|\, X_j \right] \right. \\ &\left. + \sum_i c_i \left(\mathrm{E}\left(\frac{R_i}{N+1} \,\middle|\, X_i \right) - \overline{F}(X_i) \right) \phi'\left(\overline{F}(X_i) \right). \right. \end{aligned}$$

The second term is bounded in second mean as in the lemma; the first term is equal to

$$\frac{1}{N+1} \sum_i c_i \sum_{j \neq i} \mathrm{E}\left(1\{X_j \leq X_i\} \phi'\left(\overline{F}(X_i) \right) \mid X_j \right) + \text{constant}.$$

If we replace $(N+1)$ by N, write out the conditional expectation, add the diagonal terms, and remove the constant, then we obtain the second term in the definition of $\hat{T}$. The difference between these two expressions is bounded above in second mean as in the lemma.

To conclude the proof it suffices to show that the difference between T_1 and its Hájek projection is negligible. We employ the Hoeffding decomposition. Because each of the variables $R_i\phi'\big(\overline{F}(X_i)\big)$ is contained in the space $\sum_{|A|\le 2} H_A$, the difference between T_1 and its Hájek projection is equal to the projection of T_1 onto the space $\sum_{|A|=2} H_A$. This projection has second moment

$$\frac{1}{(N+1)^2}\sum_{|A|=2}\mathrm{E}\left(P_A\sum_i c_i\sum_k 1\{X_k\le X_i\}\phi'\big(\overline{F}(X_i)\big)\right)^2.$$

The projection of the variable $1\{X_k\le X_i\}\phi'\big(\overline{F}(X_i)\big)$, which is contained in the space $H_{\{k,i\}}$, onto the space $H_{\{a,b\}}$ is zero unless $\{a,b\}\subset\{k,i\}$. Thus, the expression in the preceding display is equal to

$$\frac{1}{(N+1)^2}\sum_{a<b}\mathrm{E}\Big(c_b 1\{X_a\le X_b\}\phi'\big(\overline{F}(X_b)\big)+c_a 1\{X_b\le X_a\}\phi'\big(\overline{F}(X_a)\big)\Big)^2.$$

This is bounded by the upper bound of the lemma, as desired. The proof is complete. ■

As a consequence of the lemma, the sequences $(T_N-\mathrm{E}T_N)/\mathrm{sd}\,T_N$ and $(\hat{T}_N-\mathrm{E}\hat{T}_N)/\mathrm{sd}\,\hat{T}_N$ have the same limiting distribution (if any) if

$$\frac{\sum_{i=1}^N(c_{Ni}-\bar{c}_N)^2}{N\mathrm{var}\,\hat{T}_N}\to 0.$$

This condition is certainly satisfied if the observations are identically distributed. Then the rank vector is uniformly distributed on the permutations, and the explicit expression for $\mathrm{var}\,T_N$ given by Lemma 13.1 shows that the left side (with $\mathrm{var}\,T_N$ instead of $\mathrm{var}\,\hat{T}_N$) is of the order $O(1/N)$. Because this leaves much too spare, the condition remains satisfied under small departures from identical distributions, but the general situation requires a calculation.

Under the conditions of the lemma we have the approximation

$$\mathrm{E}T_N\approx\bar{c}_N\sum_{i=1}^N\phi\left(\frac{i}{N+1}\right)+\sum_{i=1}^N(c_{Ni}-\bar{c}_N)\mathrm{E}\phi\big(\overline{F}_N(X_i)\big).$$

The square of the difference is bounded by the upper bound of the lemma.

The preceding lemma is restricted to smooth score-generating functions. One possibility to extend the result to more general scores is to show that the difference between the rank statistics of interest and suitable approximations by rank statistics with smooth scores is small. The following lemma is useful for this purpose, although it is suboptimal if the observations are identically distributed. (For a proof, see Theorem 3.1, in [68].)

13.24 ***Lemma (Variance inequality).*** *For nondecreasing coefficients $a_{N1}\le\cdots\le a_{NN}$ and arbitrary scores $c_{N1},\ldots,c_{NN}$,*

$$\mathrm{var}\sum_{i=1}^N c_{Ni}a_N,R_{Ni}\le 21\max_{1\le i\le N}(c_{Ni}-\bar{c}_N)^2\sum_{i=1}^N(a_{Ni}-\bar{a}_N)^2.$$

13.5 Permutation Tests

Rank tests are examples of *permutation tests*. General permutation tests also possess a distribution-free level but still use the values of the observations next to their ranks. In this section we illustrate this for the two-sample problem.

Suppose that the null hypothesis H_0 that two independent random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are identically distributed is rejected for large values of a test statistic $T_N(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$. Write $Z_{(1)}, \ldots, Z_{(N)}$ for the values of the pooled sample stripped of its original order. ($N = m + n$.) Under the null hypothesis each permutation $Z_{\pi_1}, \ldots, Z_{\pi_N}$ of the N values is equally likely to lead back to the original observations. More precisely, the conditional null distribution of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ given $Z_{(1)}, \ldots, Z_{(N)}$ is uniform on the $N!$ permutations of the latter sample. Thus, it would be reasonable to reject H_0 if the observed value $T_N(x_1, \ldots, x_m, y_1, \ldots, y_n)$ is among the $100\alpha\%$ largest values $T_N(z_{\pi_1}, \ldots, z_{\pi_N})$ as π ranges over all permutations. Then we obtain a test of level α, conditionally given the observed values and hence also unconditionally.

Does this procedure work? Does the test have the desired power? The answer is affirmative for statistics T_N that are sums, in the sense that, asymptotically, the permutation test is equivalent to the test based on the normal approximation to T_N. If the latter test performs well, then so does the permutation test.

We consider statistics of the form, for a given measurable function f,

$$T_N(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = \frac{1}{m}\sum_{i=1}^{m} f(X_i) - \frac{1}{n}\sum_{j=1}^{n} f(Y_j).$$

These statistics include, for instance, the score statistics for testing that the two samples have distributions p_0 and p_θ, respectively, for which we take f equal to the score function $\dot{p}_0/p_0$ of the model. Because a permutation test is conditional on the observed values, and T_N is fixed once $\sum_j f(Y_j)$ and $\sum_i f(Z_i)$ are fixed, it would be equivalent to consider statistics of the form $\sum_j f(Y_j)$.

Let $(\pi_{N1}, \ldots, \pi_{NN})$ be uniformly distributed on the $N!$ permutations of the numbers $1, 2, \ldots, N$, and be independent of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$.

13.25 ***Theorem.*** *Let both* $\mathrm{E}f^2(X_1)$ *and* $\mathrm{E}f^2(Y_1)$ *be finite, and suppose that* $m, n \to \infty$ *such that* $m/N \to \lambda \in (0, 1)$. *Then, given almost every sequence* $X_1, X_2, \ldots, Y_1, Y_2, \ldots$, *the sequence* $\sqrt{N}T_N(Z_{\pi_{N1}}, \ldots, Z_{\pi_{NN}})$ *is asymptotically normal with mean zero. Under the null hypothesis the asymptotic variance is equal to* $\operatorname{var} f(X_1)/\big(\lambda(1-\lambda)\big)$.

Proof. Conditionally on the values of the pooled sample, the statistic $NT_N(Z_{\pi_{N1}}, \ldots, Z_{\pi_{NN}})$ is distributed as the simple linear rank statistic $\sum_{i=1}^{N} c_{Ni} a_{N,R_{Ni}}$ with coefficients and scores

$$c_{Ni} = f(Z_i), \qquad a_{Ni} = \begin{cases} \frac{N}{m}, & i \le m \\ -\frac{N}{n}, & i > m \end{cases}$$

Here $R_{N1}, \ldots, R_{NN}$ are the antiranks of $\pi_{N1}, \ldots, \pi_{NN}$ defined by the equation $\sum c_{N,\pi_{Ni}} a_{Ni} = \sum c_{Ni} a_{N,R_{Ni}}$ (for any numbers c_{Ni} and a_{Ni}).

The coefficients satisfy relation (13.7) for almost every sequence $X_1, X_2, \ldots, Y_1, Y_2, \ldots$, because, by the law of large numbers,

$$\overline{c_N^k} \xrightarrow{\text{as}} \lambda \mathrm{E} f^k(X_1) + (1-\lambda)\mathrm{E} f^k(Y_1), \qquad k = 1, 2,$$
$$\frac{1}{N} \max_{1 \le i \le N} c_{Ni}^2 \xrightarrow{\text{as}} 0.$$

The scores are generated as $a_{Ni} = \phi_N\big(i/(N+1)\big)$ for the functions

$$\phi_N(u) = \begin{cases} \frac{N}{m}, & u \le \frac{m}{N+1}, \\ -\frac{N}{n}, & u > \frac{m}{N+1}. \end{cases}$$

These functions depend on N, unlike the situation of Theorem 13.5, but they converge to the fixed function $\phi = \lambda^{-1} 1_{[0,\lambda)} - (1-\lambda)^{-1} 1_{(\lambda,1]}$. By a minor extension of Theorem 13.5, the sequence $\sum c_{Ni} a_{N,R_{Ni}}$ is asymptotically equivalent to $\sum (c_{Ni} - \bar{c}_N)\phi(U_i)$, for a uniform sample $U_1, \ldots, U_N$. The (asymptotic) variance of the latter variable is easy to compute. ■

By the central limit theorem, under the null hypothesis,

$$\sqrt{N} T_N(X_1, \ldots, X_m, Y_1, \ldots, Y_n) \rightsquigarrow N(0, \sigma^2), \qquad \sigma^2 = \frac{\operatorname{var} f(X_1)}{\lambda(1-\lambda)}.$$

The limit is the same as the conditional limit distribution of the sequence $\sqrt{N} T_N(Z_{\pi_{N1}}, \ldots, Z_{\pi_{NN}})$ under the null hypothesis. Thus, we have a choice of two sequences of tests, both of asymptotic level α, rejecting H_0 if:

- $\sqrt{N} T_N(X_1, \ldots, X_m, Y_1, \ldots, Y_n) \ge z_\alpha \sigma$; or
- $\sqrt{N} T_N(X_1, \ldots, X_m, Y_1, \ldots, Y_n) \ge c_N(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$, where $c_N(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ is the upper α-quantile of the conditional distribution of $\sqrt{N} T_N\ (Z_{\pi_{N1}}, \ldots, Z_{\pi_{NN}})$ given $Z_{(1)}, \ldots, Z_{(N)}$.

The second test is just the permutation test discussed previously. By the preceding theorem the "random critical values" $c_N(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ converge in probability to $z_\alpha \sigma$ under H_0. Therefore the two tests are asymptotically equivalent under the null hypothesis. Furthermore, this equivalence remains under "contiguous alternatives" (for which again $c_N(X_1, \ldots, X_m, Y_1, \ldots, Y_n) \xrightarrow{\text{P}} z_\alpha \sigma$; see Chapter 6), and hence the local asymptotic power functions as discussed in Chapter 14 are the same for the two sequences of tests.

The preceding theorem also shows that the sequence of "critical values" $c_N(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ remains bounded in probability under every alternative. Because $\sqrt{N} T_N(X_1, \ldots, X_m, Y_1, \ldots, Y_n) \rightsquigarrow \infty$ if $\mathrm{E} f(X_1) > \mathrm{E} f(Y_1)$, the power at any alternative with this property converges to 1. Thus, permutation tests are an attractive alternative to both rank and classical tests. Their main drawback is computational complexity. The dependence of the null distribution on the observed values means that it cannot be tabulated and must be computed for every new data set.

*13.6 Rank Central Limit Theorem

The rank central limit theorem Theorem 13.5, is slightly special in that the scores a_{Ni} are assumed to be of one of the forms (13.3) or (13.4). In this section we record what is commonly viewed as the rank central limit theorem. For a proof see [67]. For given coefficients and scores, let

$$C_n^2 = \sum_{i=1}^{n} (c_{Ni} - \bar{c}_N)^2, \qquad A_n^2 = \sum_{i=1}^{n} (a_{Ni} - \bar{a}_N)^2.$$

13.26 ***Theorem (Rank central limit theorem).*** *Let $T_N = \sum c_{Ni} a_{N,R_{Ni}}$ be the simple linear rank statistic with coefficients and scores such that $\max_{1\le i\le N} |a_{Ni} - \bar{a}_N|/A_N \to 0$ and $\max_{1\le i\le N} |c_{Ni} - \bar{c}_N|/C_N \to 0$, and let the rank vector R_N be uniformly distributed on the set of all $N!$ permutations of $\{1, 2, \ldots, N\}$. Then the sequence $(T_N - \mathrm{E}T_N)/\mathrm{sd}\, T_N$ converges in distribution to a standard normal distribution if and only if, for every $\varepsilon > 0$,*

$$\sum\sum_{(i,j):\sqrt{N}|a_{Ni}-\bar{a}_N||c_{Ni}-\bar{c}_N|>\varepsilon A_N C_N} \frac{|a_{Ni} - \bar{a}_N|^2 |c_{Ni} - \bar{c}_N|^2}{A_N^2 C_N^2} \to 0.$$

Notes

The classical reference on rank statistics is the book by Hájek and Šidák [71], which still makes wonderful reading and gives extensive references. Its treatment of rank statistics for nonidentically distributed observations is limited to contiguous alternatives, as in the first sections of this chapter. The papers [43] and [68] remedied this, shortly after the publication of the book. Section 13.4 reports only a few of the results from these papers, which, as does the book, use the projection method. An alternative approach to obtaining the limit distribution of rank statistics, initiated by Chernoff and Savage in the late 1950s and refined many times, is to write them as functions of empirical measures and next apply the von Mises method. We discuss examples of this approach in Chapter 20. See [134] for a more comprehensive treatment and further references.

PROBLEMS

1. This problem asks one to give a precise meaning to the notion of a *locally most powerful test*. Let T_N be a rank statistic based on the "locally most powerful scores." Let $\alpha = \mathrm{P}_0(T_N > c_\alpha)$ for a given number c_α. (Then α is a *natural level* of the test statistic, a level that is attained without randomization.) Then there exists $\varepsilon > 0$ such that the test that rejects the null hypothesis if $T_N > c_\alpha$ is most powerful within the class of all rank tests at level α uniformly in the alternatives $\theta \in (0, \varepsilon)$.
   (i) Prove the statement.
   (ii) Can the statement be extended to arbitrary levels?
2. Find the asymptotic distribution of the median test statistic under the null hypothesis that the two samples are identically distributed and continuous.
3. Show that $\sqrt{n}$ times Spearman's rank correlation coefficient is asymptotically standard normal.
4. Find the scores for a locally most powerful two-sample rank test for location for the Laplace family of densities.

5. Find the scores for a locally most powerful two-sample rank test for location for the Cauchy family of densities.

6. For which density is the Wilcoxon signed rank statistic locally most powerful?

7. Show that Spearman's rank correlation coefficient is a linear combination of Kendall's τ and the U-statistic with (asymmetric) kernel $h(x, y, z) = \text{sign}(x_1 - y_1)\,\text{sign}(x_2 - z_2)$. This decomposition yields another method to prove the asymptotic normality.

8. The symmetrized *Siegel-Tukey test* is a two-sample test with score vector of the form $a_N = (1, 3, 5, \ldots, 5, 3, 1)$. For which type of alternative hypothesis would you use this test?

9. For any a_{Ni} given by (13.3), show that $\bar{a}_N = \int_0^1 \phi(u)\,du$.