Chapter 5. Hypothesis Testing

1 Nested Hypotheses

In this chapter we provide a theoretical discussion on testing of statistical hypotheses. Neyman and Pearson (1933) presented Neyman-Pearson Fundamental Lemma which unfolded the various complex problems in testing statistical hypotheses. In 1928, Neyman and Pearson proposed a general recipe, which is named as the likelihood ratio test, for doing hypothesis testing. We shall treat this test and two related test statistics, each based on the maximum likelihood method in this chapter. For the general case only asymptotic distributions of the test statistics have been established. Let $X_1, \ldots, X_n$ be iid with distribution F_θ belonging to a family $\mathcal{F} = \{F_\theta, \theta \in \Theta\}$, where $\Theta \subset R^s$. Let the distributions F_θ possess densities or mass functions $f(x;\theta)$. Assume that the information matrix $\mathbf{I}(\theta)$ exists and positive definite.

Suppose we are concerned with statistical tests of $r < s$ independent equality restrictions on the $(s \times 1)$ parameter vector θ^0, which we represent by the implicit side relations

$$R_i(\theta) = 0, \quad i = 1, 2, \ldots, r. \tag{1}$$

The vector that satisfy these equations form an $(s-r)$-dimensional subspace Θ^0 of the parameter space Θ, and we shall consider the null hypothesis that θ^0 lies in this subspace.

Or in the set-up of hypotheses testing, we consider the null hypothesis $H_0 : \theta^0 \in \Theta^0$ (to be tested), where Θ^0 is a subset of Θ Θ^0 is determined by a set of r $(\leq s)$ restrictions given by (1). The restrictions under review are thus set within the context of a wider parent model, which provides the maintained hypothesis and defines the alternative hypothesis. In the case of a simple hypothesis $H_0 : \theta = \theta^0$, we have $\Theta^0 = \{\theta^0\}$, and the function $R_i(\theta)$ may be taken to be

$$R_i(\theta) = \theta_i - \theta_i^0, \quad 1 \leq i \leq s.$$

In the case of a composite hypothesis, the set Θ^0 contains more than one element and we necessarily have $r < s$.

In this classical setting the null hypothesis is sometimes called a *nested* hypothesis, with obvious reference to the relative position of Θ^0 and Θ. In the following discussions, θ can range freely over Θ, so that functions of θ are several times over differentiable in respect of all its elements, at least in the neighborhood of θ^0. This property continues to hold under H_0. Since Θ^0 and $\Theta - \Theta^0$ together form the entire well-behaved parameter space Θ, we can differentiate functions of θ at $\theta^0 \in \Theta^0$ in all directions, including those leading to a passage into the alternative parameter space $\Theta - \Theta^0$. We shall make use of this facility in deriving the three tests of this section.

Note that each restriction, $R_i(\theta) = 0$, is in some (re-)parametrization equivalent to putting one parameter equal to zero, or suppressing it. This usually means a simplification of the model, and in this sense hypotheses express *simplifying* assumptions, as is reflected by the corresponding reduction in the number of dimensions of the parameter space. Starting from a loosely specified very general model with a surfeit of parameters we may arrange successive simplifications in a definite order so as to nested subspaces of ever smaller dimensionality within one another. In obeisance to the *principle of parsimony* we should then move down along this sequence, paring down the number of parameters and thus gradually carving the final sparse specification out of the overblown

parent model in which it is concealed, testing all the while to see how far we can go. But simplifying assumptions or nested hypotheses and the need for relevant statistical tests do, of course, also arise outside this specific context in the natural pursuit of parsimony.

2 Constrained Estimation

In previous section, we discuss statistical tests of r independent equality restrictions on the s parameter vector θ^0. The simplest way of estimating θ^0 subject to (1) is to eliminate r parameters. It can be done by adopting a transformation of $\eta = R(\theta)$ while taking care to define the first r elements of the vector function in such a way that they correspond to the restrictions (1) or $R_i(\theta) = 0$ for $1 \leq i \leq r$. The remaining functions $R_i(\theta)$ for $i = r+1, \ldots, s$ can be chosen at will, provided the existence of the inverse transformation $\theta = R^{-1}(\eta)$ is assured. Whenever possible the identity is a popular choice. Under H_0 we now have

$$\eta^0 = \begin{bmatrix} 0 \\ \eta_*^0 \end{bmatrix}$$

and the remaining $s-r$ elements of the subvector η_* can be estimated without constraint. This will yield an estimator $\hat{\eta_*}$ with covariance matrix $Var(\hat{\eta_*})$. The constrained estimator of the full parameter vector is then

$$\hat{\eta} = \begin{bmatrix} 0 & \hat{\eta_*^0} \end{bmatrix}, \quad Var(\tilde{\eta}) = \begin{bmatrix} 0 & 0 \\ 0 & Var(\hat{\eta_*}) \end{bmatrix}.$$

Then we can find the constrained estimator of the original parameter vector θ^0.

This is a practical method of estimation, but it is not very helpful when we wish to examine the asymptotic distribution of the constrained estimate; for this purpose we turn to constrained maximization of the likelihood function by the Lagrange Multiplier method.

To begin with, we rewrite the restrictions (1) as an $r \times 1$ vector function $g_R(\theta) = 0$. Likewise we write

$$G_R(\theta) = \left(\frac{\partial}{\partial \theta_j} R_i(\theta) \right)_{r \times s}.$$

The new maximand is $L(\theta) - g_R(\theta)^T \mu$ with μ a vector of r Lagrange multipliers. Differentiation yields $s + r$ first-order conditions that must be satisfied by the constrained estimators $\tilde{\theta}$ and $\tilde{\mu}$, namely,

$$Q(\tilde{\theta}) - G_R(\tilde{\theta})^T \tilde{\mu} = 0, \quad g_R(\tilde{\theta}) = 0, \tag{2}$$

where

$$Q(\theta)^T = \left(\frac{\partial}{\partial \theta_j} L(\theta) \right)_{1 \times s}.$$

Now we examine these estimators under H_0, when they are appropriate; as MLE are consistent, and as the sample size increases $\tilde{\theta}$ will converge to $\theta^0 \in \Theta^0$. We suppose indeed that $\tilde{\theta}$ is sufficiently close to θ^0 to justify several large-sample approximations, as follows:

$$\begin{aligned} G_R(\tilde{\theta})^T \tilde{\mu} &\approx G_R(\theta^0)^T \tilde{\mu} && (3) \\ Q(\tilde{\theta}) &\approx Q(\theta^0) + H(\theta^0)(\tilde{\theta} - \theta^0) && (4) \\ g_R(\tilde{\theta}) &\approx G_R(\theta^0)(\tilde{\theta} - \theta^0), && (5) \end{aligned}$$

where

$$H(\theta) = \left(\frac{\partial^2}{\partial\theta_j \partial\theta_k} L(\theta) \right)_{s\times s}.$$

H_0 is used in (5) where we take it that $g_R(\theta^0) = 0$.

Upon substitution of these approximations into (2) and some rearrangement of the terms we obtain a system of $r + s$ simultaneous linear equations

$$\begin{bmatrix} -H(\theta^0) & G_R(\theta^0)^T \\ G_R(\theta^0) & 0 \end{bmatrix} \begin{bmatrix} \tilde{\theta} - \theta^0 \\ \tilde{\mu} \end{bmatrix} \approx \begin{bmatrix} Q(\theta^0) \\ 0 \end{bmatrix}$$

or again

$$\begin{bmatrix} -\frac{1}{n} H(\theta^0) & G_R(\theta^0)^T \\ G_R(\theta^0) & 0 \end{bmatrix} \begin{bmatrix} \sqrt{n}(\tilde{\theta} - \theta^0) \\ \frac{1}{\sqrt{n}}\tilde{\mu} \end{bmatrix} \approx \begin{bmatrix} \frac{1}{\sqrt{n}} Q(\theta^0) \\ 0 \end{bmatrix}. \tag{6}$$

It is known that $-n^{-1}H(\theta^0) \xrightarrow{P} \mathbf{I}(\theta^0)$ by the law of large numbers. Upon substitution in (6) this gives

$$\begin{bmatrix} \sqrt{n}(\tilde{\theta} - \theta^0) \\ \frac{1}{\sqrt{n}}\tilde{\mu} \end{bmatrix} \approx \begin{bmatrix} \mathbf{I}(\theta^0) & G_R(\theta^0)^T \\ G_R(\theta^0) & 0 \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{\sqrt{n}} Q(\theta^0) \\ 0 \end{bmatrix}. \tag{7}$$

It follows directly from the multivariate Lindberg-Levy CLT that

$$\frac{1}{\sqrt{n}} Q(\theta^0) \xrightarrow{d} N(0, \mathbf{I}(\theta^0)).$$

Observe that

$$\begin{bmatrix} \mathbf{I}(\theta^0) & G_R(\theta^0)^T \\ G_R(\theta^0) & 0 \end{bmatrix}^{-1}$$
$$= \begin{bmatrix} \mathbf{I}^{-1}(\theta^0) - \mathbf{I}^{-1}(\theta^0) G_R(\theta^0)^T [A(\theta^0)]^{-1} G_R(\theta^0) \mathbf{I}^{-1}(\theta^0) & \mathbf{I}^{-1}(\theta^0) G_R(\theta^0)^T [A(\theta^0)]^{-1} \\ A^{-1}(\theta^0) G_R(\theta^0) \mathbf{I}^{-1}(\theta^0) & [A(\theta^0)]^{-1} \end{bmatrix},$$

where $A(\theta^0) = G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T$. It follows that the $r + s$ vector on the left of (7) is also asymptotically normal with zero mean and with covariance matrix

$$\begin{bmatrix} \mathbf{I}(\theta^0) & G_R(\theta^0)^T \\ G_R(\theta^0) & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{I}(\theta^0) & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{I}(\theta^0) & G_R(\theta^0)^T \\ G_R(\theta^0) & 0 \end{bmatrix}^{-1} \tag{8}$$
$$= \begin{bmatrix} \mathbf{I}^{-1}(\theta^0) - \mathbf{I}^{-1}(\theta^0) G_R(\theta^0)^T [A(\theta^0)]^{-1} G_R(\theta^0) \mathbf{I}^{-1}(\theta^0) & \cdots \\ \cdots & [G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T]^{-1} \end{bmatrix}.$$

This yields that the asymptotic variance of $\sqrt{n}(\tilde{\theta} - \theta^0)$ is

$$\mathbf{I}^{-1}(\theta^0) - \mathbf{I}^{-1}(\theta^0) G_R(\theta^0)^T \left[G_R(\theta^0)\mathbf{I}^{-1}(\theta^0) G_R(\theta^0)^T \right]^{-1} G_R(\theta^0)\mathbf{I}^{-1}(\theta^0) \tag{9}$$
$$= \mathbf{I}^{-1/2}(\theta^0) \left\{ I - \mathbf{I}^{-1/2}(\theta^0) G_R(\theta^0)^T \left[G_R(\theta^0)\mathbf{I}^{-1}(\theta^0) G_R(\theta^0)^T \right]^{-1} G_R(\theta^0)\mathbf{I}^{-1/2}(\theta^0) \right\} \mathbf{I}^{-1/2}(\theta^0).$$

The matrix in parentheses has the same structure, and hence much the same properties, as the "projection matrix" of the linear regression model in its generalized least squares version. It is idempotent, as is readily verified, and hence its rank equals to its trace. This trace is the difference of the traces of two terms. The first is a unit matrix of order

s, with trace s; the second term is itself idempotent and of rank r, since it includes $G_R(\theta^0)$, and hence of trace r. Altogether the rank of (9) is $s - r$.

As for $\tilde{\mu}$, we seldom explicitly determine these estimates, and their is little interest in their asymptotic covariance matrix; for the discussion of the Lagrange multiplier test we mention that the asymptotic variance of $n^{-1/2}\tilde{\mu}$ is

$$\left[G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T\right]^{-1}. \tag{10}$$

3 Hypothesis Testing By Likelihood Methods

Let H_0 denote a null hypothesis to be tested. Typically, we may represent H_0 as a specified family $\mathcal{F}_0$ of distributions for the data. For any test procedure T, we shall denote by T_n the version based on a sample of size n. The function

$$\gamma_n(T, F) = P_F(T_n \text{ rejects } H_0),$$

defined for distribution function F, is called the *power function* of T_n (or of T). For $F \in \mathcal{F}_0$, $\gamma_n(T, F)$ represents the probability of a Type I error. The quantity

$$\alpha_n(T, F) = \sup_{F \in \mathcal{F}_0} \gamma_n(T, F)$$

is called the *size* of the test. For $F \notin \mathcal{F}_0$, the quantity

$$\beta_n(T, F) = 1 - \gamma_n(T, F)$$

represents the probability of a Type II error. Usually, attention is confined to *consistent* tests: for fixed $F \notin \mathcal{F}_0$, $\beta_n(T, F) \to 0$ as $n \to \infty$. Also, usually attention is confined to *unbiased* tests: for $F \notin \mathcal{F}_0$, $\gamma_n(T, F) \geq \alpha_n(T, \mathcal{F}_0)$.

A general way to compare two such test procedures is through their power functions. In this regard we shall use the concept of *asymptotic relative efficiency* (ARE). For two test procedures T_A and T_B, suppose that a performance criterion is tightened in such a way that the respective sample sizes n_A and n_B for T_A and T_B to perform "equivalently" tend to ∞ but have ratio n_A/n_B tending to some limit. Then the limit represents the ARE of procedure T_B relative to procedure T_A and is denoted by $e(T_B, T_A)$.

The earliest approach to ARE was introduced by Pitman (1949). In this approach, two tests sequences $T = \{T_n\}$ and $U = \{U_n\}$ are compared as the Type I and Type II error probabilities tend to positive limits α and β, respectively. In order that $\alpha_n \to \alpha > 0$ and simultaneously $\beta_n \to \beta > 0$, it is necessary to consider $\beta_n(\cdot)$ evaluated at an alternative $F^{(n)}$ converging at a suitable rate to the null hypothesis $\mathcal{F}_0$.

In justification of this approach, we might argue that large sample sizes would be relevant in practice only if the alternative of interest were close to the null hypothesis and thus hard to distinguish with only a small sample.

3.1 Test Statistics for A Simple Null Hypothesis

Although the theory of this section is of most value for composite null hypotheses, it is convenient to begin with simple null hypothesis. Consider testing $H_0 : \theta = \theta^0$.

A *likelihood ratio* statistic,

$$\Lambda_n = \frac{Lik(\theta^0; \mathbf{x})}{\sup_{\theta \in \Theta} Lik(\theta; \mathbf{x})}$$

was introduced by Neyman and Pearson (1928). Clearly, Λ_n takes values in the interval $[0, 1]$ and H_0 is to be rejected for sufficiently small values of Λ_n. Equivalently, the test may be carried out in terms of the statistic

$$\lambda_n = -2 \log \Lambda_n.$$

For finite n, the null distribution of λ_n will generally depend on n and on the form of pdf of X. However, there is for regular problems a uniform limiting result as $n \to \infty$. It turns out to be more convenient for asymptotic considerations.

Expanding λ_n in a Taylor series, we get

$$\begin{aligned} \lambda_n &= -2 \left\{ -\sum_{i=1}^{n} \log f(X_i, \hat{\theta}) + \sum_{i=1}^{n} \log f(X_i, \theta^0) \right\} \\ &= 2 \left\{ \frac{1}{2} (\theta^0 - \hat{\theta})^T \left(-\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(x; \theta) \Big|_{\theta=\theta^*} \right) (\theta^0 - \hat{\theta}) \right\}, \end{aligned}$$

where $\hat{\theta}$ lies between $\hat{\theta}$ and θ^0. Since θ^* is consistent,

$$\lambda_n = n(\hat{\theta} - \theta^0)^T \left(-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta_j \partial \theta_k} L(\theta) \Big|_{\theta=\theta_0} \right) (\hat{\theta} - \theta^0) + o_P(1).$$

By the asymptotic normality of $\hat{\theta}$ and the convergence of $-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta_j \partial \theta_k} L(\theta)|_{\theta=\theta^0}$ to $\mathbf{I}(\theta^0)$. λ_n has, under H_0, a limiting chi-squared distribution on s degrees of freedom.

Let $\hat{\theta}_n$ denote a consistent, asymptotically normal, and asymptotically efficient sequence of solutions of the likelihood equations. Denote the efficient scores

$$q(\theta) = (q_1(x; \theta), \ldots, q_s(x; \theta))^T$$

where

$$q_j(x; \theta) = \frac{\partial}{\partial \theta_j} \log f(x; \theta).$$

Replace the matrix $\left(-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta_j \partial \theta_k} L(\theta)|_{\theta=\theta^0} \right)$ by $\mathbf{I}(\hat{\theta}_n)$, we get a second statistic,

$$W_n = n(\hat{\theta}_n - \theta^0)^T \mathbf{I}(\hat{\theta}_n)(\hat{\theta}_n - \theta^0),$$

which was introduced by Wald (1943).

Write $Q(\theta) = \sum_{i=1}^{n} q(X_i; \theta)$. A second large-sample equivalent to λ_n can be obtained by

$$\hat{\theta} - \theta^0 = \mathbf{I}^{-1}(\theta^0) Q(\theta^0) + o_P(n^{-1/2}).$$

A third statistic,

$$V_n = n[n^{-1} Q(\theta^0)]^T \mathbf{I}^{-1}(\theta^0)[n^{-1} Q(\theta^0)] = n^{-1} Q(\theta^0)^T \mathbf{I}^{-1}(\theta^0) Q(\theta^0),$$

was introduced by Rao (1948).

The three statistics differ somewhat in computational features. Note that Rao's statistic does not require explicit computation of the maximum likelihood estimates. Nevertheless all three statistics have the same limit chi-squared distribution with degree of freedom s under the null hypothesis. The limiting distribution can be found by the following lemma.

Lemma 1 *Under regularity conditions,*
(i) $n^{-1/2}Q(\theta^0) \xrightarrow{d} N(0, \mathbf{I}(\theta^0))$; *(ii)* $n^{1/2}(\hat{\theta} - \theta^0) \xrightarrow{d} N(0, \mathbf{I}^{-1}(\theta^0))$;
(iii) $n(\hat{\theta}_n - \theta^0)^T \mathbf{I}(\theta)(\hat{\theta}_n - \theta^0) \xrightarrow{d} \chi^2_s$; *(iv)* $n^{-1}Q(\theta^0)^T \mathbf{I}^{-1}(\theta^0)Q(\theta^0) \xrightarrow{d} \chi^2_s$;
(v) $\lambda_n - n(\hat{\theta}_n - \theta^0)^T \mathbf{I}(\theta)(\hat{\theta}_n - \theta^0) \xrightarrow{P} 0$.

Remark. Consider a sequence of n independent trials, with s possible outcomes for each trials. Let θ_j denote the probability of occurrence of the jth outcome in any given trial. Let N_j denote the number of occurrences of the jth outcome in the series of n trials. The MLE of θ_j's are N_j/n. The three test statistics λ_n, W_n and V_n for testing $H_0 : \theta = \theta^0$ against $H_A : \theta \neq \theta^0$ are easily seen to be

$$\begin{aligned} \lambda_n &= 2\sum_{j=1}^{s} N_j \log(\frac{N_j}{n\theta_j^0}), \\ W_n &= \sum_{j=1}^{s} \frac{(N_j - n\theta_j^0)^2}{N_j}, \\ V_n &= \sum_{j=1}^{s} \frac{(N_j - n\theta_j^0)^2}{n\theta_j^0}. \end{aligned}$$

Both W_n and V_n are referred to as chi-squared goodness of fit statistics; the latter often called the Pearson chi-squared distribution. The large sample properties was first derived by Pearson (1900).

Let us now consider the behavior of λ_n, W_n and V_n under "local" alternatives, that is, for a sequence $\{\theta_n\}$ of the form

$$\theta_n = \theta_0 + n^{-1/2}\Delta,$$

where $\Delta = (\Delta_1, \ldots, \Delta_s)^T$. Suppose that the convergences expressed in the above lemma may be established uniformly in Θ for θ in a neighborhood of θ^0. It then would follow that

$$\begin{aligned} n^{1/2}(\hat{\theta} - \theta^0) &= n^{1/2}(\hat{\theta} - \theta_n) + \Delta \xrightarrow{d} N(\Delta, \mathbf{I}^{-1}(\theta^0)), \\ n^{-1/2}Q(\theta^0) &= n^{1/2}(\hat{\theta} - \theta_n)\mathbf{I}(\theta^0) + o_{P_{\theta_n}}(1) \xrightarrow{d} N(\mathbf{I}(\theta^0)\Delta, \mathbf{I}(\theta^0)), \end{aligned}$$

and

$$\lambda_n - W_n \xrightarrow{P_{\theta_n}} 0,$$

It then follow that the statistics λ_n, W_n and V_n each converge in distribution to $\chi^2_s(\Delta^T \mathbf{I}(\theta^0)\Delta)$.

Therefore, under appropriate regularity conditions, the statistics λ_n, W_n and V_n are asymptotically *equivalent* in distribution, both under the null hypothesis and under local alternatives converging sufficiently fast. However, at fixed alternatives these equivalences are not anticipated to hold.

3.2 Likelihood Confidence Region

Due to the duality between tests and confidence regions, families of tests can generate confidence regions. Let $\{\delta(\mathbf{x}, \theta)\}$ be a family of tests such that $\{\delta(\mathbf{x}, \theta^0)\}$ is a test of level of significance α for testing $H_0 : \theta = \theta^0$. Note that $\delta(\mathbf{x}, \theta)$ takes values either 1 or 0. When $\delta(\mathbf{x}, \theta) = 1$, we reject. Otherwise, we cannot reject H_0.

Define the subset $C(\mathbf{x})$ of Θ by $C(\mathbf{x}) = \{\theta : \delta(\mathbf{x}, \theta) = 0\}$. This is just the set of all θ that would not be rejected if we observe $\mathbf{X} = \mathbf{x}$ and used the given family of tests. Based on the likelihood ratio tests, we can construct a $1-\alpha$ likelihood-based confidence region of θ by $C(\mathbf{x}) = \{\theta : \lambda_n \leq \chi_s^2(1-\alpha)\}$.

Remark. Under regularity conditions, asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta^0)$ is usually normal which leads to confidence ellipsoids for θ. However, the confidence regions of θ obtained by the likelihood ratio tests are not necessarily ellipsoids.

3.3 Likelihood Ratio Test

This test is based on the likelihood function. If H_0 holds, and hence $\theta^0 \in \Theta^0 = \{\theta \in \Theta, R_i(\theta) = 0, 1 \leq i \leq r\}$, the unconstrained maximum $L(\hat{\theta})$ should be close to the constrained maximum $L(\tilde{\theta})$. We therefore consider the *likelihood ratio*

$$\Lambda_n = \frac{\sup_{\theta \in \Theta^0} Lik(\theta; \mathbf{x})}{\sup_{\theta \in \Theta} Lik(\theta; \mathbf{x})} = \frac{Lik(\tilde{\theta}; \mathbf{x})}{Lik(\hat{\theta}; \mathbf{x})}.$$

As all likelihoods are positive, and as the constrained maximum cannot exceed the unconstrained maximum, $0 < \Lambda_n \leq 1$. Equivalently, we use the quantity

$$\lambda_n = -2 \log \Lambda_n = 2[L(\hat{\theta}) - L(\tilde{\theta})]$$

is therefore always nonnegative, and we shall show that under H_0 it is asymptotically distributed as chi-square with r degrees of freedom, or

$$\lambda_n \xrightarrow{d} \chi_r^2$$

λ_n can thus serve as a test statistic for H_0.

Recall that $Q(\theta) = \sum_{i=1}^n q(x_i; \theta)$. Once more we examine a Taylor series expansion

$$L(\tilde{\theta}) - L(\hat{\theta}) = Q(\hat{\theta})^T(\tilde{\theta} - \hat{\theta}) + \frac{1}{2}(\tilde{\theta} - \hat{\theta})^T H(\hat{\theta})(\tilde{\theta} - \hat{\theta}) + o_P(1), \tag{11}$$

where

$$H(\theta) = \left(\sum_{i=1}^n \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(x_i; \theta) \right)_{s \times s}.$$

The first term on the right-hand side vanishes since $Q(\hat{\theta}) = 0$; by the consistency of $\hat{\theta}$

$$-\frac{1}{n} H(\hat{\theta}) \xrightarrow{P} \mathbf{I}(\theta^0).$$

We simplify (11) accordingly, and substitute the result in λ_n. This gives

$$\lambda_n = [\sqrt{n}(\tilde{\theta} - \hat{\theta})]^T \mathbf{I}(\theta^0)[\sqrt{n}(\tilde{\theta} - \hat{\theta})] + o_P(1). \tag{12}$$

Then it follows from (7) that under H_0

$$\sqrt{n}(\tilde{\theta} - \theta^0) \approx \mathbf{I}^{-1}(\theta^0) \left\{ I - G_R(\theta^0)^T [G_R(\theta^0) \mathbf{I}^{-1}(\theta^0) G_R(\theta^0)^T]^{-1} G_R(\theta^0) \mathbf{I}^{-1}(\theta^0) \right\} \frac{1}{\sqrt{n}} Q(\theta^0)$$

and that

$$\sqrt{n}(\hat{\theta} - \theta^0) = \mathbf{I}^{-1}(\theta^0) \frac{1}{\sqrt{n}} Q(\theta^0) + o_P(1)$$

so that

$$\sqrt{n}(\tilde{\theta}-\hat{\theta}) \approx -\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T[G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T]^{-1}G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)\frac{1}{\sqrt{n}}Q(\theta^0). \quad (13)$$

Recall that

$$\frac{1}{\sqrt{n}}Q(\theta^0) \xrightarrow{d} N(0,\mathbf{I}(\theta^0))$$

so that

$$\frac{1}{\sqrt{n}}\mathbf{I}^{-1/2}(\theta)Q(\theta^0) \xrightarrow{d} \epsilon \sim N(0,I). \quad (14)$$

Here ϵ is a vector of s standard normal variate that are uncorrelated and hence independent. By (13) and (14), moreover,

$$\sqrt{n}(\tilde{\theta}-\hat{\theta}) \approx -\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T[G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T]^{-1}G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)\mathbf{I}^{1/2}(\theta^0)\epsilon.$$

Upon substituting this into (12) we finally obtain

$$\lambda_n \approx \epsilon^T\mathbf{I}^{-1/2}(\theta^0)G_R(\theta^0)^T[G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T]^{-1}G_R(\theta^0)\mathbf{I}^{-1/2}(\theta^0)\epsilon. \quad (15)$$

This is a quadratic form in independent standard Normal variates, with a nonstochastic idempotent coefficient matrix that is of rank r because of the order of $G_R(\theta^0)$; and it is therefore distributed as a chi-square with r degrees of freedom. And this is what we set out to prove.

3.4 Wald's Test

Note that $R_i(\theta) = \theta_i - \theta_i^0$, $1 \le i \le s$, for the simple null hypothesis $H_0 : \theta = \theta^0$. For the composite hypotheses, this test is based on the vector $\mathbf{b}_\theta = (R_1(\theta), \ldots, R_r(\theta))^T$ and the estimate $\hat{\theta}$ which maximizes $L(\theta)$ without subjecting to the restrictions. If H_0 holds the $R_i(\theta^0)$ are zero, and the $R_i(\hat{\theta})$ should presumably be close to zero. Recall that $\sqrt{n}(\hat{\theta}-\theta^0)$ under H_0 is asymptotical normal with mean zero and covariance $\mathbf{I}^{-1}(\theta^0)$. This implies that $\sqrt{n}\mathbf{b}_{\hat{\theta}}$ under H_0 is asymptotical normal with mean zero ($\mathbf{b}_{\theta^0} = 0$) and covariance

$$G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T.$$

Under H_0, and we use this in the quadratic form

$$\begin{aligned}\sqrt{n}\mathbf{b}_{\hat{\theta}}^T\left[G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T\right]^{-1}\mathbf{b}_{\hat{\theta}}\sqrt{n}\\ = \quad n\mathbf{b}_{\hat{\theta}}^T\left[G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T\right]^{-1}\mathbf{b}_{\hat{\theta}} \xrightarrow{d} \mathcal{X}_r^2.\end{aligned}$$

Since

$$\left[G_R(\hat{\theta})\mathbf{I}^{-1}(\hat{\theta})G_R(\hat{\theta})^T\right]^{-1} \xrightarrow{P} \left[G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T\right]^{-1},$$

we thus have

$$W_n = n\mathbf{b}_{\hat{\theta}}^T\left\{G_R(\hat{\theta})\mathbf{I}^{-1}(\hat{\theta})G_R(\hat{\theta})^T\right\}^{-1}\mathbf{b}_{\hat{\theta}} \xrightarrow{d} \mathcal{X}_r^2. \quad (16)$$

by the consistency of $\hat{\theta}$ and Slutsky's theorem. This establishes the asymptotic distribution of Wald's test statistic under the null hypothesis.

We can therefore construct the test statistic from the original unconstrained estimate $\hat{\theta}$ and its covariance matrix estimate with the help of the restriction function R_i of (1) and their first derivatives which form G_R.

3.5 Lagrange Multiplier Test

This test is also known as the Rao efficient score test or as the chi-square test. It is based on the score vector $n^{-1}Q(\theta)$ and the estimate $\tilde{\theta}$ which maximizers $L(\theta)$ subject to the restrictions $R_i(\theta) = 0$, $1 \leq i \leq r$. When we evaluate this vector at the constrained estimate $\tilde{\theta}$ the result is $n^{-1}Q(\tilde{\theta})$, and under H_0 this should be close to the value of the score vector at the unconstrained estimate, which is of course zero. Rao (1948) introduced the statistic

$$V_n = n^{-1}Q(\tilde{\theta})^T[var(Q(\tilde{\theta}))]^{-1}Q(\tilde{\theta}).$$

Before we work on the details, we will motivate this test statistic first. Assume that the specification of Θ_0 may be equivalently be given as functions of β where $\beta = (\beta_1, \ldots, \beta_{s-r})$ ranges through an open subset in R^{s-r}. In terms of the MLE $\hat{\beta}$ of β, $\tilde{\theta}$ may be represented as

$$\tilde{\theta} = \mathbf{g}(\hat{\beta}) = (g_1(\hat{\beta}), \ldots, g_s(\hat{\beta})).$$

Denoting by $\mathbf{J}(\beta)$ the information matrix for the β-formulation of the model and

$$t_\beta = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(X_i; \mathbf{g}(\beta))}{\partial \beta_1}, \cdots, \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(X_i; \mathbf{g}(\beta))}{\partial \beta_{s-r}} \right)^T.$$

Then Rao's statistic is

$$V_n = n t_{\hat{\beta}}^T \mathbf{J}^{-1}(\hat{\beta}) t_{\hat{\beta}}$$

or in the original formulation

$$V_n = n^{-1}Q(\tilde{\theta})^T[var(Q(\tilde{\theta}))]^{-1}Q(\tilde{\theta}).$$

In order to examine the asymptotic distribution of $Q(\tilde{\theta})$ we once more start off with a Taylor series

$$Q(\tilde{\theta}) = Q(\hat{\theta}) + H(\hat{\theta})(\tilde{\theta} - \hat{\theta}) + o_P(\sqrt{n}) \tag{17}$$

and use $Q(\hat{\theta}) = 0$ and $n^{-1}H(\hat{\theta}) = -\mathbf{I}(\theta^0) + o_P(1)$, as before. This yields

$$\frac{1}{\sqrt{n}}Q(\tilde{\theta}) = \mathbf{I}(\theta^0)\sqrt{n}(\hat{\theta} - \tilde{\theta}) + o_P(1)$$

so that the asymptotic variance of $\frac{1}{\sqrt{n}}Q(\tilde{\theta})$ is

$$G_R(\theta^0)^T[G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T]^{-1}G_R(\theta^0)$$

by (13) and (14). These conclude

$$\begin{aligned}
&\frac{1}{\sqrt{n}}Q(\tilde{\theta})^T \left\{ G_R(\theta^0)^T[G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T]^{-1}G_R(\theta^0) \right\}^{-1} Q(\tilde{\theta})\frac{1}{\sqrt{n}} \\
&= \sqrt{n}(\hat{\theta} - \tilde{\theta})^T \left\{ G_R(\theta^0)^T[G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T]^{-1}G_R(\theta^0) \right\}^{-1} (\tilde{\theta} - \hat{\theta})\sqrt{n} + o_P(1) \\
&= \epsilon^T \mathbf{I}^{1/2}(\theta^0) \left\{ G_R(\theta^0)^T[G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T]^{-1}G_R(\theta^0) \right\}^{-1} \mathbf{I}^{1/2}(\theta^0)\epsilon + o_P(1) \qquad (18)
\end{aligned}$$

and this by (15) is the likelihood ratio test statistic, which we know to be asymptotically chi-square (r) distributed. This will also hold if we replace $\mathbf{I}(\theta^0)$ in (18) by a consistent estimate, as in the test statistic.

$$V_n = \frac{1}{n}Q(\tilde{\theta})^T \left\{ G_R(\tilde{\theta})^T[G_R(\tilde{\theta})\mathbf{I}^{-1}(\tilde{\theta})G_R(\tilde{\theta})^T]^{-1}G_R(\tilde{\theta}) \right\}^{-1} Q(\tilde{\theta}). \tag{19}$$

This is based on the second expression in (18), and we have replaced θ^0 by the constrained estimator $\tilde{\theta}$ - just as the score vectors Q here take their form from the unconstrained estimation problem, but their values from the constrained estimator $\tilde{\theta}$. In order to calculate the test statistic, $\tilde{\theta}$ is the only estimate we need determine.

By (2) and (7)

$$Q(\tilde{\theta}) \approx G_R(\tilde{\theta})^T \tilde{\mu}$$

and

$$\tilde{\mu} \approx [G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)G_R(\theta^0)^T]^{-1} G_R(\theta^0)\mathbf{I}^{-1}(\theta^0)Q(\theta^0)$$

so that

$$V_n \approx \frac{1}{\sqrt{n}}\tilde{\mu}^T \left[G_R(\tilde{\theta})\mathbf{I}^{-1}(\tilde{\theta})G_R(\tilde{\theta})^T\right] \frac{1}{\sqrt{n}}\tilde{\mu}.$$

Note that $[G_R(\tilde{\theta})\mathbf{I}^{-1}(\tilde{\theta})G_R(\tilde{\theta})^T]^{-1}$ is the asymptotic variance of $\frac{1}{\sqrt{n}}\tilde{\mu}$. Under H_0 the likelihood ratio test statistic may be regarded as a standardized quadratic form in the estimated Lagrange Multipliers. This may explain the usual name of the test.

3.6 Discussion

All we have shown is that *under the null hypothesis* the three test statistics have the same asymptotic distribution; in the alternative case they need not at all have the same asymptotic behavior. Again, in a finite sample the three test statistics usually take different values, and at the present level of generality nothing can be said about their exact distribution, under the null hypothesis or otherwise. Without further knowledge of the finite-sample properties of the tests in a particular application they are of little use, since it is hard to tell what significance levels from an asymptotic distribution mean if the evidence comes from a small sample. In certain applications it has, however, proved possible to derive the exact distribution of one or other of the present test statistics or of its transformation. Failing this, the finite-sample performance of the tests may always be established by simulation studies.

In certain cases the asymptotic test statistic is thus taken as the starting point of a further investigation, as a suggestion for an exact approach. This applies in particular to the Likelihood Ratio test; as it is the oldest of the three it has been studied more intensively than the other two. It was introduced by Fisher in the 1920s and adopted in Neyman and Pearson's testing methodology of a decade later. The next test was that of Wald (1943), and the most recent is the Lagrange Multiplier test (Rao 1948).

If we do not have the benefit of further analyses, and a large sample that inspires some confidence in the validity of asymptotic results, the choice between the three tests may be affected by expediency. Each requires different computations. For the likelihood ratio test we must maximize the likelihood function twice, with and without restrictions; for the Wald test we need unconstrained parameter estimates with their covariance matrix; and for Rao's score test, constrained estimates and their covariance matrix. Some ingredients may be easier to obtain than others.

Let $T = \{T_n\}$ and $U = \{U_n\}$ be two sequences of tests for the same problem based on sample sizes n. The limiting ratio of the sample sizes n_U and n_T required to achieve the same limiting power $\gamma_n(T, F)$ evaluated at the same sequence of alternatives, when the significance levels of the two test sequences also have the same limit, is the Pitman ARE of T relative to U,

$$e_P(T, U) = \lim_n \frac{n_U}{n_T}.$$

If, for example, $e_P(T,U) = 1/2$, this means that the sequences U_n requires approximately half as many observations as the sequence T_n to achieve the same asymptotic results.

4 Pitman Efficiency

Suppose that the distribution F under consideration may be indexed by a set $\Theta \subset R$, and consider a simple null hypothesis

$$H_0 : \theta = \theta_0$$

to be tested against alternatives

$$\theta > \theta_0.$$

Consider the comparison of test sequences $T = \{T_n\}$ satisfying the following conditions, relative to a neighborhood $\theta_0 \leq \theta \leq \theta_0 + \delta$ of the null hypothesis.

Pitman Conditions

- (P1) For some continuous strictly increasing distribution function G, and functions $\mu_n(\theta)$ and $\sigma_n(\theta)$, the F_θ-distribution of $(T_n - \mu_n(\theta))/\sigma_n(\theta)$ converges to G uniformly in $[\theta_0, \theta_0 + \delta]$:

$$\sup_{\theta_0 \leq \theta \leq \theta_0 + \delta} \sup_{-\infty < t < \infty} \left| P\left(\frac{T_n - \mu_n(\theta)}{\sigma_n(\theta)} \right) \right| \to 0, \qquad n \to \infty.$$

- (P2) For $\theta \in [\theta_0, \theta_0 + \delta]$, $\mu_n(\theta)$ is k times differentiable, with $\mu_n^{(1)}(\theta_0) = \cdots = \mu_n^{(k-1)}(\theta_0) = 0 < \mu_n^{(k)}(\theta_0)$.

- (P3) For some function $d(n) \to \infty$ and some constant $c > 0$,

$$\sigma_n(\theta_0) \sim c \frac{\mu_n^{(k)}(\theta_0)}{d(n)}, \qquad n \to \infty.$$

- (P4) For $\theta_n = \theta_0 + O([d(n)]^{-1/k})$,

$$\mu_n^{(k)}(\theta_n) \sim c \mu_n^{(k)}(\theta_0), \qquad n \to \infty.$$

- (P5) For $\theta_n = \theta_0 + O([d(n)]^{-1/k})$,

$$\sigma_n(\theta_n) \sim c \sigma_n(\theta_0), \qquad n \to \infty.$$

Theorem 1 *(Pitman-Noether). (i) Let $T = \{T_n\}$ satisfy (P1)-(P5). Consider testing H_0 by critical regions $\{T_n > u_{\alpha_n}\}$ with*

$$\alpha_n = P_{\theta_0}(T_n > u_{\alpha_n}) \to \alpha,$$

where $0 < \alpha < 1$. For $0 < \beta < 1 - \alpha$, and $\theta_n = \theta_0 + O([d(n)]^{-1/k})$, we have

$$\beta_n(\theta_n) = P_{\theta_n}(T_n \leq u_{\alpha_n}) \to \beta$$

if and only if

$$\frac{(\theta_n - \theta_0)^k}{k!} \frac{d(n)}{c} \to G^{-1}(1-\alpha) - G^{-1}(\beta). \tag{20}$$

(ii) Let $T_A = \{T_{An}\}$ and $T_B = \{T_{Bn}\}$ satisfy (P1)-(P5) with common G, k and $d(n)$ in (P1)-(P3). Let $d(n) = n^q$, $q > 0$. Then the Pitman ARE of T_A relative to T_B is given by

$$e_P(T_A, T_B) = \left(\frac{c_B}{c_A}\right)^{1/q}.$$

Proof. Check that, by (P1),

$$\left|\beta_n(\theta_n) - G\left(\frac{u_{\alpha n} - \mu_n(\theta_n)}{\sigma_n(\theta_n)}\right)\right| \to 0, \quad n \to \infty.$$

Then $\beta_n(\theta_n) \to \beta$ if and only if

$$\frac{u_{\alpha n} - \mu_n(\theta_n)}{\sigma_n(\theta_n)} \to G^{-1}(\beta). \tag{21}$$

Likewise (check), $\alpha_n \to \alpha$ if and only if

$$\frac{u_{\alpha n} - \mu_n(\theta_0)}{\sigma_n(\theta_0)} \to G^{-1}(1-\alpha). \tag{22}$$

It follows (check, utilizing (P5)) that (21) and (22) together are equivalent to (22) and

$$\frac{\mu_n(\theta_n) - \mu_n(\theta_0)}{\sigma_n(\theta_0)} \to G^{-1}(1-\alpha) - G^{-1}(\beta) \tag{23}$$

together. By (P2) and (P3),

$$\frac{\mu_n(\theta_n) - \mu_n(\theta_0)}{\sigma_n(\theta_0)} \sim \frac{\mu_n^{(k)}(\tilde{\theta}_n)}{\mu_n^{(k)}(\theta_0)} \cdot \frac{(\theta_n - \theta_0)^k}{k!} \cdot \frac{d(n)}{c_A}, \quad n \to \infty,$$

where $\theta_0 \le \tilde{\theta}_n \le \theta_n$. Thus, by (P4), (23) is equivalent to (20). This completes the proof of (i).

Now consider tests based on T_A and T_B, having sizes $\alpha_{An} \to \alpha$ and $\alpha_{Bn} \to \alpha$. Let $0 < \beta < 1 - \alpha$. Let $\{\theta_n\}$ be a sequence of alternatives of the form

$$\theta_n = \theta_0 + A[d(n)]^{-1/k}.$$

It follows by (i) that if $h(n)$ is the sample size at which T_B peforms "equivalently" to T_A with sample size n, that is, at which T_B and T_A have the same limiting power $1 - \beta$ for the given sequence of alternatives, so that

$$\beta_{An}(\theta_n) \to \beta, \quad \beta_{Bh(n)}(\theta_n) \to \beta,$$

then we must have $d(h(n))$ proportional to $d(n)$ and

$$\frac{(\theta_n - \theta_0)^k}{k!}\frac{d(n)}{c_A} \sim \frac{(\theta_n - \theta_0)^k}{k!}\frac{d(h(n))}{c_B},$$

or

$$\frac{d(h(n))}{d(n)} \to \frac{c_B}{c_A}.$$

For $d(n) = n^q$, this yields $(h(n)/n)^q \to (c_B/c_A)$, proving (ii).

4.1 Rank Tests for Comparing Two Treatments

For comparing a new treament or procedure with the standard method, N subjects (patients, students, etc.) are divided at random into a group of n who will receive a new treatment and a control group of m who will be treated by the standard method. At the termination of the study, the subjects are ranked either directly or according to some response that measures the success of the treatment such as a test score in an educational or pyschological investigation. The hypothesis H_0 of no treatment effect is rejected, and the superiority of the new treatment acknowledged, if the ranking the n treated subjects rank sufficiently high. (Here it is assumed that the success of the treatment is indicated by an increased response; if instead the aim is to decrease the response, H_0 is rejected when the n treated subjects rank sufficiently low.)

Let the ranks of the treated subjects be denoted by $S_1, \ldots, S_n$, where we shall assume that they are numbered in increasing order. Denote the sum of the treatment ranks $W_S = S_1 + \cdots + S_n$. The hypothesis H_0 is then rejected and the treatment judged to be effective when W_S is sufficiently large, say, when $W_S \geq c$. Here the constant c is determined by the equation

$$P_{H_0}(W_S \geq c) = \alpha.$$

The test defined above is known as the *Wilcoxon rank-sum test.*

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent, the X's identically distributed with distribution F and the Y's identically distributed with distribution G. Here the Y's are responses to a treatment. Then $H_0 : F = G$ and $H_a : Y$ is stochastically larger than X, i.e., $G(t) \leq F(t)$ for all t but $G \neq F$.

Let the ranks of the X's be denoted by $R_1, \ldots, R_m$. If we substitute R's for X's and S's for Y's in the two-sample t-test statistic, we obtain

$$\left(\frac{nm}{N}\right)^{1/2} \frac{\frac{1}{n}\sum_{i=1}^{n} S_i - \frac{1}{m}\sum_{j=1}^{m} R_j}{(N-2)^{-1}\left[\{\sum_{i=1}^{n}(S_i - \frac{N+1}{2})^2 + \sum_{j=1}^{m}(R_j - \frac{N+1}{2})^2\right]\}^{1/2}}.$$

This statistic is equivalent to the Wilcoxon statistic W_S, the sum of the ranks of the treatment group. Write W_{XY} as the number of pairs (X_i, Y_j) with $X_i < Y_j$. It can be shown that

$$W_S - \frac{1}{2}n(n+1) = W_{XY}.$$

W_{XY} is usually known as the Mann-Whitney statistic. Let $\phi(X_i, Y_j) = 1$ if $X_i < Y_j$, and 0 otherwise. Then

$$W_{XY} = \sum_{i=1}^{m}\sum_{j=1}^{n} \phi(X_i, Y_j) \tag{24}$$

Then we shall prove that W_{XY} is asymptotically normal as m and n tend to infinity. The method of proof consists in replacing the variable W_{XY} by a sum of independent random variables, which is asymptotically equivalent to W_{XY} and to which the central limit theorem can then be applied. It is natural for this purpose to try a sum of the form

$$S = \sum_{i=1}^{m} a_i(X_i) + \sum_{j=1}^{n} b_j(Y_j) \tag{25}$$

but how should one choose the functions a_i and b_j? The following "projection mathod" introduced in a different context by Hajek (1961), produces the a_i and b_j most likely to succeed in the sense of minimizing $E(W_{XY} - S)^2$. This approach is due to Hoeffding

(1948), and is applicable to a large class of statistics, the so-called U-statistics. Note that

$$\theta(F,G) = \int F dG = P(X \leq Y).$$

An unbiased estimator of $\theta(F,G)$ is

$$U = \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n} I(X_i \leq Y_j),$$

which is the W_{XY}. A statistic can be written in the form is called a U-statistics. Note that the popularity of this projection method is due to Hajek (1968), who gives the following result.

Lemma 2 *(Hoeffding) Let $Z_1, \ldots, Z_n$ be independent random variables and $S = S(Z_1, \ldots, Z_n)$ any statistic satisfying $E(S^2) < \infty$. Then the random variable*

$$S^* = \sum_{i=1}^{n} E(S|Z_i) - (n-1)E(S)$$

satisfies $E(S^*) = E(S)$ and

$$E(S - S^*)^2 = Var(S) - Var(S^*).$$

The random variables S^* is called the *projection* of S on $Z_1, \ldots, Z_n$. Note that it is conveniently a sum of *independent and identically distributed* random variables. In cases that $E(S - S^*)^2 \to 0$ at a suitable rate as $n \to \infty$, the asymptotic normality of S may be established by applying classical theory to S^*.

Proof of Hoeffding's Lemma. Without loss of generality, we can assume that $E(S) = 0$. Consider the problem of finding the sum

$$T = \sum_{i=1}^{n} k_i(Z_i) \tag{26}$$

for which $E(S - T)^2$ is as small as possible; the minimizing T may be considered the "projection" of S onto the linear space formed by the functions T. Let

$$r_i(z_i) = E(S|Z_i = z_i) \tag{27}$$

be the conditional expectation of S given $Z_i = z_i$, and let

$$S^* = \sum_{i=1}^{n} r_i(Z_i). \tag{28}$$

That S^* is the desired minimizing function is an immediate consequence of the following identity, which holds for all statistics T and S with mean zero and satisfying (26) for which the required expectation exist:

$$E(S - T)^2 = E(S - S^*)^2 + E(S^* - T)^2. \tag{29}$$

To prove this identity, write

$$E(S - T)^2 = E[(S - S^*) + (S^* - T)]^2.$$

Squaring the right-hand side proves (29) if it can be shown that

$$E[(S - S^*)(S^* - T)] = 0. \tag{30}$$

Since the left-hand side of (30) is the sum of the expectations of

$$[r_i(Z_i) - k_i(Z_i)](S - S^*) \tag{31}$$

it is enough to show that the expectation of (31) given Z_i is zero for all i. We shall prove this by showing that the conditional expectation of (31) given Z_i is zero. In the conditional expectation of this product, the first factor can be taken out of the expectation sign since it depends only on Z_i, so that it is finally only necessary to show that the conditional expectation of $S - S^*$ given Z_i is zero. Now

$$E[(S - S^*)|Z_i] = E\{S - r_i(Z_i) - \sum_{j \neq i} r_j(Z_j)|Z_i\}.$$

From the definition of $r_i(Z_i)$, it is seen that the conditional expectation of $S - r_i(Z_i)$ given Z_i is zero. On the other hand, since Z_i and Z_j are independent, the conditional expectation of $r_j(Z_j)$ given Z_i is equal to the unconditional expectation of $r_j(Z_j)$, which by the definition of r_j is equal to $E(S)$ and hence equal to zero. This completes the proof of (30) and therefore of (29).

A useful special case of (29) is obtained by putting $T = 0$, which gives after arrangement

$$E(S - S^*)^2 = E(S^2) - E(S^{*2}) = Var(S) - Var(S^*). \tag{32}$$

Before we apply Hoeffeding lemma to the W_{XY}-statistic (24), we will calculate the expectation and variance of W_{XY}. Set $\theta = (F, G)$,

$$E_\theta[\phi(X, Y)] = P_\theta[X < Y]$$

and we obtain

$$E_\theta(W_{XY}) = mnp \tag{33}$$

where $p = P_\theta[X < Y]$. Similarly, we have

$$Var_\theta(W_{XY}) = nmp(1 - p) + nm(n - 1)(q_1 - p^2) + nm(m - 1)(q_2 - p^2) \tag{34}$$

where $q_1 = P_\theta[X_1 < \min(Y_1, Y_2)]$ and $q_2 = P_\theta[Y_1 > \max(X_1, X_2)]$.

Note that under H_0, if F is continuous, $p = 1/2$ while $q_1 = q_2 = 1/3$, since, among three independent identically distributed variables, each one is equally likely to be the minimum or the maximum. We then have $E_\theta(W_{XY}) = mn/2$ and $Var_\theta(W_{XY}) = mn(N + 1)/12$ under H_0.

Put

$$\psi(x, y) = \phi(x, y) - p. \tag{35}$$

Note that

$$E[\psi(X_\alpha, Y_\beta)|X_i = x] = \begin{cases} E\psi(x, Y_\beta) & \text{if } \alpha = i \\ 0 & \text{if } \alpha \neq i \end{cases}$$

and

$$E[\psi(X_\alpha, Y_\beta)|Y_j = y] = \begin{cases} E\psi(X_\alpha, y) & \text{if } \beta = j \\ 0 & \text{if } \beta \neq j \end{cases}$$

Put $\psi_{10}(x) = E_Y\psi(x,Y)$ and $\psi_{01}(y) = E_X\psi(X,y)$. The projection of $W_{XY} - mnp$ by Hoeffeding Lemma is $n\sum_{i=1}^m \psi_{10}(X_i) + m\sum_{j=1}^n \psi_{01}(Y_j)$. Consider

$$U = \sqrt{m}\left[\frac{1}{m}\sum_{i=1}^m \psi_{10}(X_i) + \frac{1}{n}\sum_{j=1}^n \psi_{01}(Y_j)\right]$$

and $S = \sqrt{m}[(mn)^{-1}W_{XY} - p]$. Note that

$$\begin{aligned} Var(S) &\rightarrow q_1 - p^2 + \frac{m}{n}(q_2 - p^2), \\ Var(U) &= Var(\psi_{10}(X)) + \frac{m}{n}Var(\psi_{01}(Y)), \\ E(S-U)^2 &= Var(S) - Var(U). \end{aligned}$$

Observe that for $j \neq k$, $Var(\psi_{10}(X)) = q_1 - p^2$ and $Var(\psi_{01}(Y) = q_2 - p^2$.
(i.e. $E\psi(x_1, Y_j)\psi(x_1, Y_k) = [\psi_{10}(x_1)]^2$ and

$$E_X[\psi_{10}(X)]^2 = E\psi(X,Y_j)\psi(X,Y_k) = Cov(\psi(X,Y_j), \psi(X,Y_k)).$$

We then conclude that $E(S-U)^2 \rightarrow 0$.

Theorem 2 *Suppose that F and G are continuous and that $0 < P_\theta[X < Y] < 1$. Then*

$$\frac{S - E_\theta(S)}{\sqrt{Var_\theta(S)}} \xrightarrow{d} N(0,1) \;\; as \; \min(n,m) \rightarrow \infty.$$

Remark. Reject H_0 when

$$\frac{W_{XY} - \frac{1}{2}nm}{\sqrt{\frac{1}{12}nm(N+1)}} \geq z(1-\alpha).$$

4.2 Pitman efficiency of the Wilcoxon rank-sum test to the two-sample t-test

We turn now to the comparison of the performance of the Wilcoxon and two-sample t tests. At first sight it would appear that a good reason for using the Wilcoxon is that it has a guaranteed probability of type I error and a good reason against using the Wilcoxon is its inefficient use of the data. We assume that the X's and Y's have the same variance σ^2 and means μ_1 and μ_2. Although the t test does not have a guaranteed probability of type I error, if n and m are moderately large, H_0 is true, and F has a finite second moment, then the probability of type I error of the t test is fairly close to that specified by the normal model.

Recall that the two-sample t statistic is given by

$$T = \sqrt{\frac{nm}{N}}\frac{\bar{Y} - \bar{X}}{s_2} \tag{36}$$

where

$$s_2 = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2}{N-2}. \tag{37}$$

We start by obtaining an approximation to the critical value and power of the t test. Note that $s_2^2 \xrightarrow{P} \sigma^2$ as $\min(n,m) \to \infty$. It follows from Slutsky's theorem and central limit theorem that when $\mu_1 = \mu_2$, T converges in law to a $N(0,1)$ random variable as $\min(n,m) \to \infty$. Then the t test that rejects H_0 when $T \geq t_{N-2}(1-\alpha)$ has approximately level α regardless of the shape of F and G and $z(1-\alpha)$ is an approximate critical value as we claimed above.

If $\mu_1 \neq \mu_2$, let $\delta = (\mu_2 - \mu_1)/\sigma$. Then, arguing as above, if $\sqrt{nm/N}\delta$ stays bounded $T - \sqrt{nm/N}\delta$ has approximately a $N(0,1)$ random distribution for all F and G with $\sigma^2 < \infty$. We then can approximate the probability $P_\theta(T \geq t_{N-2}(1-\alpha))$ by

$$\beta_T = P_\theta[T \geq z(1-\alpha)] = 1 - \Phi(z(1-\alpha) - \sqrt{nm/N}\delta) = \Phi(z(\alpha) + \sqrt{nm/N}\delta).$$

For Wilcoxon test,

$$\begin{aligned}
\beta_N &= P_\theta\left[W_{XY} \geq \frac{1}{2}nm + z(1-\alpha)\sqrt{\frac{1}{12}nm(N+1)}\right] \\
&= P_\theta\left[\frac{W_{XY} - E_\theta(W_{XY})}{\sqrt{var_\theta(W_{XY})}} \geq \frac{nm(\frac{1}{2}-p) + z(1-\alpha)\sqrt{\frac{1}{12}nm(N+1)}}{\sqrt{var_\theta(W_{XY})}}\right] \\
&\approx \Phi\left(\frac{nm(\frac{1}{2}-p) + z(1-\alpha)\sqrt{\frac{1}{12}nm(N+1)}}{\sqrt{var_\theta(W_{XY})}}\right).
\end{aligned}$$

Consider the case that $X \sim N(\mu_1, \sigma^2)$, $Y \sim N(\mu_2, \sigma^2)$, $n = m$ and $\alpha = 0.05$. Note that $\delta = (\mu_2 - \mu_1)/\sigma = 0.5$. Suppose we want to have $\beta = 0.9$.
For t-test, solve

$$-1.645 + \sqrt{\frac{\frac{N}{2}\frac{N}{2}}{N}}0.5 = 1.282$$

and get $N = 16 \cdot (2.927)^2 \approx 140$.
For Wilcoxon test:

$$p = P_\theta(X < Y) = \Phi\left(\frac{\mu_2 - \mu_1}{\sqrt{2}}\sigma\right), \qquad q_1 = P\left(Z_1 < \frac{\Delta}{\sqrt{2}}, Z_2 < \frac{\Delta}{\sqrt{2}}\right),$$
$$q_2 = P\left(Z_1 < \frac{\Delta}{\sqrt{2}}, Z_3 < \frac{\Delta}{\sqrt{2}}\right),$$

where $Z_1 = [X_1 - Y_1 - (\mu_1 - \mu_2)]/\sqrt{2}\sigma$, $Z_2 = [X_1 - Y_2 - (\mu_1 - \mu_2)]/\sqrt{2}\sigma$, $Z_3 = [X_2 - Y_1 - (\mu_1 - \mu_2)]/\sqrt{2}\sigma$. Note that $(Z_1, Z_2) \sim N(0,0,1,1,1/2)$, $(Z_1, Z_3) \sim N(0,0,1,1,1/2)$. When $\Delta = 0.5$, $p = 0.638$, $q_1 = q_2 = 0.483$, we have $\beta_W \approx \Phi(-1.729 + 0.355\sqrt{N/2}) = 0.9$. Hence, $N \approx 144$.

References

[1] Cramer, J.S. (1986). *Econometric applications of Maximum Likelihood methods.* Cambridge University Press.

[2] Hajek, J. (1968). Asymptotically normality of simple linear rank statistics under alternatives. *Ann. Math. Statist.* **39** 325-346.

[3] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** 293-325.

[4] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, San Francisco.

[5] Neyman. J. and Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* **20A** 175-240 and 263-294.

[6] Neyman. J. and Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc.* **A 236** 333-380.

[7] Pearson, K. (1900). On the criterion that a given system of derivations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag. Ser.* (5) **50** 157-172.

[8] Pitman, E.J.G. (1949). *Lecture Notes on Nonparametric Statistical Inference.* Columbia University.

[9] Rao, C.R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. Proceedings of the Cambridge Philosophical Society **44** 50-57.

[10] Rao, C.R. (1973). *Linear Statistical Inference and its Applications.* 2nd ed., Wiley, New York.

[11] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York.

[12] Simar, L. (1983). Protecting against gross errors: the aid of Bayesian methods. In *Specifying Statistical Models: From Parametric to Non-Parametric Using Bayesian or Non-Bayesian Approaches.* Edited by J.R. Florens, M. Mouchart, J.P. Raoult, L. Simar, and A.F.M. Smith. Lecture Notes in Statistics: Vol. 16. Springer-Verlag, New York.

[13] Wald, A. (1943). Test of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society.* **54** 595-601.