24 Classical Nonparametrics

The development of nonparametrics originated from a concern about the approximate validity of parametric procedures based on a specific narrow model when the model is questionable. Procedures which are reasonably insensitive to the exact assumptions that one makes are called robust. Such assumptions may be about a variety of things. They may be about an underlying common density assuming that the data are iid; they may be about the dependence structure of the data itself; in regression problems, they may be about the form of the regression function, etc. For example, if we assume that our data are iid from a certain $N(\theta, 1)$ density, then we have a specific parametric model for our data. Statistical models are always, at best, an approximation. We do not believe that the normal model is *the correct* model. So, if we were to use a procedure that had excellent performance under the normal model, but fell apart at models similar to normal, but different from the normal in aspects that a statistician would find hard to pin down, then the procedure would be considered risky. For example, tails of underlying densities are usually hard to pin down. Nonparametric procedures provide a certain amount of robustness to departure from a narrow parametric model, at the cost of a suboptimal performance at the parametric model. It is important to understand, though, that what we commonly call nonparametric procedures do not provide robustness with regard to all characteristics. For instance, a nonparametric test may retain the type I error rate approximately under various kinds of models, but may not retain good power properties under different kinds of models. The implicit robustness is limited, and it always comes at the cost of some loss of efficiency at fixed parametric models. There is a trade-off.

As a simple example, consider the t-test for the mean μ of a normal distribution. If normality holds, then under the null hypothesis, $H_0 : \mu = \mu_0$,

$$P_{\mu_0}\left(\frac{\sqrt{n}(\overline{X} - \mu_0)}{s} > t_{\alpha,n-1}\right) = \alpha,$$

for all n, μ_0, and σ. However, if the population is not normal, neither the size nor the power of the t-test remains the same as under the normal case. If these hange substantially, we have a robustness problem. However, as we will later see, by making a minimal number of assumptions (specifically, no parametric assumptions) we can develop procedures with some sort of a safety net. Such methods would qualify for being called nonparametric methods.

There are a number of texts that discuss classical nonparametric estimators and tests in various problems. We recommend Hajek and Sidak (1967), Hettmansperger (1984), Randles and Wolfe (1979), and Lehmann and Romano (2005), in particular. A recent article in the nature of a review of asymptotic theory of common nonparametric tests is Jurevckova(1995). Other specific references are given in the sections.

24.1 Some Early Illustrative Examples

We start with three examples to explain the ideas of failure of narrowly focused parametric procedures at broader nonparametric models, and of the possibility of other procedures which have some limited validity independent of specific parametric models.

Example 24.1. Let F be a cdf on $\mathcal{R}$. For $0 < p < 1$, let ξ_p denote the pth percentile of the distribution F. That is,

$$\xi_p = \inf\{x : F(x) \geq p\}.$$

Let F_n be the empirical cdf given by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \leq x\}}.$$

Suppose we estimate the percentile ξ_p by inverting the empirical cdf. That is,

$$\hat{\xi}_p = F_n^{-1}(p) = \inf\{x : F_n(x) \geq p\}.$$

Then, it can be shown that under minimal assumptions, the estimator $\hat{\xi}_p$, a distribution-free estimate of the corresponding population quantile, is strongly consistent for ξ_p; see Hettmansperger (1984), e.g. Thus, $F_n^{-1}(p)$ gives us at least consistency without requiring any rigid parametric assumptions. It would qualify for being called a nonparametric procedure.

Example 24.2. Consider the t confidence interval, $\overline{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$, denoted by C_n. If $X_1, \ldots, X_n$ are iid observations from $N(\mu, \sigma^2)$ then $P_{\mu,\sigma}(C_n \ni \mu) \equiv 1 - \alpha$ for all n, μ, σ. That is, the coverage probability is exact. But what happens if $X_1, \ldots, X_n$ are iid observations from a general distribution F? More precisely, what can be

asserted about the coverage probability, $P_F(C_n \ni \mu(F))$? If we can assume that $\mathbb{E}_F X^2 < \infty$ then it can be shown that

$$\lim_{n\to\infty} P_F(C_n \ni \mu(F)) = 1 - \alpha.$$

That is, for fixed F and $\epsilon > 0$, there exists a number $N = N(F, \alpha, \epsilon) > 0$ such that

$$n > N \Longrightarrow |P_F(C_n \ni \mu(F)) - (1 - \alpha)| < \epsilon.$$

However, the asymptotic validity is not uniform in F. That is, if $\mathfrak{F}$ denotes the set of all cdf's with finite second moment, then

$$\lim_{n\to\infty} \inf_{F\in\mathfrak{F}} P_F(C_n \ni \mu) = 0.$$

This is an example of the failure of a parametric procedure under completely nonparametric models.

Example 24.3. This is a classical example of a nonparametric confidence interval for a quantile. Let θ denote the median of a distribution F. Suppose $X_{(1)} < \ldots < X_{(n)}$ are the order statistics of an iid sample from a continuous cdf F. For a fixed k, $0 \le k \le \frac{n-1}{2}$, using the notation $\{U_i\}$ to denote an iid $U[0,1]$ sample,

$$\begin{aligned} P_F(X_{(k+1)} \le \theta \le X_{(n-k)}) &= P_F(F(X_{(k+1)} \le F(\theta) \le F(X_{(n-k)}) \\ &= P(U_{(k+1)} \le 0.5 \le U_{(n-k)}) \\ &= P(k+1 \le S_n < n-k) \\ &= P(k+1 \le \text{Bin}(n, 0.5) < n-k) \end{aligned}$$

where $S_n = \#\{i : U_i \le 0.5\}$. We can choose k such that this probability is $\ge 1 - \alpha$. This translates to a nonparametric confidence interval for θ. Notice that the only assumption we have used here is that F is continuous (this assumption is needed to perform the quantile transformation).

But, the nonparametric interval would not perform as well as a t-interval if F is a normal cdf.

24.2 Sign Test

This is perhaps the earliest example of a nonparametric testing procedure. In fact, the test was apparently discussed by Laplace in the 1700's. The Sign test is a test for the median of any continuous distribution, without requiring any other assumptions.

Let $X_1, \ldots, X_n$ be iid samples from an (absolutely) continuous distribution F. Let $\theta = \theta(F)$ be the median of the distribution. Consider testing $H_0 : \theta = \theta_0$ versus the one-sided alternative $H_1 : \theta > \theta_0$. Define the statistic

$$S_n = \sum_{i=1}^{n} I_{\{X_i > \theta_0\}} \tag{24.1}$$

Then large values of S_n would indicate that H_1 is true, and so the Sign test rejects the null when $S_n > k = k(n, \alpha)$, where this k is chosen so that $P_{H_0}(S_n > k) \leq \alpha$. Under H_0, S_n has a Bin(n,1/2) distribution and so $k = k(n, \alpha)$ is just a quantile from the appropriate Binomial distribution. Thus, the Sign test is a size-α test for the median θ for any sample size n and any continuous cdf.

The next question is how does the Sign test perform relative to a competitor; e.g. the t-test. Of course, to make a comparison with the t test, we must have F such that the mean exists and equals the median. A good ground for comparison is when $F = N(\theta, \sigma^2)$.

Suppose that $X_1, \ldots, X_n$ are iid observations from $N(\theta, \sigma^2)$ for some unknown θ and σ. We wish to test $H_0 : \theta = \theta_0$ against a one- or two-sided alternative. Each of the tests reject H_0 if $T_n \geq c_n$, where T_n is the appropriate test statistic. The two power functions for the case of one sided alternatives are, respectively,

$$P_{\theta,\sigma}\left(\frac{\sqrt{n}(\overline{X} - \theta_0)}{s} > t\right) \quad \text{and} \quad P_{\theta,\sigma}\left(S_n > k(n, \alpha, \theta_0)\right).$$

The former probability is a non-central t probability and the latter is a Binomial probability. We wish to compare the two power functions.

The point is that, at a fixed alternative θ, if α remains fixed, then for large n, the power of both tests is approximately 1 and there would be no way to practically compare the two tests. Perhaps we can see how the powers compare for $\theta \approx \theta_0$. The idea is to take $\theta = \theta_n \to \theta_0$ at such a rate that the limiting power of the tests is strictly between α and 1. If the two powers converge to different values then we can take the ratio of the limits as a measure of efficiency. We have discussed this concept of efficiency, namely the Pitman efficiency, in detail in Chapter 22.

Example 24.4. Let $X_1, \ldots, X_n$ be iid observation from $N(\theta, \sigma^2)$. Suppose we wish to test $H_0 : \theta = \theta_0$. Let T denote the t-test and S denote the Sign test. Then $e_P(S, T) = \frac{2}{\pi} \approx 0.637 < 1$. That is, the precision that the t-test achieves with 637 observations is achieved by the Sign test with 1000 observations. This reinforces our earlier comment that while nonparametric procedures enjoy a certain amount of

validity at broad models, they cannot compete with parametric optimal procedures at specified parametric models.

The Sign test, however, cannot get arbitrarily bad with respect to the t test, under some restrictions on the cdf F, as is shown by the following result, although the t test can be arbitrarily bad with respect to the Sign test.

Theorem 24.1. (Hodges-Lehmann, 1956) Let $X_1, \ldots, X_n$ be iid observations from any distribution with density $f(x-\theta)$ where $f(0) > 0$, f is continuous at 0 and $\int z^2 f(z)\, dz < \infty$. Then $e_P(S,T) \geq \frac{1}{3}$ and $e_P(S,T) = \frac{1}{3}$ when f is any symmetric uniform density.

Remark: We learn from this result that the Sign test has an asymptotic efficiency with respect to the t test that is bounded away from zero for a fairly large class of location parameter cdfs, but that the minimum efficiency is only $\frac{1}{3}$, which is not very good. We will later discuss alternative nonparametric tests for the location parameter problem, which have much better asymptotic efficiencies.

24.3 Consistency of the Sign Test

Definition 24.1. Let $\{\varphi_n\}$ be a sequence of tests for $H_0 : F \in \Omega_0$ versus $H_1 : F \in \Omega_1$. Then $\{\varphi_n\}$ is consistent against the alternatives Ω_1 if

(i) $\mathbb{E}_F(\varphi_n) \to \alpha \in (0,1)\ \forall\ F \in \Omega_0$,

(ii) $\mathbb{E}_F(\varphi_n) \to 1\ \forall\ F \in \Omega_1$.

As in estimation, consistency is a rather weak property of a sequence of tests. However, something must be fundamentally wrong with the test for it not to be consistent. If a test is inconsistent against a large class of alternatives, then it is considered an undesirable test.

Example 24.5. For a parametric example, let $X_1, \ldots, X_n$ be an iid sample from the Cauchy distribution, $C(\theta, 1)$. For all $n \geq 1$, we know that $\overline{X}$ also has the $C(\theta, 1)$ distribution. Consider testing the hypothesis $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ by using a test which rejects for large $\overline{X}$. The cutoff point, k, is found by making $P_{\theta=0}(\overline{X} > k) = \alpha$. But k is simply the αth quantile of the $C(0,1)$ distribution. Then the power of this test is given by

$$P_\theta(\overline{X} > k) = P(C(\theta,1) > k) = P(\theta + C(0,1) > k) = P(C(0,1) > k - \theta).$$

This is a fixed number, not depending on n. Therefore, the power $\not\to 1$ as $n \to \infty$, and so the test is not consistent even against parametric alternatives.

Remark: A test based on the median would be consistent in the $C(\theta, 1)$ case.

The following theorem gives a sufficient condition for a sequence of tests to be consistent.

Theorem 24.2. Consider a testing problem $H_0 : F \in \Omega_0$ vs. $H_1 : F \in \Omega_1$. Let $\{V_n\}$ be a sequence of test statistics and $\{k_n\}$ a sequence of numbers such that

$$P_F(V_n \geq k_n) \to \alpha < 1 \ \forall F \in \Omega_0.$$

For a test which rejects H_0 when $V_n \geq k_n$, suppose:

- Under any $F \in \Omega_0 \cup \Omega_1$, $V_n \stackrel{P}{\Rightarrow} \mu(F)$, some suitable functional of F
- For all $F \in \Omega_0$, $\mu(F) = \mu_0$ and for all $F \in \Omega_1$, $\mu(F) > \mu_0$
- Under H_0, $\frac{\sqrt{n}(V_n - \mu_0)}{\sigma_0} \stackrel{\mathcal{L}}{\Rightarrow} N(0,1)$ for some $0 < \sigma_0 < \infty$.

Then the sequence of tests is consistent against $H_1 : F \in \Omega_1$.

Proof. We can take $k_n = \frac{\sigma_0 z_\alpha}{\sqrt{n}} + \mu_0$, where z_α is a standard normal quantile. With this choice of $\{k_n\}$,

$$P_F(V_n \geq k_n) \to \alpha \ \ \forall F \in \Omega_0.$$

The power of the test is

$$Q_n = P_F(V_n \geq k_n) = P_F(V_n - \mu(F) \geq k_n - \mu(F)).$$

Since we assume $\mu(F) > \mu_0$, it follows that $k_n - \mu(F) < 0$ for all large n and for all $F \in \Omega_1$. Also, $V_n - \mu(F)$ converges in probability to 0 under any F, and so $Q_n \to 1$. Since the power goes to 1, the test is consistent against any alternative F in Ω_1. $\square$

Corollary 24.1. If F is an absolutely continuous cdf with unique median $\theta = \theta(F)$, then the Sign test is consistent for tests on θ.

Proof. Recall that the Sign test rejects $H_0 : \theta(F) = \theta_0$ in favor of $H_1 : \theta(F) > \theta_0$ if $S_n = \sum I_{\{X_i > \theta_0\}} \geq k_n$. If we choose $k_n = \frac{n}{2} + z_\alpha \sqrt{\frac{n}{4}}$ then, by the ordinary Central Limit Theorem, we have

$$P_{H_0}(S_n \geq k_n) \to \alpha.$$

Then the consistency of the Sign test follows from the above theorem by letting

(i) $k_n = \frac{1}{2} + z_\alpha \sqrt{\frac{1}{4n}}$

(ii) $V_n = \frac{S_n}{n}$

(iii) $\mu_0 = \frac{1}{2}$, $\sigma_0 = \frac{1}{2}$

(iv) $\mu(F) = 1 - F(\theta_0) > \frac{1}{2}$ for all F in the alternative.

□

24.4 Wilcoxon Signed-Rank Test

Recall that Hodges and Lehmann proved that the Sign test has a small positive lower bound of $\frac{1}{3}$ on the Pitman efficiency with respect to the t-test in the class of densities with a finite variance, which is not satisfactory (see Theorem 24.1). The problem with the Sign test is that it only considers whether an observation is $> \theta_0$ or $\leq \theta_0$, but not the magnitude. A nonparametric test which incorporates the magnitudes as well as the signs is called the Wilcoxon Signed-Rank Test; see Wilcoxon (1945).

Definition 24.2. Given a generic set of n numbers $z_1, \ldots, z_n$, the rank of a particular z_i is defined as

$$R_i = \#\{k : z_k \leq z_i\}.$$

Suppose that $X_1, \ldots, X_n$ are the observed data from some location parameter distribution $F(x - \theta)$ and assume that F is symmetric. Let $\theta = \text{Med}(F)$. We want to test $H_0 : \theta = 0$ against $H_1 : \theta > 0$. We start by ranking $|X_i|$ from the smallest to the largest, giving the units ranks $R_1, \ldots, R_n$. Then the Wilcoxon Signed-Rank statistic is defined to be the sum of these ranks which correspond to originally positive observations. That is,

$$T = \sum_{i=1}^{n} R_i I_{\{X_i > 0\}} \tag{24.2}$$

If we define $W_i = I_{|X|_{(i)}}$ corresponds to some positive X_j, then we have an alternative expression for T, namely,

$$T = \sum_{i=1}^{n} i W_i \tag{24.3}$$

To do a test, we need the null distribution of T. It turns out that, under H_0, the $\{W_i\}$ have a relatively simple joint distribution.

Theorem 24.3. Under H_0, $W_1, \ldots, W_n$ are IID Bernoulli $\frac{1}{2}$ variables.

This, together with the representation of T above and Lyapunov's CLT (which we recall below) leads to the asymptotic null distribution of T. See Hettmansperger (1984) for the formal details in the proofs.

Theorem 24.4. (Lyapunov's CLT) For $n \geq 1$, let $X_{n1}, \ldots, X_{nn}$ be a sequence of independent random variables such that $\mathbb{E}X_{ni} = 0 \; \forall i$, $\mathrm{Var}\left(\sum_i X_{ni}\right) = 1$ and $\mathbb{E}|X_{ni}|^3 \to 0$. Then

$$\sum_{i=1}^{n} X_{ni} \stackrel{\mathcal{L}}{\Rightarrow} N(0,1).$$

Thus, under H_0, the statistic T is a sum of independent, but not iid, random variables. It follows from Lyapunov's Theorem, stated above, that T is asymptotically normal. Clearly

$$E_{H_0}T = \frac{n(n+1)}{4} \quad \text{and} \quad \mathrm{Var}_{H_0}T = \frac{n(n+1)(2n+1)}{24}.$$

The above results imply the following theorem.

Theorem 24.5. Let $X_1, \ldots, X_n$ be iid observations from $F(x-\theta)$, where F is continuous, and $F(x) = 1 - F(-x)$ for all x. Under $H_0 : \theta = 0$,

$$\frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \stackrel{\mathcal{L}}{\Rightarrow} N(0,1).$$

Therefore, the Signed-Rank test can be implemented by rejecting the null hypothesis, $H_0 : \theta = 0$ if

$$T > \frac{n(n+1)}{4} + z_\alpha \sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

The other option would be to find the *exact* finite sample distribution of T under the null. This can be done in principle, but the CLT approximation works pretty well.

We work out the exact distribution of T_n under the null, due to its classic nature. Recall that $T_n = \sum_{i=1}^{n} iW_i$ where W_i are iid Bernoulli $\frac{1}{2}$ random variables. Let $M = \frac{n(n+1)}{2}$. The probability generating function of T_n is

$$\psi_n(t) = E_{H_0}t^{T_n} = \sum_{k=0}^{M} t^k P(T_n = k).$$

If we can find $\psi_n(t)$ and its power series representation then we can find $P(T_n = k)$ by equating the coefficients of t^k from each side. But,

$$\begin{aligned}\sum_{k=0}^{M} t^k P(T_n = k) &= E_{H_0} t^{T_n} = E_{H_0} t^{\sum kW_k} = \prod_k E_{H_0} t^{kW_k} \\ &= \prod_{k=1}^{n} \left(\frac{1}{2} + \frac{1}{2}t^k\right) = \frac{1}{2^n}\prod_{k=1}^{n}(1+t^k) = \frac{1}{2^n}\sum_{k=0}^{M} c_{k,n} t^k\end{aligned}$$

where the sequence $\{c_{k,n}\}$ is determined from the coefficients of t^k in the expansion of the product. From here, we can get the distribution of T_n by setting $P(T_n = k) = c_{k,n}/2^n$. This cannot be done by hand, but is easily done in any software package, unless n is large, such as $n > 30$. Tables of $P_{H_0}(T_n = k)$ are also widely available.

Remark: If $X_i \overset{\text{iid}}{\sim} F(x - \theta)$ where $F(\cdot)$ is symmetric but $\theta \neq 0$ (i.e. under the alternative) then T_n no longer has the representation of the form $T_n = \sum_{j=1}^{n} Z_j$ for independent $\{Z_j\}$. In this case, deriving the asymptotic distribution of T_n is more complicated. We will do this later by using the theory of U-statistics.

Meanwhile, for sampling from a completely arbitrary continuous distribution, say $H(x)$, there are formulas for the mean and variance of T_n; see Hettmansperger (1084) for proofs. These formulas are extremely useful and we provide them next.

Theorem 24.6. Let H be a continuous cdf on the Real line. Suppose $X_1, X_2, X_3 \overset{\text{iid}}{\sim} H$. Define the four quantities:

$$\begin{aligned}p_1 &= P_H(X_1 > 0) = 1 - H(0) \\ p_2 &= P_H(X_1 + X_2 > 0) = \int_{-\infty}^{\infty} [1 - H(-x_2)]\, dH(x_2) \\ p_3 &= P_H(X_1 + X_2 > 0, X_1 > 0) = \int_{0}^{\infty} [1 - H(-x_1)]\, dH(x_1) \\ p_4 &= P_H(X_1 + X_2 > 0, X_1 + X_3 > 0) = \int_{-\infty}^{\infty} [1 - H(-x_1)]^2\, dH(x_1)\end{aligned}$$

Then, for the Wilcoxon Signed-Rank statistic T_n,

$$\begin{aligned}E_H(T_n) &= np_1 + \frac{n(n-1)}{2}p_2 \\ \text{Var}_H(T_n) &= np_1(1-p_1) + \frac{n(n-1)}{2}p_2(1-p_2) + 2n(n-1)(p_3 - p_1p_2) + \\ &\quad + n(n-1)(n-2)(p_4 - p_2^2)\end{aligned}$$

Example 24.6. Suppose H is symmetric; i.e. $H(-x) = 1 - H(x)$. In this case, $H(0) = 1/2$ and so $p_1 = 1/2$. Also, $p_2 = 1/2$ as $X_1 + X_2$ is symmetric if X_1 and X_2 are independent and symmetric. Therefore,

$$E_H(T_n) = \frac{n}{2} + \frac{n(n-1)}{2} \times \frac{1}{2} = \frac{n(n+1)}{4}.$$

Notice that this matches the expression given earlier. Likewise, $p_3 = 3/8$ and $p_4 = 1/3$. Plugging into the variance formula above we get

$$\mathrm{Var}_H(T_n) = \frac{n(n+1)(2n+1)}{24}.$$

Again, this matches the variance expression we derived earlier.

Remark: It can be shown that for any continuous H, $p_3 = \frac{p_1^2 + p_2}{2}$.

Since T_n takes into account the magnitude as well as the sign of the sample observations, we expect that overall it may have better efficiency properties than the Sign test. The following striking result was proved by Hodges and Lehmann in 1956.

Theorem 24.7. (Hodges-Lehmann, 1956) Define the family of cdf's $\mathfrak{F}$ as

$$\mathfrak{F} = \left\{ F : F \text{ is continuous, } f(z) = f(-z), \sigma_F^2 = \int z^2 f(z)\, dz < \infty \right\}.$$

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F(x-\theta)$. Then the Pitman efficiency of the Wicoxon Signed-Rank test, T, with respect to the t-test, t, is

$$e_P(T, t) = 12\sigma_F^2 \left(\int f^2(z)\, dz \right)^2.$$

Furthermore,

$$\inf_{F \in \mathfrak{F}} e_P(T, t) = \frac{108}{125} = .864,$$

attained at F such that $f(x) = b(a^2 - x^2)$, $|x| < a$, where $a = \sqrt{5}$ and $b = 3\sqrt{5}/20$.

Remark: Notice that the worst case density f is not one of heavy tails, but one with no tails at all (i.e. it has a compact support). Also note that the minimum Pitman efficiency is .864 in the class of symmetric densities with a finite variance, a very respectable lower bound.

Example 24.7. The following table shows the value of the Pitman efficiency for several distributions that belong to the family of cdf's $\mathfrak{F}$ defined in the theorem above. They are obtained by direct calculation using the formula given above. It is interesting that even in the normal case, the Wilcoxon test is 95% efficient with respect to the t test.

F	$e_P(T,t)$
$N(0,1)$	0.95
$U(-1,1)$	1.00
$f(x)=\frac{x^2}{4}e^{-\lvert x\rvert}$	1.26

24.5 Robustness of the t-Confidence Interval

If $X_1,\ldots,X_n \overset{\text{iid}}{\sim} N(\theta,\sigma^2)$ then an exact $100(1-\alpha)\%$ confidence interval for θ is the famous t-confidence interval, C_n, with limits given by

$$\overline{X} \pm t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}},$$

with the property $P_{\theta,\sigma}(C_n \ni \theta) = 1-\alpha\,\forall\, n,\theta,\sigma$. However, if the population is non-normal, then the exact distribution of the statistic $t_n = \frac{\sqrt{n}(\overline{X}-\theta)}{s}$ is *not* t. Consequently, the coverage probability may not be $1-\alpha$, even approximately, for finite n. Asymptotically, the $1-\alpha$ coverage property holds for any population with a finite variance.

Precisely, if $X_1,\ldots,X_n \overset{\text{iid}}{\sim} F$ with $\mu = E_F X_1$ and $\sigma^2 = \text{Var}_F X_1 < \infty$ then

$$t_n = \frac{\sqrt{n}(\overline{X}-\mu)}{s} = \frac{\sqrt{n}(\overline{X}-\mu)/\sigma}{s/\sigma} \overset{\mathcal{L}}{\Rightarrow} N(0,1),$$

since the numerator converges in law to $N(0,1)$ and the denominator converges in probability to 1. Furthermore, for any given α, $t_{\frac{\alpha}{2},n-1} \to z_{\frac{\alpha}{2}}$ as $n\to\infty$. Hence,

$$P_F(C_n \ni \mu) = P_F\left(|t_n| \le t_{\frac{\alpha}{2},n-1}\right) \longrightarrow P(|Z| \le z_{\alpha/2}) = 1-\alpha.$$

That is, given a specific F and fixed α and ϵ

$$1-\alpha-\epsilon \le P_F(C_n \ni \mu) \le 1-\alpha+\epsilon$$

for all $n \ge N = N(F,\alpha,\epsilon)$.

However, if we know only that F belongs to some large class of distributions, $\mathfrak{F}$, then there are no guarantees about the uniform validity of the coverage. In fact, it is possible to have

$$\lim_{n\to\infty}\inf_{F\in\mathfrak{F}} P_F(C_n \ni \mu) = 0.$$

The t confidence interval is a very popular procedure, routinely used for all types of data in practical statistics. An obviously important question is whether this is a safe practice, or more precisely, when is it safe. The literature has some surprises. The t-interval is not unsafe, as far as coverage is concerned, for heavy-tailed data, at least when symmetry is present. The paper Logan et al.(1973) has some major surprises as regards the asymptotic behavior of the t statistic for heavy-tailed data. However, the t-interval can have poor coverage properties when the underlying distribution is skewed.

Example 24.8. A Mathematica simulation was done to check the coverage probabilities of the nominal 95% t-interval for various distributions. The table below summarizes the simulation.

n	$N(0,1)$	$U(0,1)$	$C(0,1)$	Exp(1)	$\log N(0,1)$
10	0.95	0.949	0.988	0.915	0.839
25	0.95	0.949	0.976	0.916	0.896

The table indicates that for symmetric distributions, heavy-tailed or light-tailed, the t-interval does not have a significant coverage bias, for large samples. In fact, for heavy-tails, the coverage of the t-interval could be $> 1 - \alpha$ (of course, at the expense of interval width). However, for skewed data, the t-interval has a deficiency in coverage, even for rather large samples.

The previous example helps motivate our discussion. Now we get to the current state of the theory on this issue.

Definition 24.3. The family of scale mixtures of normal distributions is defined as

$$\mathfrak{F} = \left\{ f(x) = \int_0^\infty \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{x^2}{2\tau^2}}\, dG(\tau), \text{ where G is a Probability measure} \right\}.$$

The family contains many symmetric distributions on the Real line that are heavier tailed than the normal. In particular, t, Double Exponential, Logistic, hyperbolic cosine, etc. all belong to the family $\mathfrak{F}$. Even all symmetric stable distributions belong to $\mathfrak{F}$.

Here is a rather surprising result which says that the coverage of the t confidence interval in the normal scale mixture class is often better than the claimed nominal level.

Theorem 24.8. (Benjamini, 1983) Let $X_i = \mu + \sigma Z_i$ where $Z_i \overset{\text{iid}}{\sim} f \in \mathfrak{F}$ for $i = 1, \ldots, n$. Let C_n be the t-interval given by the formula: $\overline{X} \pm t\frac{s}{\sqrt{n}}$, where t stands for the t percentile $t_{\alpha/2,n-1}$. Then

$$\inf_{f \in \mathfrak{F}} P_{\mu,\sigma,f}(C_n \ni \mu) = P_{\mu,\sigma,\varphi}(C_n \ni \mu),$$

provided $t \geq 1.8$, where φ is the standard normal density.

Remark: The cutoff point 1.8 is not a mathematical certainty; so the theorem is partially numerical. If $\alpha = 0.05$, then $t \geq 1.8$ for all $n \geq 2$. If $\alpha = 0.10$ the result holds for all $n \leq 11$.

This theorem states that the t-interval is safe when the tails are heavy. The natural question now arises what can happen when the tails are light?

The following theorem (see Basu and DasGupta(1995))uses the family of symmetric unimodal densities given by

$$\mathfrak{F}_{su} = \{f : f(z) = f(-z),\ f \text{ is unimodal } \}$$

Theorem 24.9. Let $X_i = \mu + \sigma Z_i$ where $Z_i \overset{\text{iid}}{\sim} f \in \mathfrak{F}_{su}$.

(a) If $t < 1$ then

$$\inf_{f \in \mathfrak{F}_{su}} P_{\mu,\sigma,f}(C_n \ni \mu) = 0\ \forall n \geq 2.$$

(b) For all $n \geq 2$, there exists a number τ_n such that, for $t \geq \tau_n$,

$$\inf_{f \in \mathfrak{F}_{su}} P_{\mu,\sigma,f}(C_n \ni \mu) = P_{\mu,\sigma,U[-1,1]}(C_n \ni \mu).$$

Remark: The problem of determining the value of the above infimum for $1 \leq t \leq \tau_n$ remains unsolved. The theorem also shows that t-intervals with small nominal coverage are arbitrarily bad over the family $\mathfrak{F}_{su}$, while those with a high nominal coverage are quite safe, because the t interval performs quite well for uniform data.

Example 24.9. The values of τ_n cannot be written down by a formula, but difficult calculations can be done to get them, for a given value of n, as shown in the following table. In the table, $1 - \alpha$ is the nominal coverage when the coefficient t equals τ_n.

n	2	5	7	10
τ_n	1.00	1.92	2.00	2.25
$1 - \alpha$	0.50	0.85	0.90	0.95

For example, for all $n \geq 10$, the infimum in the above theorem is attained at symmetric uniform densities, if $\alpha = 0.05$. The next table shows the actual values of the infimum coverage in the symmetric unimodal class for various sample sizes and significance levels.

$\alpha \downarrow$, $n \rightarrow$	2	3	5	7	10
0.2	0.75	0.77	-	-	-
0.1	0.86	0.87	0.89	-	-
0.05	0.92	0.92	0.93	0.94	0.945
0.01	0.98	0.98	0.98	0.983	0.983

Remark: The numerics and the theorems presented indicate that the coverage of the t-interval can have a significant negative bias if the underlying population F is skewed, although for any F with finite variance, we know it to be asymptotically correct. That is,

$$\lim_{n \rightarrow infty} P_F(C_n \ni \mu) = 1 - \alpha.$$

They also indicate that for data from symmetric densities, regardless of tail, the t-interval is quite safe.

We can give a theoretical explanation for why the t-interval is likely to have a negative bias in coverage for skewed F. This explanation is provided by looking at a higher order expansion of the CDF of the t statistic under a general F with some moment conditions. This is the previously described *Edgeworth Expansion* in chapter 13. We recall it below.

Theorem 24.10. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ with $\mathbb{E}_F X_1^4 < \infty$. Assume that F satisfies the Cramér condition. Define

$$\begin{aligned} t_n &= \frac{\sqrt{n}(\overline{X} - \mu(F))}{s} \\ \gamma &= \frac{E_F(X - \mu)^3}{\sigma^3(F)} \\ \kappa &= \frac{E(X - \mu)^4}{\sigma^4(F)} - 3 \end{aligned}$$

Then

$$P_F(t_n \leq t) = \Phi(t) + \frac{p_1(t, F)\varphi(t)}{\sqrt{n}} + \frac{p_2(t, F)\varphi(t)}{n} + o(n^{-1}),$$

where

$$\begin{aligned} p_1(t,F) &= \frac{\gamma(1+2t^2)}{6} \\ p_2(t,F) &= t\left[\frac{\kappa(t^2-3)}{12} - \frac{\gamma^2}{18}(t^4+2t^2-3) - \frac{1}{4}(t^2+3)\right] \end{aligned}$$

Corollary 24.2. If C_n denotes the t-interval then, by the above theorem

$$\begin{aligned} P_F(C_n \ni \mu(F)) &= P_F(|t_n| \le t) \\ &= P_F(t_n \le t) - P_F(t_n \le -t) \\ &= 2\Phi(t) - 1 + \frac{2t}{n}\varphi(t)\left\{\frac{\kappa}{12}(t^2-3) - \frac{\gamma^2}{18}(t^4+2t^2-3) - \frac{1}{4}(t^2+1)\right\} \end{aligned}$$

The corollary shows that when $|\gamma|$ is large, the coverage is likely to have a negative bias and fall below $1-\alpha \approx 2\Phi(t)-1$.

Going back to the asymptotic correctness of the t-interval, for any F with a finite variance, we now show that the validity is not uniform in F.

Theorem 24.11. Let $\mathfrak{F} = \{F : \mathrm{Var}_F(X) < \infty\}$. Then

$$\inf_{F\in\mathfrak{F}} P_F(C_n \ni \mu(F)) = 0, \ \forall\, n \ge 2.$$

Proof. Fix n and take a number c such that $e^{-n} < c < 1$. Let $p_n = p_n(c) = \frac{-\log(c)}{n}$. Take the two point distribution $F = F_{n,c}$ with

$$P_F(X = p_n) = 1 - p_n \ \text{ and } \ P_F(X = p_n - 1) = p_n.$$

Then $\mu(F) = \mathbb{E}_F(X) = 0$ and $\mathrm{Var}_F(X) < \infty$. Now, if all the sample observations are equal to p_n, then the t-interval is just the single point p_n, and hence,

$$\begin{aligned} P_F(C_n \not\ni \mu(F)) &\ge P_F(X_i = p_n, \forall\, i \le n) \\ &= (1-p_n)^n = \left(1 + \frac{\log(c)}{n}\right)^n \end{aligned}$$

But this implies that for any fixed $n \ge 2$,

$$\sup_{F\in\mathfrak{F}} P_F(C_n \not\ni \mu(F)) \ge (1 + \frac{\log(c)}{n})^n \Rightarrow \inf_{F\in\mathfrak{F}} P_F(C_n \ni \mu(F)) = 0$$

by now letting $c \to 1$. □

Remark: The problem here is that we have no control over the skewness in the class $\mathfrak{F}$. In fact, the skewness of the two point distribution F used in the proof is

$$\gamma_F = \frac{2p_n - 1}{\sqrt{p_n(1-p_n)}} \to -\infty, \text{ as } c \to 1.$$

It turns out that with minimal assumptions like a finite variance, no intervals can be produced which are *uniformly* (in F) asymptotically correct and, yet, non-trivial.

To state this precisely, recall the duality between testing and confidence set construction. If $\{\varphi_n\}$ is any (nonrandomized) sequence of test functions, then inversion of the test produces a confidence set C_n for the parameter. The coverage of C_n is related to the testing problem by

$$P_{\theta_0}(C_n \ni \theta_0) = 1 - E_{\theta_0}(\varphi_n).$$

Bahadur and Savage (1956) proved that for sufficiently rich convex families of distribution functions F, there cannot be any tests for the mean which have uniformly small type I error probability and non-trivial (non-zero) power at the same time. This result is considered to be one of the most important results in testing and interval estimation theory.

24.6 The Bahadur-Savage Theorem

Theorem 24.12. (Bahadur & Savage, 1956) Let $\mathfrak{F}$ be any family of cdf's such that

(a) $E_F|X| < \infty$ for all $F \in \mathfrak{F}$,

(b) For any real number r, there is an $F \in \mathfrak{F}$ such that $\mu(F) = E_F(X) = r$, and

(c) If $F_1, F_2 \in \mathfrak{F}$ then for any $0 < \lambda < 1$, $\lambda F_1 + (1-\lambda)F_2 \in \mathfrak{F}$.

Suppose $X_1, X_2, \ldots, X_n$ are iid from some $F \in \mathfrak{F}$ and $C_n = C_n(X_1, X_2, \ldots, X_n)$ a (measurable) set. If there exists an $F_0 \in \mathfrak{F}$ such that $P_{F_0}(C_n$ is bounded from below) $= 1$, then, $\inf_{F \in \mathfrak{F}} P_F(C(n) \ni \mu(F)) = 0$.

Remark: Examples of families of distributions which satisfy the conditions of the Bahadur-Savage theorem are

- The family of all distributions with a finite variance;
- The family of all distributions with all moments finite;

- The family of all distributions with an (unknown) compact support.

It is a consequence of the Bahadur-Savage theorem that in general, we cannot achieve uniform asymptotic validity of the t-interval over rich convex classes. It is natural to ask what additional assumptions will ensure that the t-interval is uniformly asymptotically valid, or, more generally, what assumptions are needed for any uniformly asymptotically valid interval to exist at all. Here is a positive result; notice how the skewness is controlled in this next result. See Lehmann and Romano (2005) for a proof.

Theorem 24.13. Fix a number $b \in (0, \infty)$. Define the family of cdf's

$$\mathfrak{F}_b = \left\{ F : \frac{E_F|X - \mu(F)|^3}{\sigma^3(F)} \leq b \right\}.$$

Then the t-interval is uniformly asymptotically correct over $\mathfrak{F}_b$.

24.7 Kolmogorov-Smirnov & Anderson Confidence Intervals

A second theorem on existence of uniformly asymptotically valid intervals for a mean is due to T.W. Anderson (see Lehmann and Romano(2005)). This construction makes the assumption of a known compact support. The construction depends on the classical goodness-of-fit test due to Kolmogorov and Smirnov, summarized below; see Chapter 26 for more details.

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ and we wish to test $H_0 : F = F_0$. The common sense estimate of the unknown cdf is the empirical cdf F_n. From the Glivenko-Cantelli theorem, we know that

$$\|F_n - F\|_\infty = \sup_x |F_n(x) - F(x)| \longrightarrow 0, \ a.s.$$

However, the statistic

$$D_n = \sqrt{n}\|F_n - F\|_\infty$$

has a non-degenerate limit distribution, and for every n, if the true cdf F is continuous, then D_n has the remarkable property that its distribution is completely independent of F.

The quickest way to see this property is to notice the identity:

$$D_n \overset{\mathcal{L}}{=} \sqrt{n} \max_{1 \leq i \leq n} \max \left\{ \frac{i}{n} - U_{(i)}, U_{(i)} - \frac{i-1}{n} \right\},$$

where $U_{(1)} \leq \ldots \leq U_{(n)}$ are order statistics of an independent sample from $U(0,1)$ and the relation $=_{\mathcal{L}}$ denotes "equality in law".

Therefore, given $\alpha \in (0,1)$, there is a well-defined $d = d_{\alpha,n}$ such that, for any continuous cdf F, $P_F(D_n > d) = \alpha$. Thus,

$$\begin{aligned}
1-\alpha &= P_F(D_n \leq d) \\
&= P_F\left(\sqrt{n}\|F_n - F\|_\infty \leq d\right) \\
&= P_F\left(|F_n(x) - F(x)| \leq \frac{d}{\sqrt{n}}\,\forall\, x\right) \\
&= P_F\left(-\frac{d}{\sqrt{n}} \leq F_n(x) - F(x) \leq \frac{d}{\sqrt{n}}\,\forall\, x\right) \\
&= P_F\left(F_n(x) - \frac{d}{\sqrt{n}} \leq F(x) \leq F_n(x) + \frac{d}{\sqrt{n}}\,\forall\, x\right)
\end{aligned}$$

This gives us a "confidence band" for the true cdf F. More precisely, the $100(1-\alpha)\%$ Kolmogorov-Smirnov confidence band for the cdf F is:

$$KS_{n,\alpha}:\ \max\left\{0, F_n(x) - \frac{d}{\sqrt{n}}\right\} \leq F(x) \leq \min\left\{1, F_n(x) + \frac{d}{\sqrt{n}}\right\}.$$

Remark: The computation of $d = d_{\alpha,n}$ is quite non-trivial but tables are available. See Chapter 26.

Anderson constructed a confidence interval for $\mu(F)$ using the Kolmogorov-Smirnov band for F. The interval is constructed as follows.

$$C_A = \{\mu : \mu = \mu(H) \text{ for some } H \in KS_{n,\alpha}\}.$$

That is, this *interval* contains all μ that are the mean of a *KS-plausible distribution*. With the compactness assumption, the following theorem holds; see Lehmann and Romano (2005).

Theorem 24.14. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$. Suppose F is continuous and supported on a known compact interval $[a,b]$. Then, for any $\alpha \in (0,1)$ and for any n,

$$P_F(C_A \ni \mu(F)) \geq 1 - \alpha.$$

This interval can be computed by finding the associated means for the upper and lower bounds of the KS confidence band.

Remark: So again, with suitable assumptions, in addition to finiteness of variance, uniformly asymptotically valid intervals for the mean exist.

24.8 Hodges-Lehmann Confidence Interval

The Wilcoxon Signed-Rank statistic T_n can be used to construct a point estimate for the point of symmetry of a symmetric density and, out of it, one can construct a confidence interval.

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, where F has a symmetric density, centered at θ. For any pair i, j with $i \leq j$, define the Walsh average $W_{ij} = \frac{1}{2}(X_i + X_j)$ (see Walsh(1959)). Then the Hodges-Lehmann estimate $\hat{\theta}$ is defined as

$$\hat{\theta} = \text{med}\left\{W_{ij} : 1 \leq i \leq j \leq n\right\}.$$

A confidence interval for θ can be constructed using the distribution of $\hat{\theta}$. The interval is found from the following connection with the null distribution of T_n.

Let a be a number such that $P_{\theta=0}(T_n \geq N - a) \leq \frac{\alpha}{2}$, where $N = \frac{n(n+1)}{2}$ is the number of Walsh averages. Let $W_{(1)} \leq \ldots \leq W_{(N)}$ be the ordered Walsh averages. Then, for all continuous symmetric F,

$$P_F\left(W_{(a+1)} \leq \theta(F) \leq W_{(N-a)}\right) \geq 1 - \alpha.$$

This is the Hodges-Lehmann interval for θ.

Remark: We cannot avoid calculation of the N Walsh averages for this method. Furthermore, we must use a table to find a. However, we can approximate a by using asymptotic normality of T_n:

$$\tilde{a} = \frac{n(n+1)}{4} - \frac{1}{2} - z_{\alpha/2}\sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

Alternatively, we can construct a confidence interval for θ based on the Hodges-Lehmann estimate using its asymptotic distribution; see Hettmansperger (1984).

Theorem 24.15. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F(x - \theta)$, where f, the density of F is symmetric around zero. Let $\hat{\theta}$ be the Hodges-Lehmann estimator of θ. Then, if $f \in \mathcal{L}^2$,

$$\sqrt{n}(\hat{\theta} - \theta) \overset{\mathcal{L}}{\Rightarrow} N(0, \tau_F^2),$$

where

$$\tau_F^2 = \frac{1}{12\|f\|_2^4}.$$

Clearly, this asymptotic result can be used to construct a confidence interval for θ in the usual way. That is,

$$P_{\theta,F}\left(\hat{\theta} - \frac{z_{\alpha/2}}{\sqrt{12n\|f\|_2^4}} \leq \theta \leq \hat{\theta} + \frac{z_{\alpha/2}}{\sqrt{12n\|f\|_2^4}}\right) \to 1 - \alpha.$$

Of course, the point of nonparametrics is to make minimal assumptions about the distribution F. Therefore, in general, we do not know f and, hence, we cannot know $\|f\|_2$. However, if we can estimate $\|f\|_2$ then we can simply plug it in to the asymptotic variance formula.

24.9 Power of the Wilcoxon Test

Unlike the null case, the Wilcoxon Signed-Rank statistic T does not have a representation as a sum of independent random variables under the alternative. So the asymptotic non-null distribution of T, which is very useful for approximating the power, does not follow from the CLT for independent summands. However, T still belongs to the class of U-statistics, and hence our previously described CLTs for U-statistics can be used to derive the asymptotic nonnull distribution of T, and thereby get an approximation to the power of the Wilcoxon Signed-Rank test.

Example 24.10. We have previously seen exact formulas for $E_H T_n$ and $\mathrm{Var}_H T_n$ under an arbitrary distribution H. These are now going to be useful for approximation of the power. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F(x - \theta)$ and we want to test $H_0 : \theta = 0$. Take an alternative $\theta > 0$. The power of T_n at θ is

$$\begin{aligned}
\beta(\theta) &= P_\theta(T_n > k_{n,\alpha}) \\
&= P_\theta\left(\frac{T_n - E_\theta(T_n)}{\sqrt{\mathrm{Var}_\theta(T_n)}} > \frac{k_{n,\alpha} - E_\theta(T_n)}{\sqrt{\mathrm{Var}_\theta(T_n)}}\right) \\
&\approx 1 - \Phi\left(\frac{k_{n,\alpha} - E_\theta(T_n)}{\sqrt{\mathrm{Var}_\theta(T_n)}}\right)
\end{aligned}$$

where the normal approximation is made from the CLT for U-statistics. Whether the approximation is numerically accurate is a separate issue.

24.10 Exercises

Exercise 24.1. Prove that the quantile $F_n^{-1}(p)$ is a strongly consistent estimate of $F^{-1}(p)$ under very minimal assumptions.

Exercise 24.2. For each of the following cases, explicitly determine inclusion of how many order statistics of the sample gives an exact nonparametric confidence interval for the median of a density : $n = 20, \alpha = .05, n = 50, \alpha = .05, n = 50, \alpha = .01$.

Exercise 24.3. Find the Pitman efficiency of the Sign test wrt the t test for a triangular, a double exponential, and a logistic density.

Exercise 24.4. * Find the Pitman efficiency of the Sign test wrt the t test for a t density with a general degree of freedom ≥ 3 and plot it.

Exercise 24.5. * Is the Sign test consistent for *any* continuous CDF F ? Prove, or give a concrete counterexample.

Exercise 24.6. * Find the third and the fourth moments of the Wilcoxon signed rank statistic, and hence derive an expression for its skewness and kurtosis. Do they converge to the limiting normal case values ?

Exercise 24.7. Tabulate the exact distribution of the Wilcoxon signed rank statistic when $n = 3, 5, 10$.

Exercise 24.8. * Analytically evaluate the coefficients p_1, p_2, p_3, p_4 when H is a double exponential density centered at a general θ.

Exercise 24.9. Simulate the coverage probability the nominal 95% t confidence interval when the underlying true density is the mixture $.9N(0, 1) + .1N(0, 9)$, the double exponential, and the t density with 5 degrees of freedom. Use $n = 10, 25, 50$.

Exercise 24.10. * Suppose $U \sim U[0, 1]$ and that the underlying true density is U^β. How does the coverage probability of the t interval behave when β is a large positive number ?

Exercise 24.11. * Analytically approximate the coverage probability of the nominal 95% t confidence interval when the underlying true density is an Extreme value density $e^{-e^x}e^x$ by using the Edgeworth expansion of the t statistic.

Exercise 24.12. * Rigorously establish a method of explicitly computing the Anderson confidence interval.

Exercise 24.13. Find the limiting distribution of the Hodges-Lehmann estimate when the underlying true density is a uniform; a triangular; a normal; and a double exponential. Do you see any relation to the tail ?

Exercise 24.14. * By using the asymptotic nonnull distribution, compute an approximate value of the power of the Wilcoxon signed rank test in the $N(\theta, 1)$ model, and plot it. Superimpose it on the power of the t test. Compare.

24.11 References

Bahadur,R.R. and Savage,L.J.(1956). The nonexistence of certain statistical procedures in nonparametric problems,Ann.Math.Statist.,27,1115-1122.

Basu,S. and DasGupta,A.(1995). Robustness of standard confidence intervals for location parameters under departure from normality,Ann.Stat.,23,4,1433-1442.

Benjamini,Y.(1983). Is the t test really conservative when the parent distribution is long tailed ?,JASA,78,383,645-654.

Hajek,J. and Sidak,Z.(1967). Theory of Rank Tests,Academic Press,New York.

Hettmansperger,T.(1984). Statistical Inference Based on Ranks,John Wiley,New York.

Hodges,J.L. and Lehmann,E.L.(1956). The efficiency of some nonparametric competitors of the t test,Ann.math.Statist.,27,324-335.

Jurevckova,J.(1995). Jaroslav Hajek and asymptotic theory of rank tests,Mini Symposium in Honor of Jaroslav Hajek,Kybernetika(Prague),31,3,239-250.

Lehmann,E.L. and Romano,J.(2005). Testing Statistical Hypotheses,Third Ed.,Springer, New York.

Logan,B.F.,Mallows,C.L.,Rice,S.O., and Shepp,L.(1973). Limit distributions of self-normalized sums,Ann.Prob.,1,788-809.

Pitman,E.J.G.(1948). Lecture Notes on Nonparametric Statistical Inference,Columbia University.

Randles,R.H. and Wolfe,D.A.(1979). Introduction to the Theory of Nonparametric Statistics,John Wiley,New York.

Walsh,J.E.(1959). Comments on "The simplest signed-rank tests",JASA,54,213-224.

Wilcoxon,F.(1945). Individual comparisons by ranking methods,Biometrics Bulletin ,1,6,80-83.

25 Two-Sample Problems

Often in applications, we wish to compare two distinct populations with respect to some property. For example, we may want to compare the average salaries of men and women at an equivalent position. Or, we may want to compare the average effect of one treatment with that of another. We may want to compare their variances instead of the mean, or we may even want to compare the distributions themselves. Problems such as these are called two sample problems. In some sense, the two sample problem is more important than the one sample problem. We recommend Hajek and Sidak(1967), Hettmansperger(1984), Randles and Wolfe(1979) and Lehmann and Romano(2005) for further details on the material in this chapter. Additional specific references are given in the sections.

We start with the example of a common two sample parametric procedure in order to introduce a well known hard problem called the Behrens-Fisher problem.

Example 25.1. (Two-sample t-test). Let $X_1, \ldots, X_m \overset{\text{iid}}{\sim} N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu_2, \sigma^2)$, where all $m+n$ observations are independent. Then the two sample t-statistic is

$$T_{m,n} = \frac{\overline{X} - \overline{Y}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}, \quad \text{where } s^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}.$$

Under $H_0 : \mu_1 = \mu_2$, $T_{m,n} \sim t_{m+n-2}$. If $m, n \to \infty$, then $T_{m,n} \overset{\mathcal{L}}{\Rightarrow} N(0,1)$.

More generally, if $X_1, \ldots, X_m \overset{\text{iid}}{\sim} F$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} G$ and F, G have equal mean and variance, then by the CLT and Slutsky's Theorem, we still have $T_{m,n} \overset{\mathcal{L}}{\Rightarrow} N(0,1)$, as $m, n \to \infty$. The asymptotic level and the power of the two-sample t-test are the same for any F, G with equal variance, as they would be when F, G are both normal.

Of course, the assumption of equal variance is not a practical one. However, the corresponding problem with unequal variances, known as the Behrens-Fisher problem, has many difficulties. We discuss it in detail next.

25.1 Behrens-Fisher Problem

Suppose $X_1, \ldots, X_m \overset{\text{iid}}{\sim} N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu_2, \sigma_2^2)$, where all $m+n$ observations are independent. We wish to test $H_0 : \mu_1 = \mu_2$ in the presence of possibly unequal variances. We analyze four proposed solutions to this problem.

There is by now a huge literature on the Behrens-Fisher problem. We recommend Lehmann(1986), Scheffe(1970), and Linnik(1963) for overall exposition of the Behrens-Fisher problem. Here are the four ideas we want to explore.

I. Let $\Delta = \mu_1 - \mu_2$. Then

$$\overline{Y} - \overline{X} \sim N\left(\Delta, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right).$$

Also,

$$\frac{1}{\sigma_1^2}\sum_{i=1}^{m}(X_i - \overline{X})^2 + \frac{1}{\sigma_2^2}\sum_{j=1}^{n}(Y_j - \overline{Y})^2 \sim \chi^2_{m+n-2}.$$

Now, define

$$t = t_{m,n} = \frac{\sqrt{m+n-2}(\overline{Y} - \overline{X} - \Delta)}{\sqrt{\left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)\left[\sigma_1^{-2}\sum_i (X_i - \overline{X})^2 + \sigma_2^{-2}\sum_j (Y_j - \overline{Y})^2\right]}}.$$

Letting $\theta = \sigma_2^2/\sigma_1^2$ we can simplify the above expression to get

$$t = \frac{\sqrt{m+n-2}(\overline{Y} - \overline{X} - \Delta)}{\sqrt{\left(1 + \frac{m}{n}\theta\right)\left[\frac{m-1}{m}s_1^2 + \frac{n-1}{m\theta}s_2^2\right]}}.$$

However, t is not a "statistic" because it depends on the unknown θ. This is unfortunate because if θ were known (i.e. if we know the ratio of the two variances) then the statistic t could be used to test the hypothesis $\Delta = \Delta_0$..

II. Consider the two-sample t-statistic suitable for the equal variance case. That is, consider

$$T = T_{m,n} = \frac{\overline{Y} - \overline{X} - \Delta_0}{\sqrt{s_u^2\left(\frac{1}{m} + \frac{1}{n}\right)}},$$

where $s_u^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}$. We know that, under $H_0 : \Delta = \Delta_0$, the distribution of T is exactly t_{m+n-2} only if $\sigma_1 = \sigma_2$.

But what happens for large samples, in the case $\sigma_1 \neq \sigma_2$? By a simple application of Slutsky's Theorem, it is seen that, when $\sigma_1 \neq \sigma_2$, if $m, n \to \infty$ in such a way that $\frac{m}{m+n} \to \rho$, then

$$T_{m,n} \overset{\mathcal{L}}{\Rightarrow} N\left(0, \frac{(1-\rho) + \rho\theta}{\rho + (1-\rho)\theta}\right), \quad \text{under } H_0.$$

Notice that, if $\rho = \frac{1}{2}$, then $T_{m,n} \stackrel{\mathcal{L}}{\Rightarrow} N(0,1)$. That is, if m and n are large and $m \approx n$, then $T_{m,n}$ can be used to construct a test. However, if m and n are very different, one must also estimate θ.

III. We next consider the likelihood ratio test for the Behrens-Fisher problem $H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \neq \mu_2$. The LRT statistic is $\lambda = -2 \log \Lambda$, where

$$\Lambda = \frac{\sup_{H_0} l(\mu_1, \mu_2, \sigma_1, \sigma_2)}{\sup_{H_0 \cup H_1} l(\mu_1, \mu_2, \sigma_1, \sigma_2)},$$

where $l(.)$ denotes the likelihood function. Then H_0 is rejected for large values of λ. The statistic Λ itself is a complicated function of the data. Of course, the denominator is found by plugging in the unconstrained MLE's

$$\hat{\mu}_1 = \overline{X}, \;\; \hat{\mu}_2 = \overline{Y}, \;\; \hat{\sigma}_1^2 = \frac{1}{m}\sum_i (X_i - \overline{X})^2, \;\; \hat{\sigma}_2^2 = \frac{1}{n}\sum_j (Y_j - \overline{Y})^2.$$

Let $\hat{\mu}$ be the MLE of the common mean $\mu \, (= \mu_1 = \mu_2)$, under H_0. Then, the MLE's of the two variances under H_0 are

$$\hat{\sigma}_1^2 = \frac{1}{m}\sum_i (X_i - \hat{\mu})^2 \;\; \text{and} \;\; \hat{\sigma}_2^2 = \frac{1}{n}\sum_j (Y_j - \hat{\mu})^2.$$

It can be shown that the MLE $\hat{\mu}$ is *one* of the roots of the cubic equation:

$$A\mu^3 + B\mu^2 + C\mu + D = 0,$$

where

$$\begin{aligned}
A &= -(m+n) \\
B &= (m+2n)\overline{X} + (n+2m)\overline{Y} \\
C &= \frac{m(n-1)}{n} s_2^2 + \frac{n(m-1)}{m} s_1^2 \\
D &= m\overline{X}\left(\frac{n-1}{n} s_2^2 + \overline{Y}^2\right) + n\overline{Y}\left(\frac{m-1}{m} s_1^2 + \overline{X}^2\right)
\end{aligned}$$

In the event that the above equation has three real roots, the actual MLE has to be picked by examination of the likelihood function. The MLE is the unique root if the above equation has only one real root.

Therefore, the numerator of Λ is not analytically expressible, but at least asymptotically it can be used, because we have a CLT for λ under the null (see Chapter 21).

IV. The final proposed solution of the Behrens-Fisher problem is due to Welch (Welch(1949)). We know that under the null, $\overline{Y} - \overline{X} \sim N(0, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})$. Welch considered the statistic

$$W = W_{m,n} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{\overline{X} - \overline{Y} \div \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}} \div \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \qquad (\star)$$

It is clear that W is *not* of the form $N(0,1)/\sqrt{\chi_d^2/d}$, with the two variables being independently distributed. Let D^2 be the square of the denominator in the right-hand side of $(\star)$. Welch wanted to write $D^2 \approx \chi_f^2/f$ by choosing an appropriate f. Since the means already match, i.e., $E_{H_0}(D^2)$ is already 1, Welch decided to match the second moments. The following formula for f then results:

$$f = \frac{(\lambda_1\sigma_1^2 + \lambda_2\sigma_2^2)^2}{\frac{\lambda_1^2\sigma_1^4}{m-1} + \frac{\lambda_2^2\sigma_2^4}{n-1}}, \quad \text{where } \lambda_1 = \frac{1}{m}, \lambda_2 = \frac{1}{n}.$$

If we plug in s_1^2 and s_2^2 for σ_1^2 and σ_2^2, respectively, we get Welch's random degree of freedom t-test. That is, we perform a test based on Welch's procedure by comparing W to a critical value from the $t_{\hat{f}}$ distribution, where $\hat{f} = f(s_1^2, s_2^2)$.

Example 25.2. The behavior of Welch's test has been studied numerically and theoretically. As regards the size of the Welch test for normal data, the news is good. The deviation of the actual size from the nominal α is small, even for small or moderate m, n. The following table was taken from Wang (1971). Let M_α denote the maximum deviation of the size of the Welch test from α. In this example, $\alpha = 0.01$, and $\theta = \frac{\sigma_2^2}{\sigma_1^2}$.

m	n	$1/\theta$	$M_{0.01}$
5	21	2	0.0035
7	7	1	0.0010
7	13	4	0.0013
7	19	2	0.0015
13	13	1	0.0003

Regarding the power of the Welch test, it is comparative to the likelihood ratio test. See, for example, Table 2 in Best & Rayner (1987).

Remark: Pfanzagal(1974) has proved that the Welch test has some local asymptotic power optimality property. It is also very easy to implement. Its size is very

close to the nominal level α. Due to these properties, the Welch test has become quite widely accepted as the standard solution to the Behrens-Fisher problem. It is not clear, however, that the Welch test is even size robust when the individual groups are not normal.

In the case when no reliable information is known about the distributional shape, we may want to use a nonparametric procedure. That is our next topic.

25.2 Wilcoxon Rank-Sum and Mann-Whitney Test

Suppose $X_1, \ldots, X_m \overset{\text{iid}}{\sim} F(x-\mu)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F(y-\nu)$, where $F(\cdot)$ is symmetric about zero. In practical applications, we often want to know if $\Delta = \nu - \mu > 0$. This is called the problem of testing for treatment effect.

Let $R_i = \text{rank}(Y_i)$ among all $m+n$ observations. The Wilcoxon Rank-Sum statistic is $U = \sum_{i=1}^{n} R_i$. Large values of U indicate that there is, indeed, a treatment effect; i.e. that $\Delta > 0$ (Mann and Whitney(1947)). To execute the test, we need the smallest value ξ such that $P_{H_0}(U > \xi) \leq \alpha$.

In principle, the null distribution of U can be found from the joint distribution of $(R_1, \ldots, R_n)$. In particular, if $N = m+n$, then the marginal and two dimensional distributions are, respectively,

$$\begin{aligned} P(R_i = a) &= \frac{1}{N}, \ \ \forall\, a = 1, \ldots, N \\ P(R_i = a, R_j = b) &= \frac{1}{N(N-1)}, \ \ \forall\, a \neq b \end{aligned}$$

It follows immediately that

$$E_{H_0}(U) = \frac{n(N+1)}{2} \quad \text{and} \quad \text{Var}_{H_0}(U) = \frac{mn(N+1)}{12}.$$

This will be useful for the asymptotic distribution, which we discuss a little later. As for the exact null distribution of U, we have the following proposition; see Hettmansperger (1984).

Proposition 25.1. Let $p_{m,n}(k) = P_{H_0}(U = k + n(n+1)/2)$. Then the following recursive relation holds:

$$p_{m,n}(k) = \frac{n}{m+n} p_{m,n-1}(k-m) + \frac{m}{m+n} p_{m-1,n}(k).$$

From here, the exact finite sample distributions can be numerically computed for moderate values of m, n.

There is an interesting way to rewrite U. From its definition,

$$R_i = \#\{k : Y_k \leq Y_i\} + \#\{j : X_j < Y_i\},$$

which implies

$$U \equiv \sum_{i=1}^{n} R_i = \frac{n(n+1)}{2} + \#\{(i,j) : X_j < Y_i\}.$$

Then the statistic

$$W = U - \frac{n(n+1)}{2} = \sum_{i=1}^{n} R_i - \frac{n(n+1)}{2} = \#\{(i,j) : X_j < Y_i\}$$

is called the Mann-Whitney statistic. Note the obvious relationship between the Wilcoxon Rank-Sum statistic U and the Mann-Whitney statistic W, in particular

$$U = W + \frac{n(n+1)}{2}.$$

Therefore, $E_{H_0}(W) = \frac{mn}{2}$ and $\mathrm{Var}_{H_0}(W) = \frac{mn(N+1)}{12}$.

The Mann-Whitney test rejects H_0 for large values of W and is obviously equivalent to the Wilcoxon Rank-Sum test.

It follows from one-sample U-statistics theory that, under the null hypothesis, W is asymptotically normal; see Hettmansperger (1984) for the formal details of the proof.

Theorem 25.1. Let $X_1, \ldots, X_m \overset{\text{iid}}{\sim} F(x-\mu)$, $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F(y-\nu)$, all $m+n$ observations independent. Suppose that $F(\cdot)$ is symmetric about zero. Under the null hypothesis $H_0 : \Delta = \nu - \mu = 0$,

$$\frac{W - E_{H_0}(W)}{\sqrt{\mathrm{Var}_{H_0}(W)}} \overset{\mathcal{L}}{\Rightarrow} N(0,1).$$

Therefore, a cutoff value for the α-level test can be found via the CLT:

$$k_\alpha = \frac{mn}{2} + \frac{1}{2} + z_\alpha \sqrt{\frac{mn(N+1)}{12}},$$

where the additional $\frac{1}{2}$ that is added is a continuity correction.

Recall that a point estimate due to Hodges and Lehmann is available in the one sample case based on the Signed-Rank test. In particular, it was the median of the

Walsh averages (see Chapter 24). We can do something similar in the two-sample case.

$$W = \#\{(i,j) : X_j < Y_i\} = \#\{(i,j) : D_{ij} \equiv Y_i - X_j > 0\}.$$

This motivates the estimator $\hat{\Delta} = \text{med}\{D_{ij}\}$. It turns out that an interval containing adequately many order statistics of the D_{ij} has a guaranteed coverage for all m, n; see Hettmansperger (1984) for a proof. Here is the theorem.

Theorem 25.2. Let $k = k(m, n, \alpha)$ be the largest number such that $P_{H_0}(W \leq k) \leq \alpha/2$. Then $(D_{(k+1)}\,,\, D_{(mn-k)})$ is a $100(1-\alpha)\%$ confidence interval for Δ under the above shift model.

Remark: Tables can be used to find k for a given case; see, for example, Milton (1964). This is a widely accepted nonparametric confidence interval for the two sample location parameter problem.

A natural question would now be how does this test compare to, say, the t-test? It can be shown that the Mann-Whitney test is consistent for any two distributions such that $P(Y > X) > \frac{1}{2}$. That is, if $X_1, \ldots, X_m \overset{\text{iid}}{\sim} G$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} H$ and if $X \sim G$, $Y \sim H$ and X, Y are independent, then consistency holds if

$$P_{G,H}(Y > X) > \frac{1}{2}.$$

Pitman efficiencies, under the special shift model, can be found and they turn out to be the same as in the one-sample case. See Lehmann(1986) for details on consistency and efficiency of the Mann-Whitney test.

25.3 Two-Sample U-Statistics & Power Approximations

The asymptotic distribution of W under the alternative is useful for approximating the power of the Mann-Whitney test. Under the null, we used one-sample U-statistics theory to approximate the distribution. However, under the alternative, there are two underlying distributions, namely $F(x-\mu)$ and $F(y-\nu)$. So the asymptotic theory of one-sample U-statistics cannot be used under the alternative. We will need the theory of two-sample U-statistics. See Serfling(1980) for derivations of the basic formulae needed in this section.

Below we will consider two-sample U-statistics in a much more general scenario than the location shift model.

Definition 25.1. Fix $0 < r_1, r_2 < \infty$ and let $h(x_1, \ldots, x_{r_1}, y_1, \ldots, y_{r_2})$ be a real-valued function on $\mathbb{R}^{r_1+r_2}$. Furthermore, assume that h is permutation invariant among the x's and among the y's, separately. Let $X = X_1, \ldots, X_m \overset{\text{iid}}{\sim} F$ and $Y = Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} G$, and suppose all observations are independent. Define

$$U_{m,n} = U_{m,n}(X, Y) = \frac{1}{\binom{m}{r_1}\binom{n}{r_2}} \sum h\left(X_{i_1}, \ldots, X_{i_{r_1}}, Y_{j_1}, \ldots, Y_{j_{r_2}}\right)$$

where the sum is over all $1 \le i_1 \le \ldots \le i_{r_1} \le m, 1 \le j_1 \le \ldots \le j_{r_2} \le n$. Then $U_{m,n}$ is called a two-sample U-statistic with kernel h and indices r_1 and r_2.

Example 25.3. Let $r_1 = r_2 = 1$ and $h(x, y) = I_{y>x}$. Then it is easy to verify that $U_{m,n}$, based on this kernel, is the Mann-Whitney test statistic W.

As in the one sample case, $U_{m,n}$ is asymptotically normal, under suitable conditions. We use the following notation.

$$\begin{aligned}
\theta &= \theta(F, G) = E_{F,G} h(X_1, \ldots, X_{r_1}, Y_1, \ldots, Y_{r_2}) \\
h_{10}(x) &= E_{F,G}[h(X_1, \ldots, X_{r_1}, Y_1, \ldots, Y_{r_2}) | X_1 = x] \\
h_{01}(y) &= E_{F,G}[h(X_1, \ldots, X_{r_1}, Y_1, \ldots, Y_{r_2}) | Y_1 = y] \\
\zeta_{10} &= \mathrm{Var}_F h_{10}(X) \\
\zeta_{01} &= \mathrm{Var}_G h_{01}(Y)
\end{aligned}$$

Theorem 25.3. Assume that $E_{F,G} h^2 < \infty$ and $\zeta_{10}, \zeta_{01} > 0$. Also, assume that $m, n \to \infty$ in such a way that $\frac{m}{m+n} \to \lambda \in (0, 1)$. Then

$$\sqrt{n}(U_{m,n} - \theta) \overset{\mathcal{L}}{\Rightarrow} N(0, \sigma^2_{F,G}),$$

where

$$\sigma^2_{F,G} = \frac{r_1^2 \zeta_{10}}{\lambda} + \frac{r_2^2 \zeta_{01}}{1 - \lambda}.$$

Remark: Sometimes it is more convenient to use the true variance of $U_{m,n}$ and the version that states

$$\frac{U_{m,n} - \theta}{\sqrt{\mathrm{Var}_{F,G}(U_{m,n})}} \overset{\mathcal{L}}{\Rightarrow} N(0, 1).$$

Recall that if $r_1 = r_2 = 1$ and $h(x, y) = I_{y>x}$ then $U_{m,n}$ is the Mann-Whitney test statistic W. In this case, we have exact formulae for θ and $\mathrm{Var}_{F,G}(U_{m,n})$.

Let $X_1, X_2 \overset{\text{iid}}{\sim} F$ and $Y_1, Y_2 \overset{\text{iid}}{\sim} G$ and define

$$\begin{aligned} p_1 &= P_{F,G}(Y_1 > X_1) = \int (1 - G(x)) f(x)\, dx \\ p_2 &= P_{F,G}(Y_1 > X_1, Y_2 > X_1) = \int (1 - G(x))^2 f(x)\, dx \\ p_3 &= P_{F,G}(Y_1 > X_1, Y_1 > X_2) = \int F^2(y) g(y)\, dy \end{aligned}$$

Proposition 25.2. Let W be the Mann-Whitney statistic. Then

$$\begin{aligned} E_{F,G}(W) &= mnp_1 \\ \text{Var}_{F,G}(W) &= mn(p_1 - p_1^2) + mn(n-1)(p_2 - p_1^2) + mn(m-1)(p_3 - p_1^2) \end{aligned}$$

See Hettmansperger (1984) for the above formulae.

Remark: We can use these formulae to compute the approximate quantiles of W by approximating $\frac{W - E(W)}{\sqrt{\text{Var}(W)}}$ by a $N(0,1)$. Another option is to use $\sigma^2_{F,G}$ in place of the exact variance of W. Here, $\sigma^2_{F,G}$ is the asymptotic variance, as defined before. To use the alternative expression $\sigma^2_{F,G}$, we need to compute ζ_{10} and ζ_{01}. For this computation, it is useful to note that

$$\zeta_{10} = \text{Var}_F G(X) \quad \text{and} \quad \zeta_{01} = \text{Var}_G F(Y).$$

Remark: Note that the exact variance of W, as well as ζ_{10} and ζ_{01}, are functionals of F and G, which we cannot realistically assume to be known. In practice, all of these functionals must be estimated (usually by a plug-in estimator; e.g. in the formula for p_1, replace F by some suitable $\hat{F}_m$ and G by some suitable $\hat{G}_n$, so that, e.g. $\hat{p}_1 = \int (1 - \hat{G}_n(x))\, d\hat{F}_m(x)$).

25.4 Hettmansperger's Generalization

So far, we have considered the two cases $N(\theta_1, \sigma_1^2)$ vs. $N(\theta_2, \sigma_2^2)$ and $F(x - \mu)$ vs. $F(y - \nu)$. For the first case, we have settled on Welch's solution. For the second, we like the Mann-Whitney test. However, even this second model does not allow the scale parameters to differ. This is our next generalization.

Let $X_i \overset{\text{iid}}{\sim} F\left(\frac{x-\mu}{\rho}\right), 1 \le i \le m$, and $Y_j \overset{\text{iid}}{\sim} F\left(\frac{y-\nu}{\tau}\right), 1 \le j \le n$, where ρ and τ are unknown, and as usual, we assume that all observations are independent. We wish to test $H_0 : \mu = \nu$.

A test due to Hettmansperger (1973) with reasonable properties is the following. Let $S = \#\{j : Y_j > \text{med}(X_i)\}$ and $S^* = \#\{i : X_i < \text{med}(Y_j)\}$. Also, let $\sigma = \tau/\rho$. Then define

$$\begin{aligned} T_\sigma &= \sqrt{n}\left(\frac{S}{n} - \frac{1}{2}\right) \div \sqrt{\frac{1 + n/(m\sigma^2)}{4}} \\ T_\sigma^* &= \sqrt{m}\left(\frac{S^*}{m} - \frac{1}{2}\right) \div \sqrt{\frac{1 + m\sigma^2/n}{4}} \end{aligned}$$

Then the test statistic is $T = \min\{T_1, T_1^*\}$. This test rejects $H_0 : \mu = \nu$ if $T > z_\alpha$, a normal quantile.

Theorem 25.4. Let $X_1, \ldots, X_m \overset{\text{iid}}{\sim} F\left(\frac{x-\mu}{\rho}\right)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F\left(\frac{y-\nu}{\tau}\right)$, with all parameters unknown and all $(m+n)$ observations independent. Assume that F is absolutely continuous, 0 is the unique median of F, and that there exists $\lambda \in (0,1)$ such that $\frac{m}{m+n} \to \lambda$. Consider testing $H_0 : \mu = \nu$ against $H_1 : \nu > \mu$. Then

$$\lim_{m,n\to\infty} P_{H_0}(T > z_\alpha) \leq \alpha.$$

Proof. (Sketch) By a standard argument involving Taylor expansions, for all $\sigma > 0$, each of T_σ and T_σ^* are asymptotically $N(0,1)$. From the monotonicity in σ, it follows that $T_\sigma > T_1$ and $T_\sigma^* < T_1^*$ when $\sigma > 1$ and the inequalities are reversed for $\sigma \leq 1$. Therefore, when $\sigma > 1$,

$$P(T_1 > c) \leq P(T_\sigma > c)$$

while

$$P(T_1^* > c) \geq P(T_\sigma^* > c).$$

If we set $c \equiv z_\alpha$ then, when $\sigma > 1$ we get

$$P(T_1^* > z_\alpha) \geq P(T_\sigma^* > z_\alpha) \approx \alpha \approx P(T_\sigma > z_\alpha) \geq P(T_1 > z_\alpha),$$

where $\approx$ holds due to the asymptotic normality of T_σ and T_σ^*. Similarly, when $\sigma \leq 1$ we get the opposite string of inequalities. Consequently,

$$\begin{aligned} P(T > z_\alpha) &= P(\min\{T_1, T_1^*\} > z_\alpha) \\ &= P(T_1 > z_\alpha, T_1^* > z_\alpha) \\ &\leq \min\{P(T_1 > z_\alpha), P(T_1^* > z_\alpha)\} \\ &\approx \alpha \end{aligned}$$

This explains why the limiting size of the test is $\leq \alpha$. □

Remark: The theorem says that the Hettmansperger test is asymptotically distribution free under H_0 and asymptotically conservative with regards to size. An approximation for the true size of this test is

$$\alpha_{\text{true}} \approx 1 - \Phi\left(z_\alpha \sqrt{\frac{1+c}{\sigma^2 + c}} \max\{1, \sigma\} \right),$$

where $c = \lim \frac{n}{m}$. Note that the right-hand side is equal to α when $\sigma = 1$. See Hettmansperger (1973) for a derivation of this approximation.

Simulations show that, if $\sigma \approx 1$, then the true size is approximately α. However, if σ is of the order 2 or 1/2, then the test is severely conservative.

In summary, the test proposed by Hettmansperger is reasonable to use if

- the two populations have the same shape,
- the two populations have approximately the same scale parameters,
- $F(0) = \frac{1}{2}$, although F need not be symmetric.

25.5 The Nonparametric Behrens-Fisher Problem

This is the most general version of the two-sample location problem, with as few assumptions as possible. We want to construct a test which is, at least asymptotically, distribution free under H_0 and consistent against a broad class of alternatives. The model is as follows: Suppose $X_1, \ldots, X_m \overset{\text{iid}}{\sim} F$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} G$, where F, G are arbitrary distribution functions. To avoid the difficulties of ties, we assume that F and G are (absolutely) continuous. Let $\mu = \text{med}(F)$ and $\nu = \text{med}(G)$. We want to test $H_0 : \mu = \nu$ vs. $H_1 : \nu > \mu$. See Fligner and Policello(1981) and Brunner and Munzel(2000) for the development in this section. Johnson and Weerahandi(1988) and Ghosh and Kim(2001) give Bayesian solutions to the Behrens-Fisher problem, while Babu and Padmanabhan(2002) use resampling ideas for the nonparametric Behrens-Fisher problem.

It turns out that a Welch-type statistic, which was suitable for the ordinary Behrens-Fisher problem, is used here. But unlike the Welch statistic used in the case of two normal populations, this test will use the ranks of the X_i's and the Y_j's.

Let Q_i and R_j denote the ranks of $X_{(i)}$ and $Y_{(j)}$ among all $m + n$ observations, respectively. Here, as usual, $X_{(i)}$ and $Y_{(j)}$ denote the respective order statistics.

Also, let $P_i = Q_i - i$ and $O_j = R_j - j$. That is,

$$P_i = \#\{j : Y_j \leq X_{(i)}\},$$

$$O_j = \#\{i : X_i \leq Y_{(j)}\},$$

Our test function comes from a suitable function $T = T_{m,n}$ of the vector of ranks of the Y sample in the combined sample. We would like to choose T in such a way that whatever be F and G,

$$\frac{T - \theta_{F,G}}{\tau_{F,G}} \stackrel{\mathcal{L}}{\Rightarrow} N(0,1), \quad \text{where } \begin{cases} \theta_{F,G} = E_{F,G}(T) \\ \tau^2_{F,G} = \mathrm{Var}_{F,G}(T) \end{cases}.$$

There is some additional complexity here because $\mu = \nu$ does not force the two distributions to be the same. Thus, we need to estimate the variance $\tau^2_{F,G}$ under the null for general F, G. If we can do this and, moreover, if $\theta_{F,G} = \theta_0$, a fixed number, then we can use a standardized test statistic such as

$$\frac{T - \theta_0}{\hat{\tau}_{F,G}}.$$

So, the choice of T will be governed by the ease of finding θ_0 and of finding $\hat{\tau}_{F,G}$, under H_0. The statistic that is used is

$$T = \frac{1}{mn} \sum_{i=1}^{m} P_i.$$

It turns out that mnT is actually the same as the Mann-Whitney statistic. For such a choice of T, under H_0, $\theta_{F,G} = \theta_0 = 1/2$. Also, $\tau^2_{F,G}$ has the exact formula given earlier in this chapter. Of course, this formula involves the unknown F, G so we need to plug in the empirical cdf's $\hat{F}_m$ and $\hat{G}_n$. Doing so, we get the test statistic

$$\hat{W} = \frac{\frac{W}{mn} - \frac{1}{2}}{\hat{\tau}_{F,G}} = \frac{\sum_i P_i - \sum_j O_j}{2\sqrt{\sum_i (P_i - \overline{P})^2 + \sum_j (O_j - \overline{O})^2 + \overline{P} \times \overline{O}}}.$$

Then the test rejects H_0 for large values of $\hat{W}$. If $k = k_{m,n}(\alpha)$ is the cutoff for the Mann-Whitney statistic itself, then, specifically, the test that rejects H_0 for $\hat{W} > k$ has the property that if $F = G$ under H_0 (which means that equality of medians forces equality of distributions), then the size is still α. If $F \neq G$ under H_0, then some more assumptions are needed to maintain the size and for reasonable consistency properties.

Observe that

$$\begin{aligned}\lim_{m,n\to\infty} P_{F,G}\left(\hat{W} > k\right) &= \lim_{m,n\to\infty} P_{F,G}\left(\frac{W/mn - 1/2}{\hat{\tau}_{F,G}} > k\right) \\ &= \lim_{m,n\to\infty} P_{F,G}\left(\frac{W/mn - \theta_{F,G} + \theta_{F,G} - 1/2}{\hat{\tau}_{F,G}} > k\right) \quad (\star)\end{aligned}$$

But, from the general two-sample U-statistics theory, we know that

$$\frac{W/mn - \theta_{F,G}}{\hat{\tau}_{F,G}} \overset{\mathcal{L}}{\Rightarrow} N(0,1),$$

and $k = k_{m,n}(\alpha) \to z_\alpha$. Then, clearly, the above limit is equal to α if and only if $\theta_{F,G} = \frac{1}{2}$, under H_0. Recall now that

$$\theta_{F,G} = \int F\, dG = P_{F,G}(Y > X) = P_{F,G}(Y - X > 0).$$

This is $\frac{1}{2}$ under H_0 if and only if $Y - X$ has median zero. If $X \sim F$ and $Y \sim G$ and F, G are symmetric about some μ and ν then $Y - X$ is symmetric about 0 when $\mu = \nu$. In that case, $P_{F,G}(Y - X > 0) = \frac{1}{2}$ holds automatically and the size of $\hat{W}$ is asymptotically maintained.

Also, from $(\star)$ above, we see that the power under any F, G converges to 1 if and only if $\theta_{F,G} > \frac{1}{2}$. That is, the test $\hat{W}$ is consistent against those alternatives (F, G) for which

$$\int F\, dG = P_{F,G}(Y - X > 0) > \frac{1}{2}.$$

Remark: If we interpret our hypothesis of interest as

$$H_0 : P_{F,G}(Y - X > 0) = \frac{1}{2} \quad \text{vs. } H_1 : P_{F,G}(Y - X > 0) > \frac{1}{2},$$

then, without any assumptions, the test based on $\hat{W}$ maintains its size asymptotically and is consistent.

Example 25.4. Below are some tables giving the size and power of the $\hat{W}$ test for selected F, G, m, n. For the size values, it is assumed that $G(y) = F(y/\sigma)$, and for the power values, it is assumed that $G(y) = F(y - \Delta)$.

Size($\times$1000) Table: $\alpha = 0.05$, $m = 11$ & $n = 10$

F	σ	W	$\hat{W}$	Welch
N(0,1)	0.1	81	48	18
	0.25	69	54	52
	1	50	48	47
	4	71	54	47
	10	82	62	52
Double Exp	0.1	75	51	45
	0.25	65	54	49
	1	50	48	46
	4	67	54	45
	10	84	62	49
C(0,1)	0.1	69	51	26
	0.25	62	54	26
	1	49	48	25
	4	64	52	25
	10	80	62	30

Power ($\times 1000$) Table: $\alpha = 0.05$, $m = 25$ & $n = 20$

F	Δ	W	$\hat{W}$	Welch
N(0,1)	0.1	195	209	207
	0.2	506	520	523
	0.3	851	860	870
Double Exp	0.1	150	159	128
	0.2	403	411	337
	0.3	791	797	699
C(0,1)	0.1	140	145	49
	0.2	348	352	97
	0.3	726	723	88

Notice the nonrobustness of Welch's test, which we had commented on earlier.

25.6 Robustness of the Mann-Whitney Test

A natural question is what happens to the performance of the Mann-Whitney test itself under general F and G ? Here we consider only the asymptotic size of the test. Similar calculations can be done to explore the robustness of the asymptotic

power.

The true size of W is

$$\begin{aligned} P_{F,G}(W > k_{m,n}(\alpha)) &\approx P_{F,G}\left(\frac{W - mn/2}{\sqrt{mn(m+n+1)/12}} > z_\alpha\right) \\ &= P_{F,G}\left(\frac{W - mn\theta_{F,G} + mn(\theta_{F,G} - 1/2)}{\sqrt{mn(m+n+1)/12}} > z_\alpha\right) \end{aligned}$$

Suppose now that $\theta_{F,G} = 1/2$. As mentioned above, symmetry of each of F and G will imply $\theta_{F,G} = 1/2$, under H_0. In such a case,

$$\begin{aligned} \alpha_{\text{true}} &\equiv P_{F,G}(W > k_{m,n}(\alpha)) \\ &\approx P_{F,G}\left(\frac{W - mn\theta_{F,G}}{\sqrt{mn(m+n+1)/12}} > z_\alpha\right) \\ &= P_{F,G}\left(\frac{W - mn\theta_{F,G}}{\sqrt{v(p_1,p_2,p_3)}} \times \frac{\sqrt{v(p_1,p_2,p_3)}}{\sqrt{mn(m+n+1)/12}} > z_\alpha\right) \end{aligned}$$

where

$$v(p_1,p_2,p_3) = mn(p_1 - p_2^2) + mn(n-1)(p_2 - p_1^2) + mn(m-1)(p_3 - p_1^2)$$

where $v(p_1,p_2,p_3)$ denotes the exact variance of W, and the p_i are as defined earlier in section 25.3. Then the right-hand side has the limit

$$\longrightarrow 1 - \Phi\left(\frac{z_\alpha}{\sqrt{12[\lambda(p_3 - p_1^2) + (1-\lambda)(p_2 - p_1^2)]}}\right), \quad \text{as } m,n \to \infty,$$

and $\lambda = \lim_{m,n\to\infty} \frac{m}{m+n}$. Notice that, in general, this is not equal to α.

Example 25.5. In the case where $G(y) = F(y/\sigma)$, where F is symmetric and absolutely continuous, it follows from the known formulae for p_i that

$$\begin{aligned} \lim_\sigma (p_2 - p_1^2) &= \begin{cases} \frac{1}{4}, & \sigma \to 0 \\ 0, & \sigma \to \infty \end{cases} \\ \lim_\sigma (p_3 - p_1^2) &= \begin{cases} 0, & \sigma \to 0 \\ \frac{1}{4}, & \sigma \to \infty \end{cases} \end{aligned}$$

(see below for a proof). Plugging into the above formula for α_{true}, if $G(y) = F(y/\sigma)$ then

$$\lim_\sigma \alpha_{\text{true}} = \begin{cases} 1 - \Phi\left(\frac{z_\alpha}{\sqrt{3(1-\lambda)}}\right), & \sigma \to 0 \\ 1 - \Phi\left(\frac{z_\alpha}{\sqrt{3\lambda}}\right), & \sigma \to \infty \end{cases}$$

and the limit is between these values for $0 < \sigma < \infty$. If, for example, $\alpha = 0.05$ and $\lambda = 1/2$, then $\lim_\sigma \alpha_{\text{true}}$ varies between 0.05 and 0.087, which is reasonably robust.

For illustration, we prove the limiting behavior of α_{true} as $\sigma \to 0, \infty$ more carefully. Since we assume $G(y) = F(y/\sigma)$ and F, G are symmetric, it follows that $\theta_{F,G} = \int F\, dG = 1/2$. To evaluate the limits of α_{true}, we need only evaluate the limits of $p_2 - p_1^2$ and $p_3 - p_1^2$ as $\sigma \to 0, \infty$. First,

$$\begin{aligned} p_2 &= P_{F,G}(Y_1 > X_1, Y_2 > X_1) = \int (1 - G(x))^2\, dF(x) = \int (1 - F(x/\sigma))^2\, dF(x) \\ &= \int_{x>0} (1 - F(x/\sigma))^2\, dF(x) + \int_{x<0} (1 - F(x/\sigma))^2\, dF(x) \end{aligned}$$

Now, $F(x/\sigma) \to 1$ as $\sigma \to 0$ for all $x > 0$, and $F(x/\sigma) \to 0$ as $\sigma \to \infty$ for all $x < 0$. So, by the Lebesgue Dominated Convergence Theorem (DCT), as $\sigma \to 0$, we get $p_2 \to 0 + 1/2 = 1/2$. If $\sigma \to \infty$, then for all x, $F(x/\sigma) \to 1/2$ so, again by the DCT, $p_2 \to 1/4$. Next, for p_3, when $\sigma \to 0$, we get

$$p_3 = \int F^2(x)\, dG(x) = \int F^2(\sigma x)\, dF(x) \to \frac{1}{4}.$$

When $\sigma \to \infty$,

$$\begin{aligned} \int F^2(\sigma x)\, dF(x) &= \int_{x>0} F^2(\sigma x)\, dF(x) + \int_{x<0} F^2(\sigma x)\, dF(x) \\ &\longrightarrow \frac{1}{2} + 0 = \frac{1}{2} \end{aligned}$$

Since $p_1 = 1/2$, plugging into the formula for $\lim_\sigma \alpha_{\text{true}}$ we get the desired result,

$$\lim_\sigma \alpha_{\text{true}} = \begin{cases} 1 - \Phi\left(\frac{z_\alpha}{\sqrt{3(1-\lambda)}}\right), & \sigma \to 0 \\ 1 - \Phi\left(\frac{z_\alpha}{\sqrt{3\lambda}}\right), & \sigma \to \infty \end{cases}$$

25.7 Exercises

Exercise 25.1. Give a rigorous proof that the two sample t statistic converges to $N(0, 1)$ in distribution if the variances are equal and finite.

Exercise 25.2. Simulate a sample of size $m = n = 25$ from the $N(0, 1)$ and the $N(0, 10)$ distributions, and compute the MLEs of the common mean and the two variances.

Exercise 25.3. * Give a necessary and sufficient condition that the cubic equation for finding the MLE of the common mean of two normal populations has one real root.

Exercise 25.4. * For each of the following cases, simulate the random degree of freedom of Welch's test : $m = n = 20, \sigma_1 = \sigma_2 = 1; m = 10, n = 50, \sigma_1 = \sigma_2 = 1; m = n = 20, \sigma_1 = 3, \sigma_2 = 1$.

Exercise 25.5. Compare by a simulation the power function of Welch's test with the two sample t test when the populations are normal with variances $1, 4$ and the sample sizes are $m = n = 10, 20$. Use a small grid for the values of the means.

Exercise 25.6. * Derive an expression for what Welch's degree of freedom would have been if he had tried to match a percentile, instead of the second moment.

Exercise 25.7. Use the recursion relation given in text to analytically write the distribution of the Mann-Whitney statistic when $m = n = 3$, assuming that the null is true.

Exercise 25.8. * Give a rigorous proof that the Mann-Whitney test is consistent under the condition stated in text.

Exercise 25.9. * Analytically find the mean and the variance of the Mann-Whitney statistic under a normal shift model.

Exercise 25.10. For the normal location-scale model, approximate the true type I error rate of Hettmansperger's conservative test and investigate when it starts to diverge from the nominal value .05 with $m = n = 20, 30, 50$.

Exercise 25.11. * What is the limiting type I error of the Mann-Whitney test when the null density is a normal and the alternative is a uniform ? Use the general expression given in the text.

Exercise 25.12. * What is the limiting type I error of the Mann-Whitney test when the null density is a normal and the alternative is an exponential ?

25.8 References

Babu,G.J. and Padmanabhan,A.R.(2002). Resampling methods for the nonparametric Behrens-Fisher problem,Sankhya, Special issue in memory of D.Basu,64,3,I,678-692.

Best,D.J. and Rayner,J.C.(1987). Welch's approximate solution to the Behrens-Fisher problem,Technometrics,29,2,205-210.

Brunner,E. and Munzel,U.(2000). The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation,Biom.J,1,17-25.

Fligner,M.A. and Policello,G.E.(1981). Robust rank procedures for the Behrens-Fisher problem,JASA,1976,162-168.

Ghosh,M. and Kim,Y.H.(2001). The Behrens-Fisher problem revisited: A Bayes-Frequentist synthesis,Canad.J.Statist.,29,1,5-17.

Hajek,J. and Sidak,Z.(1967). Theory of Rank Tests,Academic Press,New York.

Hettmansperger,T.(1984). Statistical Inference Based on Ranks,John Wiley,New York.

Hettmansperger,T.(1973). A large sample conservative test for location with unknown scale parameters,JASA,68,466-468.

Johnson,R.A. and Weerahandi,S.(1988). A Bayesian solution to the multivariate Behrens-Fisher problem,JASA,83,145-149.

Lehmann,E.L.(1986).Testing Statistical Hypotheses,Second Ed.,John Wiley,New York.

Lehmann,E.L. and Romano,J.(2005). Testing Statistical Hypotheses,Third Ed., Springer,New York.

Linnik,J.V.(1963). On the Behrens-Fisher problem,Bull.Inst.Internat.Statist.,40,833-841.

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

Milton,R.C.(1964). An extended table of critical values of the Mann-Whitney (Wilcoxon) two-sample statistic,JASA,59,925-934.

Pfanzagl,J.(1974).On the Behrens-Fisher problem,Biometrika,61,39-47.

Randles,R.H. and Wolfe,D.A.(1979). Introduction to the Theory of Nonparametric Statistics,John Wiley,New York.

Scheffe,H.(1970). Practical solutions of the Behrens-Fisher problem,JASA,65,1501-1508.

Serfling,R.(1980). Approximation Theorems of Mathematical Statistics,John Wiley,New York.

Wang,Y.Y.(1971). Probabilities of the type I errors of the Welch test for the Behrens-Fisher problem,JASA,66,605-608.

Welch,B.(1949). Further notes on Mrs.Aspin's tables,Biometrika,36,243-246.

26 Goodness of Fit

Suppose $X_1, \ldots, X_n$ are iid observations from a distribution F on an Euclidean space, say $\mathbb{R}$. We would discuss two types of goodness of fit problems: i) test $H_0 : F = F_0$, a completely specified distribution; ii) $F \in \mathcal{F}$, where $\mathcal{F}$ is a suitable family of distributions, possibly indexed by some finite dimensional parameter. Problem i) would be called the *simple goodness of fit problem*, and problem ii) the *composite goodness of fit problem*, or synonymously, *goodness of fit with estimated parameters*. It is the composite problem which is of greater interest in practice, although the simple problem can potentially arise in some situations. For example, one may have a hunch that F is uniform on some interval $[a, b]$ or that F is Bernoulli with parameter $\frac{1}{2}$. The simple goodness of fit problem has generated a vast amount of literature which has had its positive impact on the composite problem. So the methodologies and the theory for the simple case are worth looking at. We start with the simple goodness of fit problem and test statistics that use the empirical CDF $F_n(x)$ (EDF). Of the enormous literature on goodness of fit, we recommend D'Agostino and Stephens(1986) for treatment of a variety of problems, including the simple null case, and Stephens(1993) as a very useful review, although primarily for the composite case, which is discussed in a later chapter. Stuart and Ord(1991) and Lehmann(1999) provide lucid presentation of some of the the principal goodness of fit techniques.

Theory and methodology of goodness of fit were revolutionized with the advances in empirical process theory and the theory of central limit theorems on Banach spaces. The influence of these developments in what seems, at first glance, to be abstract probability theory on the goodness of fit literature was two fold. First, scattered results with case specific proofs could be unified, with a very transparent understanding of what is really going on. Second, these developments led to development of new tests, because tools were now available to work out the asymptotic theory of the new test procedures. We recommend del Barrio et al. (2007) for a comprehensive overview of these modern aspects of goodness of fit. In fact, we will discuss some of it in this chapter.

26.1 Kolmogorov-Smirnov and Other Tests Based on F_n

We know that for large n, F_n is "close" to the true F. For example, by the Gilvenko-Cantelli Theorem, $\sup |F_n(x) - F(x)| \xrightarrow{\text{a.s}} 0$. So if $H_0 : F = F_0$ holds, then we

should be able to test H_0 by studying the deviation between F_n and F_0. Any choice of a discrepancy measure between F_n and F_0 would result in a test. The utility of the test would depend on whether one can work out the distribution theory of the test statistic. A collection of discrepancy measures that have been proposed are the following:

$$
\begin{aligned}
&D_n^+ = \sup_{-\infty<t<\infty}(F_n(t) - F_0(t)),\\
&D_n^- = \sup_{-\infty<t<\infty}(F_0(t) - F_n(t)) = -\inf_{-\infty<t<\infty}(F_n(t) - F_0(t)),\\
&D_n = \sup_{-\infty<t<\infty}|F_n(t) - F_0(t)| = \max(D_n^+, D_n^-),\\
&V_n = D_n^+ + D_n^-,\\
&C_n = \int (F_n(t) - F_0(t))^2 dF_0(t),\\
&A_n = \int \tfrac{(F_n(t)-F_0(t))^2}{F_0(t)(1-F_0(t))} dF_0(t),\\
&w_n = w_{n,k,g} = \int (F_n(t) - F_0(t))^k g(F_0(t)) dF_0(t),\\
&D_n(g) = \sup_{-\infty<t<\infty} \tfrac{|F_n(t)-F_0(t)|}{g(F_0(t))},
\end{aligned}
$$

where $g : [0,1] \to \mathbb{R}^+$ is some fixed function and $k \geq 1$ is a fixed positive integer. The tests corresponding to D_n, V_n, C_n, A_n are respectively known as the Kolmogorov-Smirnov, the Kuiper, the Cramér-von Mises and the Anderson-Darling test. The tests corresponding to w_n and $D_n(g)$ are usually referred to as weighted Cramér-von Mises and weighted Kolmogorov-Smirnov tests.

26.2 Computational Formulas

D_n, C_n and A_n are the most common among the test statistics listed above. It can be shown that D_n, C_n, A_n are equal to the following simple expressions. Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the order statistics of the sample and let $U_i = F_0(X_{(i)})$. Then, assuming F_0 is continuous,

$$
\begin{aligned}
&D_n = \max_{1\leq i\leq n} \max\left\{\tfrac{i}{n} - U_{(i)}, U_{(i)} - \tfrac{i-1}{n}\right\},\\
&C_n = \tfrac{1}{12n} + \sum_{i=1}^n \left(U_{(i)} - \tfrac{2i-1}{n}\right)^2,\\
&A_n = -n - \tfrac{1}{n}\left[\sum_{i=1}^n (2i-1)(\log U_{(i)} + \log(1 - U_{(n-i+1)}))\right].
\end{aligned}
$$

Remark: It is clear from these computational formulas that for every fixed n, the sampling distributions of D_n, C_n and A_n under F_0 do not depend on F_0, provided F_0 is continuous. Indeed, one can prove directly by making the quantile transformation $U = F_0(X)$ that all the test statistics listed above have sampling distributions (under H_0) independent of F_0, provided F_0 is continuous. For small n, the true sampling distributions can be worked out exactly by discrete enumeration.

The quantiles for some particular levels have been numerically worked out for a range of values of n . Accurate approximation to the 95th and the 99th percentiles of D_n for $n \geq 80$ are $\frac{1.358}{\sqrt{n}}$ and $\frac{1.628}{\sqrt{n}}$.

26.3 Some Heuristics

For an iid $U[0,1]$ sample $Z_1, \ldots, Z_n$, let $U_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{Z_i \leq t}$. Recall that we call $U_n(t)$ a *uniform empirical process*. Suppose F_0 is a fixed CDF on $\mathcal{R}$, and $X_1, \ldots, X_n$ are iid samples from F_0. Then, defining $Z_i = F_0(X_i)$, $Z_1, \ldots, Z_n$ are iid $U[0,1]$. Therefore,

$$\sup_t |F_n(t) - F_0(t)| = \sup_t \left|\frac{1}{n}\sum I_{X_i \leq t} - F_0(t)\right| = \sup_t \left|\frac{1}{n}\sum I_{F_0(X_i) \leq F_0(t)} - F_0(t)\right|$$
$$= \sup_t \left|\frac{1}{n}\sum I_{Z_i \leq F_0(t)} - F_0(t)\right| \stackrel{\mathcal{L}}{=} \sup_t |U_n(F_0(t)) - F_0(t)| = \sup_{0 \leq t \leq 1} |U_n(t) - t| .$$

and therefore for every n, $\sqrt{n}\sup_t |F_n(t) - F_0(t)| \stackrel{\mathcal{L}}{=} \sqrt{n}\sup_{0 \leq t \leq 1} |U_n(t) - t|$, under F_0. So for every n, D_n has the same distribution as $\sqrt{n}\sup_{0 \leq t \leq 1} |U_n(t) - t|$. Define $X_n(t) = \sqrt{n}(U_n(t) - t), 0 \leq t \leq 1$. Recall from chapter 12 that $X_n(0) = X_n(1) = 0$. and $X_n(t)$ converges to a Gaussian process, $B(t)$, with $E(B(t)) = 0, \forall t$ and $Cov(B(s), B(t)) = s \wedge t - st, 0 \leq s, t \leq 1$, the *Brownian bridge* on $[0,1]$. By the invariance principle, the distribution of $D_n = \sqrt{n}\sup_{0 \leq t \leq 1} |U_n(t) - t|$ converges to the distribution of $\sup_{0 \leq t \leq 1} |B(t)|$. We have seen in chapter 12 that a rigorous development requires the use of weak convergence theory on metric spaces.

26.4 Asymptotic Null Distributions of D_n, C_n, A_n and V_n

Asymptotic theory of EDF based tests is now commonly handled by using Empirical process techniques and weak convergence theory on metric spaces. We recommend Shorack and Wellner(1986),Pollard(1989),Martynov(1992), Billingsley(1999) and del Barrio et al. (2007), apart from chapter 12 in this text, for details on techniques, statistical aspects, and concrete applications. The following fundamental results follow as consequences of the invariance principle for empirical processes, which we treated in chapter 12.

Theorem 26.1. *Let $X_1, X_2, \ldots \stackrel{iid}{\sim} F_0$, and let $D_n = \sup_t |F_n(t) - F_0(t)|$ and $C_n = \int_{-\infty}^{\infty}(F_n(t) - F_0(t))^2 dF_0(t)$. Then, assuming that F_0 is continuous,*

$$\sqrt{n}D_n \stackrel{\mathcal{L}}{\Rightarrow} \sup_{0 \leq t \leq 1} |B(t)|;$$

$$nC_n \stackrel{\mathcal{L}}{\Rightarrow} \int_0^1 B^2(t)dt,$$

$$nA_n \stackrel{\mathcal{L}}{\Rightarrow} \int_0^1 \frac{B^2(t)}{t(1-t)}dt,$$

$$P_{F_0}(\sqrt{n}D_n^+ \leq \lambda_2, \sqrt{n}D_n^- \leq \lambda_1) \to P(-\lambda_1 \leq \inf_{0\leq t\leq 1} B(t) \leq \sup_{0\leq t\leq 1} B(t) \leq \lambda_2),$$

$$\sqrt{n}V_n \stackrel{\mathcal{L}}{\Rightarrow} \sup_{0\leq t\leq 1} B(t) - \inf_{0\leq t\leq 1} B(t).$$

Now, the question reduces to whether one can find the distributions of these four functionals of $B(.)$. Fortunately, the answer is affirmative. The distributions of the first and the fourth functional, i.e., the distributions of the supremum of the absolute value and of the range, can be found by applications of the *reflection principle* (see chapter 12). The other two statistics are quadratic functionals of $B(.)$, and their distributions can be found by using the *Karhunen-Loéve expansion* (see chapter 12) of $B(t)$, and then writing the integrals in these two functionals as a linear combination of independent chi-square random variables. Since the characteristic function of a chi-square distribution is known, one can also write the characteristic function of the Brownian quadratic functional itself. Rather remarkably, a Fourier inversion can be done, and one can arrive at closed form expressions for the CDFs, which are the CDFs of the limiting distributions we want. See del Barrio et al. (2007) for the technical details. We record below some of these closed form expressions for the limiting CDFs.

Corollary 26.1.

$$\lim_n P_{F_0}(\sqrt{n}D_n \leq \lambda) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2},$$

$$\lim_{n\to\infty} P_{F_0}(nC_n > x) = \frac{1}{\pi}\sum_{j=1}^{\infty}(-1)^{j+1}\int_{(2j-1)^2\pi^2}^{4j^2\pi^2}\sqrt{\frac{-\sqrt{y}}{sin(\sqrt{y})}}\frac{e^{-\frac{xy}{2}}}{y}dy,$$

$\lim_{n\to\infty} P_{F_0}(\sqrt{n}D_n^+ \leq \lambda_2, \sqrt{n}D_n^- \leq \lambda_1)$
$= 1 - \sum_{k=1}^{\infty}\left\{e^{-2[k\lambda_2+(k-1)\lambda_1]^2} + e^{-2[(k-1)\lambda_2+k\lambda_1]^2} - 2e^{-2k^2(\lambda_1+\lambda_2)^2}\right\}.$

Remark: The CDF of the limiting distribution of $\sqrt{n}V_n$ is the CDF of the sum in the joint CDF provided in the last part above. An expression for the CDF of the limiting distribution of nA_n can also be found on using the fact that it is the CDF of the infinite linear combination $\sum_{j=1}^{\infty}\frac{Y_j}{j(j+1)}$, Y_j being iid chi-squares with one degree of freedom.

26.5 Consistency and Distributions under Alternative

The tests introduced above based on the empirical CDF F_n all have the pleasant property that they are consistent against any alternative $F \neq F_0$. For example, the Kolomogorov-Smirnov statistic D_n has the property that $P_F(\sqrt{n}D_n > G_n^{-1}(1-\alpha)) \to 1, \forall F \neq F_0$, where $G_n^{-1}(1-\alpha)$ is the $(1-\alpha)$th quantile of the distribution of $\sqrt{n}D_n$ under F_0. To explain heuristically why this should be the case, consider a CDF $F_1 \neq F_0$, so that there exists a such that $F_1(a) \neq F_0(a)$. Let us suppose that $F_1(a) > F_0(a)$. First note that $G_n^{-1}(1-\alpha) \to \lambda$, where λ satisfies $P(\sup_{0\leq t\leq 1}|B(t)| \leq \lambda) = 1-\alpha$. So

$$
\begin{aligned}
&P_{F_1}\left(\sqrt{n}D_n > G_n^{-1}(1-\alpha)\right) \\
&= P_{F_1}\left(\sup_t |\sqrt{n}(F_n(t) - F_0(t))| > G_n^{-1}(1-\alpha)\right) \\
&= P_{F_1}\left(\sup_t |\sqrt{n}(F_n(t) - F_1(t)) + \sqrt{n}(F_1(t) - F_0(t))| > G_n^{-1}(1-\alpha)\right) \\
&\geq P_{F_1}\left(|\sqrt{n}(F_n(a) - F_1(a)) + \sqrt{n}(F_1(a) - F_0(a))| > G_n^{-1}(1-\alpha)\right) \\
&\to 1
\end{aligned}
$$

as $n \to \infty$, since $\sqrt{n}(F_n(a) - F_1(a)) = O_p(1)$ under F_1, $\sqrt{n}(F_1(a) - F_0(a)) \to \infty$ and as stated above, $G_n^{-1}(1-\alpha) = O(1)$.

Remark: The same argument establishes the consistency of the other EDF (empirical distribution function) based tests against all alternatives. In contrast, we will later see that Chi-square goodness of fit tests cannot be consistent against all alternatives.

The invariance principle argument that we used to derive the limit distributions under H_0 also produce the limit distributions under F, a specified alternative. The limit distributions are still the distributions of appropriate functionals of suitable Gaussian processes (see Raghavachari (1973)). First we need some notation. Let F be a specified CDF different from F_0. Without loss of generality, we assume $F(0) = 0$ and $F(1) = 1$ and $F_0(t) = t$. Let

$$
\begin{aligned}
&\alpha = \sup_{0\leq t\leq 1}|F(t) - F_0(t)| = \sup_{0\leq t\leq 1}|F(t) - t|, \\
&\alpha^+ = \sup_{0\leq t\leq 1}(F(t) - F_0(t)) = \sup_{0\leq t\leq 1}(F(t) - t), \\
&\alpha^- = \inf_{0\leq t\leq 1}(F(t) - F_0(t)) = \inf_{0\leq t\leq 1}(F(t) - t), \\
&K_1 = \{0 \leq t \leq 1 : F(t) - t = \alpha\}, \\
&K_2 = \{0 \leq t \leq 1 : t - F(t) = \alpha\}, \\
&K^+ = \{0 \leq t \leq 1 : F(t) - t = \alpha^+\}, \\
&K^- = \{0 \leq t \leq 1 : F(t) - t = \alpha^-\}.
\end{aligned}
$$

Let also W_F denote a Gaussian process on $[0,1]$ with $W_F(0) = 0, E(W_F(t)) = 0$, and $Cov(W_F(s), W_F(t)) = F(s) \wedge F(t) - F(s)F(t), 0 \le s \le t \le 1$.

Theorem 26.2. $P_F(\sqrt{n}(D_n - \alpha) \le \lambda) \to P\left(\sup_{t \in K_1} W_F(t) \le \lambda, \inf_{t \in K_2} W_F(t) \ge -\lambda\right)$.

Remark: This result also gives a proof of the consistency of the test based on D_n. For given $0 < \gamma < 1$, $P_F(\sqrt{n}D_n < G_n^{-1}(1-\gamma)) = P_F(\sqrt{n}(D_n - \alpha) < G_n^{-1}(1-\gamma) - \alpha\sqrt{n}) \to 0$ from the above theorem as $n \to \infty$ since $G_n^{-1}(1-\gamma) = O(1)$ and $\alpha > 0$.

One can likewise find the limiting distributions of the other EDF based statistics, e.g., D_n^+ and D_n^-, under an alternative. For example, $P_F(\sqrt{n}(D_n^+ - \alpha^+) \le \lambda) \to P(\sup_{t \in K^+} W_F(t) \le \lambda)$. As regards the Kuiper statistic V_n, $P_F(\sqrt{n}(V_n - (\alpha^+ + \alpha^-)) \le \lambda \to P(\sup_{t \in K^+} W_F(t) - \inf_{t \in K^-} W_F(t) \le \lambda)$.

26.6 Finite Sample Distributions and Other EDF Based Tests

Kolmogorov himself studied the problem of the finite sample distribution of D_n under H_0 (Kolomogorov (1933)). He gave recurrence relations for finding the pmf of D_n. Wald and Wolfowitz (1940, 1941) gave exact formulae easy to use for small n. Since then the exact percentiles and exact CDFs have been numerically evaluated and extensively tabulated. For the reader's convenience, we report a short table of exact percentiles of D_n for some selected values of n.

n	95th Percentile	99th Percentile
20	.294	.352
21	.287	.344
22	.281	.337
23	.275	.330
24	.269	.323
25	.264	.317
26	.259	.311
27	.254	.305
28	.250	.300
29	.246	.295
30	.242	.290
35	.224	.269
40	.210	.252
> 40	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

26.7 Some Important Inequalities

Many inequalities on order statistics, extremes, and spacings, and their moments have been derived over the years. We collect a number of key such inequalities for purposes of reference.

Smirnov (1941) found the exact distribution of the one sided statistic D_n^+. Indeed, he found the limiting distribution of $\sqrt{n}D_n^+$ under the null by taking the exact distribution for given n and then by finding the pointwise limit. Smirnov's formula for given n is

$$P_{H_0}(D_n^+ > \epsilon) = (1-\epsilon)^n + \epsilon \sum_{j=1}^{[n(1-\epsilon)]} \binom{j}{n} (1 - \epsilon - \frac{j}{n})^{n-j} (\epsilon + \frac{j}{n})^{j-1}.$$

By symmetry, therefore, one also knows the exact distribution of D_n^- for any given n. Weighted versions of D_n, C_n and A_n are also sometimes used. In particular, the weighted Kolmogorov-Smirnov statistic is $D_n(g) = \sup_{-\infty<x<\infty} \frac{|F_n(t)-F_0(t)|}{g(F_0(t))}$, and the weighted Anderson-Darling statistic is $w_n(g) = \int \frac{(F_n(t)-F_0(t))^2}{g(F_0(t))} dF_0(t)$, for some suitable function g. For specific types of alternatives, the weighted versions provide greater power than the original unweighted versions, if the weighting function g is properly chosen. It is *not true* that the weighted versions converge in law to what would seem to be the obvious limit for arbitrary g. In fact, the question of weak convergence of the weighted versions is surprisingly delicate. Here is the precise theorem; see del Barrio et al. (2007) for a proof.

Theorem 26.3. (a) Let g be a strictly positive function on $(0,1)$, nondecreasing in a neighborhood of $t = 0$, and nonincreasing in a neighborhood of $t = 1$. Assume that for some $c > 0, \int_0^1 \frac{1}{t(1-t)} e^{-c\frac{g^2(t)}{t(1-t)}} dt < \infty$. Then, $\sqrt{n}D_n(g) \stackrel{\mathcal{L}}{\Rightarrow} \sup_{0<t<1} \frac{|B(t)|}{g(t)}$, where $B(t)$ is a Brownian bridge on $[0,1]$.

(b) Let g be a strictly positive function on $(0,1)$. Assume that $\int_0^1 \frac{t(1-t)}{g(t)} dt < \infty$. Then, $nw_n(g) \stackrel{\mathcal{L}}{\Rightarrow} \int_0^1 \frac{B^2(t)}{g(t)} dt$, where $B(t)$ is a Brownian bridge on $[0,1]$.

Remark: See Csörgo et al. (1986) for part (a) and Araujo and Giné (1980) for part (b).

26.8 The Berk-Jones Procedure

Berk and Jones (1979) proposed an intuitively appealing method of testing the simple goodness of fit null hypothesis $F = F_0$ for some specified continuous F_0 in

the one dimensional iid situation. It is also based on the empirical CDF, and quite a bit of useful work has been done on finite sample distributions of the Berk-Jones test statistic. It has also led to subsequent developments of other tests for the simple goodness of fit problem, as generalizations of the Berk-Jones idea. On the other hand, there are some unusual aspects about the asymptotic behavior of the Berk-Jones test statistic, and the statistics corresponding to its generalizations. We discuss the Berk-Jones test in this section, and certain generalizations in the next section.

The Berk-Jones method is to transform the simple goodness of fit problem into a family of Binomial testing problems. More specifically, if the true underlying CDF is F, then for any given $x, nF_n(x) \sim Bin(n, F(x))$. Suppressing the x, and writing p for $F(x)$, p_0 for $F_0(x)$, for the given x, we want to test $p = p_0$. Since F_0 is specified, by the usual quantile transform method, we may assume that the observations take values in $[0, 1]$ and that F_0 is the CDF of the $U[0, 1]$ distribution. We can use a likelihood ratio test corresponding to a two-sided alternative to test this hypothesis. It will require maximization of the binomial likelihood function over all values of p, which corresponds to maximization over $F(x)$, with x being fixed, while F being an arbitrary CDF. The likelihood is maximized at $F(x) = F_n(x)$, resulting in the likelihood ratio statistic

$$\lambda_n(x) = \frac{F_n(x)^{nF_n(x)}(1 - F_n(x))^{n-nF_n(x)}}{F_0(x)^{nF_n(x)}(1 - F_0(x))^{n-nF_n(x)}}$$

$$= (\frac{F_n(x)}{F_0(x)})^{nF_n(x)}(\frac{1 - F_n(x)}{1 - F_0(x)})^{n-nF_n(x)}.$$

But, of course, the original problem is to test that $F(x) = F_0(x) \forall\, x$. So, it would make sense to take a supremum of the log-likelihood ratio statistics over x. The Berk-Jones statistic is

$$R_n = n^{-1} sup_{0 \le x \le 1} \log \lambda_n(x).$$

As always, the questions of interest are the asymptotic and fixed sample distributions of R_n under the null, and if possible, under suitable alternatives. We present some key available results on these questions below. The principal references are Berk and Jones (1979), Wellner and Koltchinskii (2003), and Jager and Wellner (2006).

To study the asymptotics of the Berk-Jones statistic, first, it is useful to draw a connection between it and the Kullback-Leibler distance between Bernoulli distributions. Let $K(p, \theta) = p(\log p - \log \theta) + (1 - p)(\log(1 - p) - \log(1 - \theta))$ be

the Kullback-Leibler distance between the Bernoulli distributions with parameters p and θ. Then, it is easily seen that $\log \lambda_n(x) = K(F_n(x), F_0(x))$, and hence, $R_n = sup_{0 \le x \le 1} K(F_n(x), F_0(x))$. Properties of the Kullback-Leibler distance and Empirical process theory are now brought together, in an entirely nontrivial way, to derive the limiting distribution of R_n under the null hypothesis. See Jager and Wellner (2006) for a proof of the next theorem.

Theorem 26.4. Let $c_n = 2 \log \log n + \frac{1}{2} \log \log \log n - \frac{1}{2} \log(4\pi), b_n = \sqrt{2 \log \log n}$. Under $H_0 : F = F_0, nR_n - \frac{c_n^2}{2b_n^2} \stackrel{\mathcal{L}}{\Rightarrow} V$, where V has the CDF $e^{-4e^{-x}}, -\infty < x < \infty$.

Approximations to finite sample percentiles of nR_n are presented in Owen (1995). Somewhat simpler, but almost as accurate, approximations for the 95th percentile of nR_n are $\frac{c_n^2}{2b_n^2} - \log(-.25 \log .95)$.

26.9 φ-Divergences and the Jager-Wellner Tests

The Kullback-Leibler connection to the Berk-Jones statistic is usefully exploited to produce a more general family of tests for the simple goodness of fit problem in Jager and Wellner (2006). We describe these tests and the asymptotic distribution theory below. Some remarks about the efficiency of these tests are made at the end of this section.

The generalizations are obtained by considering generalizations of the $K(p, \theta)$ function above. The $K(p, \theta)$ function arises from the Kullback-Leibler distance, as we explained. The more general functions are obtained from distances more general than the Kullback-Leibler distance. In the information theory literature, these more general distances are known as φ-divergences. Let P_1, P_2 be two probability measures on some space, absolutely continuous with respect to a common measure λ (such an λ always exists). Let p_1, p_2 be the densities of P_1, P_2 with respect to λ, and let $g = \frac{p_2}{p_1} I_{p_1 > 0}$. Given a nonnegative convex function φ on the nonnegative reals, let $K_\varphi(P_1, P_2) = E_{P_1}[\varphi(g)]$. These are known as φ-divergence measures between a pair of probability distributions. See Csizár(1963) for the apparently first introduction of these divergences. Divergence measures have also been used usefully in estimation, and particularly robust estimation; one reference for an overview is Basu et al. (1998).

Some examples of the φ-function in common use are:

$$\varphi(x) = (x-1) \log x; \varphi(x) = -\log x; \varphi(x) = (\sqrt{x}-1)^2; \varphi(x) = |x-1|; \varphi(x) = -x^{1-t}, 0 < t < 1.$$

For the purpose of writing tests for the goodness of fit problem, Jager and Wellner (2006) use the following one parameter family of φ-functions:

$$\varphi_s(x) = x - \log x - 1,\ s = 0;$$

$$\varphi_s(x) = x \log x - x + 1,\ s = 1;$$

$$\varphi_s(x) = \frac{1 - s + sx - x^s}{s(1-s)}, s \neq 0, 1.$$

These functions result in the corresponding divergence measures

$$K_s(p, \theta) = \theta \varphi_s(p/\theta) + (1-\theta)\varphi_s((1-p)/(1-\theta)).$$

Accordingly, one has the family of test statistics

$$S_n(s) = sup_{0 \leq x \leq 1} K_s(F_n(x), F_0(x)).$$

Or, instead of taking supremums, one can take averages and get the test statistics

$$T_n(s) = \int_0^1 K_s(F_n(x), F_0(x))dx.$$

It is interesting to note that S_n, T_n generalize well known tests for the simple goodness of fit problem; for example, in particular, $S_n(1)$ = The Berk-Jones statistic; $T_n(2)$ = The integral form of the Anderson-Darling statistic.

The central limit theorem under the null for the families of test statistics $S_n(s), T_n(s)$ is described in the following theorem; see Jager and Wellner (2006) for a proof.

Theorem 26.5. (a) Let b_n, c_n, V be as in the previous theorem. For $s \in [-1, 2], nS_n(s) - \frac{c_n^2}{2b_n^2} \stackrel{\mathcal{L}}{\Rightarrow} V$ under H_0, as $n \to \infty$;

(b) For $s \in (-\infty, 2], nT_n(s) \stackrel{\mathcal{L}}{\Rightarrow} \int_0^1 \frac{B^2(t)}{2t(1-t)}dt$ under H_0, as $n \to \infty$, where $B(t)$ is a Brownian Bridge on $[0, 1]$.

The question of comparison naturally arises. We now have a large family of possible tests, all for the simple goodness of fit problem. Which one should one use ? The natural comparison would be in terms of power. This can be done theoretically, or by large scale simulations. The theoretical study focuses on comparison of Bahadur slopes of these various statistics. However, this is considerably more subtle than one would first imagine. The problem is that sometimes, depending on what is the exact alternative, when one intuitively expects the sequence of statistics to have an obvious almost sure limit, in reality it converges in law to a nondegenerate

random variable. There is a boundary phenomenon going on. On one side of the boundary, there is an almost sure constant limit, while on the other side there is a nondegenerate weak limit. This would make comparison by Bahadur slopes essentially meaningless. However, some qualitative understanding of power comparison has been achieved; see Berk and Jones (1979), Groeneboom and Shorack (1981), and Jager and Wellner (2006). Simulations are available in Jager (2006). Berk-Jones type supremum statistics appear to come out well in these theoretical studies and the simulations, but perhaps with a truncated supremum, over $x \in [X_{(1)}, X_{(n)}]$.

26.10 The Two Sample Case

Suppose X_i, $i = 1, 2, \cdots, n$ are iid samples from some continuous CDF F_0 and Y_i, $i = 1, 2, \cdots, m$ are iid samples from some continuous CDF F, and all random variables are mutually independent. Without loss of generality, assume $F_0(t) = t$, $0 \le t \le 1$, and assume that F is a CDF on $[0, 1]$. Let F_n, G_m denote the empirical CDFs of the X_i's and the Y_i's respectively. Analogous to the one sample case, one can define two and one sided Kolmogorov-Smirnov and Kuiper test statistics

$$\begin{aligned} D_{m,n} &= \sup_{0\le t\le 1} |F_n - G_m|; \\ D^+_{m,n} &= \sup_{0\le t\le 1} (F_n - G_m); \\ D^-_{m,n} &= \sup_{0\le t\le 1} (G_m - F_n) = -\inf_{0\le t\le 1} (F_n - G_m); \\ V_{m,n} &= D^+_{m,n} + D^-_{m,n}. \end{aligned} \tag{26.1}$$

There is substantial literature on the two sample equality of distributions problem. In particular, see Kiefer(1959),Anderson(1962), and Hodges(1958). As in the one sample case, the multivariate problem is a lot harder, it being difficult to come up with distribution-free simple and intuitive tests, even in the continuous case. See Bickel(1968), and Weiss(1960) for some results.

The limiting distribution of the two-sided Kolmogorov-Smirnov (KS) statistic is as follows.

Theorem 26.6. Let $X_i, 1 \le i \le n \overset{iid}{\sim} F_0, Y_j, 1 \le j \le m \overset{iid}{\sim} F$, where F_0, F are

continuous CDFs. Consider testing $H_0 : F = F_0$. Then,

$$\lim_{m,n\to\infty} P_{H_0}(\sqrt{\frac{mn}{m+n}}D_{m,n} \le \lambda) = P(\sup_{0\le t\le 1}|B(t)| \le \lambda)$$
$$= 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2\lambda^2}$$

provided for some $0 < \gamma < 1$, $\frac{m}{m+n} \to \gamma$.

Remark: Notice that the limiting distribution of the two sample two sided K-S statistic under H_0 is the same as that of the one sample two sided K-S statistic. The reason is $\sqrt{\frac{mn}{m+n}}(F_n(t)-G_m(t)) = \sqrt{\frac{mn}{m+n}}(F_n(t)-t-(G_m(t)-t)) = \sqrt{\frac{mn}{m+n}}(F_n(t)-t)-\sqrt{\frac{mn}{m+n}}(G_m(t)-t) \stackrel{\mathcal{L}}{\Rightarrow} \sqrt{1-\gamma}B_1(t)-\sqrt{\gamma}B_2(t)$,where $B_1(.)$, $B_2(.)$ are two independent Brownian bridges. But $\sqrt{1-\gamma}B_1(t)-\sqrt{\gamma}B_2(t)$ is another Brownian bridge. Therefore, by our usual continuous mapping argument, $\sqrt{\frac{mn}{m+n}}D_{m,n} \stackrel{\mathcal{L}}{\Rightarrow} \sup_{0\le t\le 1}|B(t)|$. The asymptotic null distribution of the two-sample Kuiper statistic is also easily found and is stated next.

Theorem 26.7. Assume the same conditions as in the previous theorem. Let $B(t)$ be a standard Brownian bridge on [0,1]. Then

$$\lim_{m,n\to\infty} P_{F_0}(\sqrt{\frac{mn}{m+n}}V_{m,n} \le \lambda) = P(\sup_{0\le t\le 1}B(t) - \inf_{0\le t\le 1}B(t) \le \lambda)$$

Remark: Notice that again in the two sample case, the asymptotic null distribution of the Kuiper statistic is the same as that in the one sample case. The reason is the same as the explanation given above for the case of the K-S statistic.

The asymptotic distributions of $D_{m,n}$ and $V_{m,n}$ under an alternative are also known which we describe below.

Theorem 26.8. Suppose $\frac{m}{m+n} \to \gamma$, $0 < \gamma < 1$. Then

$$\lim_{m,n\to\infty} P_F(\sqrt{\frac{mn}{m+n}}(D_{m,n}-\alpha) \le \lambda)$$
$$= P(\sup_{t\in K_1}(\sqrt{\gamma}W_F(t) - \sqrt{1-\gamma}B(t)) \le \lambda, \inf_{t\in K_2}(\sqrt{1-\gamma}W_F(t) - \sqrt{\gamma}B(t)) \ge -\lambda),$$

where $\alpha, K_1, K_2, W_F(t)$ are as in section 26.5, and $W_F(t)$ and $B(t)$ are independent.

Theorem 26.9. Suppose $\frac{m}{m+n} \to \gamma$, $0 < \gamma < 1$. Then

$$\lim_{m,n\to\infty} P_F(\sqrt{\frac{mn}{m+n}}(V_{m,n}-\alpha^+ + \alpha^-) \le \lambda)$$
$$= P(\sup_{t\in K^+}(\sqrt{\gamma}W_F(t) - \sqrt{1-\gamma}B(t)) - \inf_{t\in K^-}(\sqrt{\gamma}W_F(t) - \sqrt{1-\gamma}B(t)) \le \lambda),$$

where $K^{\pm}$ are as in section 26.5, and $W_F(t)$ and $B(t)$ are again independent.

Remark: See Raghavachari (1973) for details and proofs of the last two theorems in this section. In examples, one of the sets K_1 and K_2, and one of the sets K^+, K^- may be empty and the other one a singleton set. This will facilitate analytical calculations of the asymptotic CDFs in the above two theorems. In general, analytical calculation may be cumbersome. For discrete cases, it is wrong to use the statistics; however if our "wrong" p-value is very small, the true p-value is even smaller than the wrong one, and so it is probably safe to reject in such a case.

26.11 Tests for Normality

Because of its obvious practical importance, there has been a substantial amount of work on devising tests for normality. Thus, suppose $X_1, X_2, \ldots, X_n$ are iid observations from a CDF F on the Real line. The problem is to test that F belongs to the family of normal distributions. Although we will discuss modifications of the Kolmogorov-Smirnov and the chi-square tests for testing this hypothesis in chapters 27 and 28, we will describe some other fairly popular tests for normality here, due to the practical interest in the problem.

The Q-Q Plot The Q-Q plot is a hugely popular graphical method for testing for normality. It appears to have been invented by the research group at the Bell Labs. See Gnanadesikan(1997). The simple rationale is that the quantile function of a general normal distribution satisfies $Q(\alpha) = \mu + \sigma z_\alpha$, with obvious notation. So a plot of $Q(\alpha)$ against z_α would be linear, with an intercept of μ and a slope σ. With given data, the order statistics of the sample are plotted against the standard normal percentiles. This plot should be roughly linear if the data are truly normal. A visual assessment of linearity is then made. To avoid singularities, the plot consists of the pairs $(\Phi^{-1}(\frac{i}{n+1}), X_{(i)}), i = 1, 2, \ldots, n$ (or some such modification of $\Phi^{-1}(\frac{i}{n+1})$).

At the hands of a skilled analyst, the Q-Q plot can provide useful information about the nature of the true CDF from which the observations are coming. For example, it can give information about the tail and skewness of the distribution, and about its unimodality. See Marden(1998,2004). Brown, DasGupta,Marden and Politis(2004) show that the basic assessment of linearity, however, is fundamentally unreliable, as most types of data would produce remarkably linear Q-Q plots, except

for a detour only in the extreme tails. If this detour in the tails is brushed aside as unimportant, then the Q-Q plot becomes a worthless tool. They give the following theorem.

Theorem 26.10. Let r_n denote the correlation coefficient computed from the bivariate pairs $(\Phi^{-1}(\frac{i}{n+1}), X_{(i)}), i = 1, 2, \ldots, n$.Let F denote the true CDF, with finite variance σ^2. Then

$$\lim_{n\to\infty} r_n = \rho_F = \frac{\int_0^1 F^{-1}(\alpha)\Phi^{-1}(\alpha)d\alpha}{\sigma}, a.s.$$

The a.s. limit is very close to 1 for all kinds of distributions, as can be seen in the Table below.

F	$\lim r_n$
Normal	1
Uniform	.9772
Double Exponential	.9811
t_3	.9008
t_5	.9832
χ^2_5	.9577
Exponential	.9032
Tukey	.9706
Logistic	.9663

They also show that if just 5% of the points from each tail are deleted, then the corresponding a.s. limits are virtually equal to 1 for even skewed data such as χ^2_5. Having said that, a formal test which uses essentially the correlation coefficient r_n above is an omnibus consistent test for normality. This test is described next.

The Shapiro-Francia-Wilk Test The Shapiro-Francia test is a slight modification of the wildly popular Shapiro-Wilk test, but is easier to describe; see Shapiro and Wilk(1965) and de Wet and Ventner(1972). It rejects the hypothesis of normality when r_n is small. If the true CDF is nonnormal, then r_n, on centering, and norming by $\sqrt{n}$ has a limiting normal distribution. If the true CDF is normal, $r_n \to 1$ faster than $\sqrt{n}$. Against all nonnormal alternatives with a finite variance, the test is consistent. See Sarkadi(1985).

Theorem 26.11. (de Wet and Ventner, 1972; Sarkadi, 1985) a. If F does not belong to the family of normal distributions, then $\sqrt{n}(r_n - \rho_F) \stackrel{\mathcal{L}}{\Rightarrow} N(0, \tau_F^2)$, where ρ_F is as above, and τ_F is a suitable functional of F.

b. If F belongs to the family of normal distributions, then $n(1-r_n) \stackrel{\mathcal{L}}{\Rightarrow} \sum_{i=1}^{\infty} c_i(W_i - 1)$, where W_i are iid χ_1^2 variables, and c_i are suitable constants.

c. If F does not belong to the family of normal distributions, then $\lim_{n\to\infty} P(r_n < 1 - \frac{c}{n}) = 1$, for any $c > 0$.

Among many other tests for normality available in the literature, tests based on the skewness and the kurtosis of the sample are quite popular; see the Exercises. There is no clear cut comparison between these tests without focusing on the type of alternative for which good power is desired.

26.12 Exercises

Exercise 26.1. * For the data set $-1.88, -1.71, -1.40, -.95, .22, 1.18, 1.25, 1.41, 1.70, 1.97$, calculate the values of each of $D_n^+, D_n^-, D_n, C_n, A_n$, and V_n, when the null distribution is $F_0 = N(0,1); F_0 = DoubleExp(0,1); F_0 = C(0,1)$. Do the computed values make sense intuitively ?

Exercise 26.2. * For $n = 15, 25, 50, 100$, simulate $U[0,1]$ data, and then plot the normalized uniform Empirical process $\sqrt{n}(U_n(t) - t)$. Comment on the behavior of your simulated trajectory.

Exercise 26.3. * By repeatedly simulating $U[0,1]$ data for $n = 50, 80, 125$, compare the simulated 95th percentile of D_n with the approximation $\frac{1.358}{\sqrt{n}}$ quoted in text.

Exercise 26.4. Show that on $D[0,1]$, the functionals $x \to sup_t|x(t)|, x \to \int x^2(t)dt$ are continuous with respect to supnorm (see Chapter 12).

Exercise 26.5. * Find the mean and the variance of the asymptotic null distribution of the two sided K-S statistic.

Exercise 26.6. * By using Corollary 26.1, find the mean and the variance of the asymptotic null distribution of nC_n.

Exercise 26.7. * By careful computing, plot and superimpose the CDFs of the asymptotic null distributions of $\sqrt{n}D_n$ and nC_n.

Exercise 26.8. * By careful computing, plot the density of the asymptotic null distribution of the Kuiper statistic.

Exercise 26.9. * Prove that each of the Kuiper, Cramér-von Mises and the Anderson-Darling tests is consistent against any alternative.

Exercise 26.10. * By using Theorem 26.4, approximate the power of the two sided K-S test for testing $H_0 : F = U[0,1]$ against a $Beta(\alpha, \alpha)$ alternative, with $\alpha = .5, 1.5, 2, 5$ and $n = 25, 50, 100$.

Exercise 26.11. * Consider the statistic N_n = Number of times F_n crosses F_0. Simulate the expected value of N_n under the null and compare to its known limiting value $\sqrt{\frac{\pi}{2}}$.

Exercise 26.12. * Simulate a sample of size $m = 20$ from the $N(0, 25)$ distribution and a sample of size $n = 10$ from the $C(0, 1)$ distribution. Test the hypothesis that the data come from the same distribution by using the two sample K-S and the two sample Kuiper statistic. Find the P-values.

Exercise 26.13. * Simulate a sample of size $n = 20$ from the $N(0, 25)$ distribution. Test the hypothesis that the data come from a standard Cauchy distribution by using the Berk-Jones statistic. Use the percentile approximation given in the text to approximately compute a P-value. Repeat the exercise for testing that the simulated data come from a $N(0, 100)$ distribution.

Exercise 26.14. * Simulate a sample of size $n = 20$ from the $N(0, 25)$ distribution and compute the values of the $S_n(s)$ and $T_n(s)$ statistics for $s = -1, 0, \frac{1}{2}, 1, 2$.

Exercise 26.15. * **Geary's 'a'** Given iid observations $X1, X_2, \ldots, X_n$ from a distribution with finite variance, let $a = \frac{\frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}|}{s}$.

a. Derive the asymptotic distribution of a in general. Using it, derive the asymptotic distribution of a when the underlying distribution is normal.

b. Hence suggest a test for normality based on Geary's a.

c. Find the *exact* mean and variance of a in finite samples under normality.

Exercise 26.16. Q-Q Plot for Exponentiality Let $X_1, X_2, \ldots, X_n$ be iid observations from a distribution on $(0, \infty)$ and let F_n be the empirical CDF. Let $t_{(i)}$ denote the order statistics of the sample, and let $Q(\alpha) = -\log(1 - \alpha)$ be the quantile function of the standard Exponential. Justify why a plot of the pairs $(t_{(i)}, Q(F_n(t_{(i)}) - .5n))$ can be used to test that $X1, X_2, \ldots, X_n$ are samples from some Exponential distribution.

Exercise 26.17. * **Test for Normality** Let $X_1, X_2, \ldots, X_n$ be iid observations from a distribution F on the Real line, and let b_1, b_2 denote the usual sample skewness and sample kurtosis coefficients, i.e., $b_1 = \frac{\frac{1}{n}\sum(X_i - \bar{X})^3}{s^3}$,and $b_2 = \frac{\frac{1}{n}\sum(X_i - \bar{X})^4}{s^4}$.

a. Show that if F is a normal distribution, then i) $\sqrt{n}b_1 \stackrel{\mathcal{L}}{\Rightarrow} N(0, 6)$, ii) $\sqrt{n}(b_2 - 3) \stackrel{\mathcal{L}}{\Rightarrow} N(0, 24)$.

b. Suggest a test for normality based on b_1; based on b_2.

c. Are these tests consistent against all alternatives, or only certain alternatives?

d. Can you suggest a test based jointly on (b_1, b_2) ?

26.13 References

Anderson,T.W.(1962).On the distribution of the two sample Cramér-von Mises criterion, Ann.Math.Statist.,33,1148-1159.

Araujo, A. and Giné, E. (1980). The Central Limit Theorem for Real and Banach Valued Random Variables, Wiley, New York.

Basu, A., Harris, I., Hjort, N. and Jones, M. (1998). Robust and efficient estimation by minimizing density power divergence, Biometrika, 85, 549-559.

Berk, R. and Jones, D. (1979). Goodness of fit statistics that dominate the Kolmogorov statistics, Z. Wahr. verw. Geb., 47, 47-59.

Bickel,P.J.(1968).A distribution free version of the Smirnov two sample test in the p-variate case,Ann.Math.Statist.,40,1-23.

Billingsley,P.(1999). Convergence of Probability Measures,Second Edition,John Wiley,New York.

Brown,L.,DasGupta,A.,Marden,J., and Politis,D.(2004). Characterizations,sub and resampling,and goodness of fit,IMS Lecture Notes Monograph Ser 45,180-206, Institute of Mathematical Statistics,Beachwood,OH.

Csizár, I. (1963). In German, Magyar Tud. Akad. Mat. Kutató Int. Közl, 8, 85-108.

Csörgo, M., Csörgo, S., Horváth, L. and Mason, D. (1986). Weighted empirical and quantile process, Ann. Prob., 14, 31-85.

D'Agostino,R. and Stephens,M.(1986).Goodness of Fit Techniques,Marcel Dekker, New York.

de Wet,T. and Ventner,J.(1972).Asymptotic distributions of certain test criteria of normality,South Afr. Statist.Jour.,6,135-149.

del Barrio, E., Deheuvels, P. and van de Geer, S. (2007). Lectures on Empirical Processes, European Mathematical Society, Zurich.

Gnanadesikan,R.(1997).Methods for Statistical Data Analysis of Multivariate Observations,John Wiley,New York.

Groeneboom, P. and Shorack, G. (1981). Large deviations of goodness of fit statistics and linear combinations of order statistics, Ann. Prob., 9, 971-987.

Hodges,J.L.(1958).The significance probability of the Smirnov two sample test Ark.Mat., 3, 469-486.

Jager, L. (2006). Goodness of fit tests based on phi-divergences, Tech. Report, Univ. Washington.

Jager, L. and Wellner, J. (2006). Goodness of fit tests via phi-divergences, In Press.

Kiefer,J.(1959).k sample analogoues of the Kolmogorov-Smirnov and the Cramér-von Mises tests,Ann.Math.Statist.,30,420-447.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione, Giorn. Ist. Ital. Attuari. 4, 83-91.

Lehmann,E.L.(1999).Elements of Large Sample Theory,Springer,New York.

Marden,J.(1998). Bivariate qq and spider-web plots,Statist.Sinica,8,3,813-816.

Marden,J.(2004). Positions and QQ plots,Stat.Sc.,19,4,606-614.

Martynov,G.V.(1992).Statistical tests based on Empirical processes and related problems,Soviet J.Math,61,4,2195-2271.

Pollard,D.(1989).Asymptotics via Empirical Processes,Statist.Sc.,4,4,341-366.

Raghavachari, M. (1973). Limiting distributions of the Kolmogorov- Smirnov type statistics under the alternative, Ann. Stat., 1, 67-73.

Sarkadi,K.(1985).On the asymptotic behavior of the Shapiro-Wilk test,Proc. 7th Conf. Prob. Th.,Brasov,Romania,IVNU Science Press.

Shapiro,S.S. and Wilk,M.B.(1965).An analysis of variance test for normality,Biometrika, 52, 591-611.

Shorack,G. and Wellner,J.(1986). Empirical Processes, with Applications to Statistics,John Wiley,New York.

Smirnov, N. (1941). Approximate laws of distribution of random variables from empirical data (in Russian), Uspekhi Mat. Nauk., 10, 179-206.

Stephens,M.(1993).Aspects of Goodness of Fit,Statistical Sciences and Data Analysis,VSP,Utrecht.

Stuart,A. and Ord,K.(1991).Kendall's Advanced Theory of Statistics,Vol II,Clarendon Press,New York.

Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population, Ann. Math. Statist., 11, 147-162.

Wald, A. and Wolfowitz, J. (1941). Note on confidence limits for continuous distribution functions, Ann. Math. Statist., 12, 118-119.

Weiss,L.(1960).Two sample tests for multivariate distributions,Ann.Math.Statist., 31,159-164.

Wellner, J. and Koltchinskii, V. (2003). A note on the asymptotic distribution of the Berk-Jones type statistics under the null distribution, In High Dimensional Probability III, 55, Progress in Prob., Birkhäuser, Basel, 321-332.

27 Chi-square Tests for Goodness of Fit

Well known competitors to EDF based statistics are Chi-square tests. They discretize the null distribution in some way, and assess the agreement of observed counts to the postulated counts. So there is obviously some loss of information and hence a loss in power. But they are versatile. Unlike EDF based tests, a Chi-square test can be used for continuous as well as discrete data, and in one dimension as well as many dimensions. Thus a loss of information is being exchanged for versatility of the principle and the ease of computation.

27.1 The Pearson χ^2 Test

Suppose $X_1, \ldots, X_n$ are IID observations from some distribution F in an Euclidean space and suppose we want to test $H_0 : F = F_0$, F_0 being a completely specified distribution. Let S be the support of F_0 and for some given $k \geq 1$, $A_{k,i}$, $i = 1, 2, \cdots, k$ form a partition of S. Let $p_{0,i} = P_{F_0}(A_{k,i})$, and $n_i = \#\{j : x_j \in A_{k,i}\}$=the observed frequency of the partition set $A_{k,i}$. Therefore, under H_0, $E(n_i) = np_{0,i}$. Karl Pearson suggested that as a measure of discrepancy between the observed sample and the null hypothesis, one compares $(n_1, ..., n_k)$ with $(np_{0i}, ..., np_{0k})$. The Pearson Chi-square statistic is defined as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{0i})^2}{np_{0i}}$$

For fixed n, certainly, χ^2 is not distributed as a Chi-square, for it is just a quadratic form in a multinomial random vector. However, the asymptotic distribution of χ^2 is χ^2_{k-1}, if H_0 holds; hence the name Pearson Chi-square for this test.
As hard as it is to believe, Pearson's chisquare test is actually more than a century old(Pearson(1900)). Cox(2000),Rao(2000) give well written accounts. Serfling(1980) and Ferguson(1996) contain theoretical developments. Greenwood and Nikulin(1996) is a masterly treatment. Modifications of Pearson's chisquare have been suggested; see, among others, Rao and Robson(1974).

27.2 Asymptotic Distribution of Pearson's Chi-square

Theorem 27.1. Suppose $X_1, X_2, \ldots, X_n$ are iid observations from some distribution F in a finite dimensional Euclidean space. Consider testing $H_0 : F = F_0$ (specified). Let χ^2 be the Pearson χ^2 statistic defined above. Then $\chi^2 \stackrel{\mathcal{L}}{\Rightarrow} \chi^2_{k-1}$

under H_0.

Proof: It is easy to see why the asymptotic null distribution of χ^2 should be χ^2_{k-1}. Define $Y = (Y_1, ..., Y_k) = (\frac{n_1 - np_{01}}{\sqrt{np_{01}}}, ..., \frac{n_k - np_{0k}}{\sqrt{np_{0k}}})$. By the multinomial CLT (see Chapter 1 exercises), $Y \stackrel{\mathcal{L}}{\Rightarrow} N_k(0, \Sigma)$, where $\Sigma = I - \mu\mu^{'}$, where $\mu^{'} = (\sqrt{p_{01}}, ...\sqrt{p_{0k}})$, trace $(\Sigma) = k - 1$. The eigenvalues of Σ are 0 with multiplicity 1 and 1 with multiplicity $(k-1)$. Notice now that Pearson's $\chi^2 = Y^{'}Y$ and if Y$\backsim N_k(0, \Sigma)$ for any general Σ then $Y^{'}Y = X^{'}P^{'}PX = X^{'}X \stackrel{\mathcal{L}}{=} \sum_{i=1}^{k} \lambda_i w_i$, where $w_i \stackrel{iid}{\sim} \chi^2_1$, λ_i are the eigenvalues of Σ and $P^{'}\Sigma P$=diag$(\lambda_1, ..., \lambda_k)$ is the spectral decomposition of Σ. So $X \sim N_k(0,$ diag$(\lambda_1, ..., \lambda_k))$ and it follows that $X^{'}X = \sum_{i=1}^{k} X_i^2 \stackrel{\mathcal{L}}{=} \sum_{i=1}^{k} \lambda_i w_i$. For our Σ, $k-1$ of $\lambda_i^{'}$s are 1 and the remaining one is zero. Since a sum of independent Chi-squares is again a Chi-square, it follows from the multinomial CLT that $\chi^2 \stackrel{\mathcal{L}}{\Rightarrow} \chi^2_{k-1}$ under H_0.

Remark: The so called Freeman-Tukey statistic is a kind of symmetrization of Pearson χ^2 with respect to the vector of observed and expected frequencies. It is defined as $FT = 4\sum_{i=1}^{k}(\sqrt{n_i} - \sqrt{np_{0i}})^2$ and it turns out that FT also converges to χ^2_{k-1} under H_0, which follows by an easy application of the delta theorem. The Freeman-Tukey statistic is sometimes preferred to Pearson's chi-square. See Stuart and Ord (1991) for some additional information.

27.3 Asymptotic Distribution Under Alternative and Consistency

Let F_1 be a distribution different from F_0 and let $p_{1i} = P_{F_1}(A_{k,i})$. Clearly, if by chance, $p_{1i} = p_{0i}$ $\forall$ i=1,..,k (which is certainly possible), then a test based on the empirical frequencies of $A_{k,i}$ cannot distinguish F_0 from F_1, even asymptotically. In such a case, the χ^2 test cannot be consistent against F_1. However, otherwise, it will be consistent as can be seen easily from the following result.

Theorem 27.2. $\frac{\chi^2}{n} \stackrel{P}{\Rightarrow} \sum_{i=1}^{k} \frac{(p_{1i} - p_{0i})^2}{p_{0i}}$ under F_1.

This is evident as $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{0i})^2}{np_{0i}} = n\sum_{i=1}^{k} \frac{(\frac{n_i}{n} - p_{0i})^2}{p_{0i}}$. But $(\frac{n_1}{n}, ..., \frac{n_k}{n}) \stackrel{P}{\Rightarrow} (p_{11}, ..., p_{1k})$ under F_1. Therefore, by the continuous mapping theorem $\frac{\chi^2}{n} \stackrel{P}{\Rightarrow} \sum_{i=1}^{k} \frac{(p_{1i} - p_{0i})^2}{p_{0i}}$.

Corollary 27.1. If $\sum_{i=1}^{k} \frac{(p_{1i} - p_{0i})^2}{p_{0i}} > 0$, then $\chi^2 \stackrel{\mathcal{P}}{\longrightarrow} \infty$ under F_1 and hence the χ^2 test is consistent against F_1.

Remark: Thus, for a fixed alternative F_1 such that the vector $(p_{11}, ..., p_{1k}) \neq (p_{01}, ..., p_{0k})$, Pearsons' χ^2 cannot have a nondegenerate limit distribution under

F_1. However, if the alternative is very close to the null, in the sense of being a Pitman alternative , there is a nondegenerate limit distribution. We have seen this phenomenon occur previously in other testing problems.

Theorem 27.3. Consider an alternative $F_1 = F_{1,n} = F_0 + \frac{1}{\sqrt{n}}G$, where the total mass of G is 0. Let $p_{1i} = p_{0i} + \frac{1}{\sqrt{n}}c_i$ where $c_i = \int_{A_{k,i}} dG$, $\sum_{i=1}^{k} c_i = 0$. Then $\chi^2 \stackrel{\mathcal{L}}{\Rightarrow} NC\chi^2(k-1, \delta^2)$ where $\delta^2 = \sum_{i=1}^{k} \frac{c_i^2}{p_{0i}}$.

Remark: This result can be used to approximate the power of the χ^2 test at a close by alternative by using the noncentral χ^2 CDF as an approximation to the exact CDF of χ^2 under the alternative.

27.4 Choice of k

A key practical question in the implementation of χ^2 tests is the choice of k and the actual partitioning sets $A_{k,i}$. Both are hard problems and despite huge literature on the topic, there are no clear cut solutions. Some major references on this hard problem are Mann and Wald(1942),Oosterhoff(1985), and Stuart and Ord(1991).
A common assumption in much of the theoretical works is to take some suitable value of k and use the partition sets $A_{k,i}$ which make $p_{0i} \equiv \frac{1}{k}$. In other words, the cells are equiprobable under H_0. Note that generally this will make the cells of unequal size (e.g, of unequal width if they are intervals). The problem then is to seek the optimum value of k. The crux of the problem in optimizing k is that a large k may or may not be a good choice, depending on the alternative. One can see this by simple moment calculations, and further comments on this are made below.

Theorem 27.4.

$$\begin{aligned}
(a)E_{H_0}(\chi^2) &= k-1; \\
(b)Var_{H_0}(\chi^2) &= \frac{1}{n}(2(n-1)(k-1) - k^2 + \sum_{i=1}^{k}\frac{1}{p_{0i}}); \\
(c)E_{F_1}(\chi^2) &= \sum_{i=1}^{k}\frac{p_{1i}(1-p_{1i})}{p_{0i}} + n(\sum_{i=1}^{k}\frac{p_{1i}^2}{p_{0i}} - 1).
\end{aligned}$$

Remark: See Sen and Singer (1993) or Serfling (1980) for simple derivations of the moments of Pearson's χ^2. The variance under the alternative has a somewhat messy expression. The formula for $\mathrm{Var}_{H_0}(\chi^2)$ indicates the problem one will have with many cells. If k is very large, then some value of p_{0i} would be small, making

$\sum_{i=1}^{k} \frac{1}{p_{0i}}$ a large number and $\text{Var}_{H_0}(\chi^2)$ quite a bit larger than $2(k-1)$. This would indicate that the χ^2_{k-1} approximation to the null distribution of χ^2 is not accurate. So even the size of the test may differ significantly from the nominal value if k is too large. Clearly, the choice of k is a rather subtle issue.

Example 27.1. Here are some values of the power of the Pearson χ^2 test, when $F_0 = N(0,1)$ and $F_1 =$ a Cauchy or another normal and when $F_0 = U[0,1]$ and $F_1 =$ a Beta distribution. The numbers in the Table are quite illuminating.

Table $(n = 50, \alpha = 0.05, p_{0i} = \frac{1}{k})$

		k			
F_0	F_1	4	6	8	15
$N(0,1)$	$C(0,\sigma), \sigma = \frac{1}{2}$	0.18	0.25	0.28	0.40
$N(0,1)$	$N(0,\sigma^2), \sigma = \frac{4}{3}$	0.32	0.32	0.30	0.24
$U[0,1]$	$Beta(\frac{2}{3}, \frac{2}{3})$	0.20	0.23	0.25	0.26
$U[0,1]$	$Beta(\frac{5}{3}, \frac{5}{3})$	0.39	0.34	0.32	0.24

For the case $F_0 = N(0,1)$, the power increases monotonically in k when the alternative is Cauchy, which is thick tailed, but actually deteriorates for the larger k, when the alternative is another normal, which is thin tailed. Similarly, when $F_0 = U[0,1]$, the power increases monotonically in k when F_1 is a U-shaped Beta distribution, but deteriorates for the larger k when F_1 is a unimodal Beta distribution. We shall later see that some general results can be given that justify such an empirical finding.
We now present some results on selecting the number of cells k.

27.5 Recommendation of Mann and Wald

Mann and Wald (1942) formulated the problem of selecting the value of k in a (somewhat complicated) paradigm and came out with an optimal rate of growth of k as $n \to \infty$. The formulation of Mann and Wald was along the following lines.

Fix a number $0 < \Delta < 1$. Let F_0 be the null distribution and F_1 a plausible alternative. Consider the class of alternatives $\mathcal{F} = \mathcal{F}_\Delta = \{F_1 : d_K(F_0, F_1) \geq \Delta\}$, where $d_K(F_0, F_1)$ is the Kolmogorov distance between F_0 and F_1. Let $\beta(F_1, n, k, \alpha) =$

$P_{F_1}(\chi^2 > \chi^2_{k-1}(\alpha))$. Mann and Wald (1942) consider $\inf_{F_1 \in \mathcal{F}_\Delta} \beta(F_1, n, k, \alpha)$ and suggest $k_n = k_n(\alpha, \Delta) = \operatorname{argmax}_k \inf_{F_1 \in \mathcal{F}_\Delta} \beta(F_1, n, k, \alpha)$ as the value of k. Actually, the criterion is a bit more complex than that; see Mann and Wald (1942) for the exact criterion. They prove that k_n grows at the rate $n^{\frac{2}{5}}$, i.e., $k_n \sim n^{\frac{2}{5}}$. Actually, they also produce a constant in this rate result. Later empirical experience has suggested that the theoretical constant is a bit too large.

A common practial recommendation influenced by the Mann-Wald result is $k = 2n^{\frac{2}{5}}$. The recommendation seems to produce values of k that agree well with practical choices of k. Here is a table.

Table

n	Integer nearest to $2n^{2/5}$
25	7
50	10
80	12
100	13

These values seem to be close to what common practice is. The important points are that k should be larger when n is large. But it is not recommended that one uses a very large value for k, and a choice in the range $5 - 15$ seems right.

27.6 Power at Local Alternatives and Choice of k

Suppose we wish to test that $X_1, X_2, \ldots, X_n$ are *i.i.d.* H, with density h. Thus the null density is h. For another density g, and $0 \leq \theta \leq 1$, consider alternatives

$$g_\theta = (1 - \theta)h + \theta g.$$

If $0 < \theta < 1$ is fixed, then the Pearson $\chi^2 \xrightarrow{\mathcal{P}} \infty$ under g_θ, as we saw previously (provided the cell probabilities are not the same under g and h). But if $\theta = \theta_n$, and θ_n converges to zero at the rate $\frac{1}{\sqrt{n}}$, then the Pearson χ^2 has a noncentral χ^2 limit distribution and the power under the alternative g_{θ_n} has a finite limit for any fixed k. The question is, if we let $k \to \infty$, then what happens to the power? If it converges to 1, letting k grow would be a good idea. If it converges to the level α,

then letting k grow arbitrarily would be a bad idea.

To describe the results, we first need some notation. We suppress k and n in the notation.

Let

$$\begin{aligned} p_{0i} &= \int_{A_{k,i}} h(x)dx \\ p_i &= \int_{A_{k,i}} g_{\theta_n}(x)dx \\ p_i^* &= \int_{A_{k,i}} g(x)dx \\ \Delta_k &= \sum_{i=1}^{k} \frac{(p_i^* - p_{0i})^2}{p_{0i}} \\ f &= \frac{g}{h} - 1 \end{aligned}$$

Then one has the following results (Kallenberg *et. al* (1985)).

Theorem 27.5. Suppose

i) $k = k(n) \to \infty$ such that $k = o(n)$,

ii) $\liminf_n \min_i (kp_{0i}) > 0$,

iii) $\lim_n n\theta_n^2$ exists and is nonzero and finite.

Then

$$\begin{aligned} \lim_n \beta(g_{\theta_n}, n, k, \alpha) &= 1, \text{ iff } \lim \tfrac{\Delta_k}{\sqrt{k}} = \infty, \\ &= \alpha, \text{ iff } \lim \tfrac{\Delta_k}{\sqrt{k}} = 0, \end{aligned}$$

where as before $\beta(.)$ denotes the power of the test.

Remark: If $0 < \lim \frac{\Delta_k}{\sqrt{k}} < \infty$, then the power would typically converge to a number between α and 1, but a general characterization is lacking. The issue about letting k grow is that the approximate noncentral χ^2 distribution for Pearson χ^2 under the alternative has a noncentrality parameter increasing in k, which would make the distribution stochastically larger. On the other hand, by increasing k, the degree of freedom also increases, which would increase the variance.

Thus, there are two conflicting effects of increasing k and it is not clear which one will win. For certain alternatives, the increase in Δ_k beats the effect of the increase in the variance and the power converges to 1. For certain other alternatives, it does not. The tail of g relative to h is the key factor. The next result makes this precise. The following result (Kallenberg et al. (1985)) connects the condition $\lim \frac{\Delta_k}{\sqrt{k}} = \infty$ (0) to the thickness of the tail of the fixed alternative g.

Theorem 27.6.

a) Suppose $\limsup_n \min_i kp_{0i} > 0$. If for some $r > \frac{4}{3}$, $\int |f|^r dH < \infty$, then $\lim \frac{\Delta_k}{\sqrt{k}} = 0$.

b) Suppose $\liminf_n \min_i kp_{0i} > 0$ and $\limsup_n \max_i kp_{0i} < \infty$. If for some $0 < r < \frac{4}{3}$, $\int |f|^r dH = \infty$, then $\lim \frac{\Delta_k}{\sqrt{k}} = \infty$.

Remark: The assumption $\liminf_n \min_i kp_{0i} > 0$ says that none of the cells $A_{k,i}$ should have very small probabilities under h. Assumption b) that $\limsup \max_i kp_{0i} < \infty$, likewise says that none of the cells should have a high probability under h. The two assumptions are both satisfied if $p_{0i} \sim \frac{1}{k}$ for all i and k.

If g has a thick tail relative to h, then for small r, $\int |f|^r dH$ would typically diverge. To the contrary, if g has a thin tail relative to h, then $\int |f|^r dH$ would typically converge even for large r. So the combined qualitative conclusion of the theorems above is that if g has thick tails relative to h, then we can afford to choose a large number of cells, and if g has thin tails relative to h, then we should not use a large number of cells. These are useful general principles.

The next two examples illustrate the phenomenon.

Example 27.2. Let $h(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and $g(x) = \frac{1}{\pi(1+x^2)}$. Note that g has thick tails relative to h. Therefore,

$$f(x) = \frac{g(x)}{h(x)} - 1 = \frac{ce^{\frac{x^2}{2}}}{1+x^2} - 1, \text{ for some } 0 < c < \infty$$

$$\Rightarrow \int |f|^r dH = \frac{1}{\sqrt{2\pi}} \int \left| \frac{ce^{\frac{x^2}{2}}}{1+x^2} - 1 \right|^r e^{-\frac{x^2}{2}} dx$$

For any $r > 1$, this integral diverges. So, from the previous theorems, $\lim \beta(g_{\theta_n}, n, k_n, \alpha) = 1$.

Example 27.3. Let $h(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and $g(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$. The larger the σ is, the thicker is the tail of g relative to h. Now, $f(x) = ce^{\frac{x^2}{2}(1-\frac{1}{\sigma^2})} - 1$, for some $0 < c < \infty$. Therefore

$$\begin{aligned} \int |f|^r dH &= \frac{1}{\sqrt{2\pi}} \int \left| ce^{\frac{x^2}{2}(1-\frac{1}{\sigma^2})} - 1 \right|^r e^{-\frac{x^2}{2}} dx \\ &\sim \int e^{\frac{x^2}{2}\left[r(1-\frac{1}{\sigma^2})-1\right]} dx. \end{aligned}$$

If $r = \frac{4}{3}$ and $\sigma^2 = 4$, $r(1 - \frac{1}{\sigma^2}) - 1 = 0$. Also, $r(1 - \frac{1}{\sigma^2}) - 1 < 0$ and the integral $\int e^{\frac{x^2}{2}\left[r(1-\frac{1}{\sigma^2})-1\right]} dx$ converges for some $r < \frac{4}{3}$, iff $\sigma^2 < 4$. On the other hand, $r(1 - \frac{1}{\sigma^2}) - 1 > 0$ and the integral $\int e^{\frac{x^2}{2}\left[r(1-\frac{1}{\sigma^2})-1\right]} dx$ diverges for some $r < \frac{4}{3}$, iff $\sigma^2 > 4$. So, if g has a "small" variance, then letting $k \to \infty$ is not a good idea, while g has a "large" variance, then one can let $k \to \infty$. Note the similarity in conclusion to the previous example.

27.7 Exercises

Exercise 27.1. * For testing that $F = N(0,1)$, and with the cells as $(-\infty,-3),(-3,-2)$, ..., $(2,3),(3,\infty)$, find explicitly an alternative F_1 such that Pearson's chisquare is not consistent.

Exercise 27.2. For testing that a p-dimensional distribution is $N(0,I)$, find $k = 10$ spherical shells with equal probability under the null.

Exercise 27.3. * For testing $H_0 : F = N(0,1)$ vs. $H_1 : F = C(0,1)$, and with the cells as $(-\infty,-4a),(-4a,-3a),(-3a,-2a),...,(3a,4a),(4a,\infty)$, find a that maximizes $\sum_{i=1}^{k} \frac{(p_{1i}-p_{0i})^2}{p_{0i}}$. Why would you want to maximize it ?

Exercise 27.4. For $k = 6,8,10,12$, $n = 15,25,40$, and with the equiprobable cells, approximately find the power of the chisquare test for testing $F = Exp(1)$ vs. $F = Gamma(2,1); Gamma(5,1)$. Do more cells help ?

Exercise 27.5. For $k = 6,8,10,12$, $n = 15,25,40$, and with the equiprobable cells, approximately find the power of the chisquare test for testing $F = N(0,1)$ vs. $F = DoubleExp(0,1)$. Do more cells help ?

Exercise 27.6. For $k = 6,8,10,12$, $n = 15,25,40$, and with the equiprobable cells, approximately find the power of the chisquare test for testing $F = C(0,1)$ vs. $F = t(m), m = 2,5,10$. Do more cells help ?

Exercise 27.7. Prove that the Freeman-Tukey statistic defined in text is asymptotically a chisquare.

Exercise 27.8. * Prove or disprove : $E_{F_1}\chi^2 \geq E_{F_0}\chi^2 \, \forall F_1 \neq F_0$.

Exercise 27.9. * Find a formula for $Var_{F_1}\chi^2$.

Exercise 27.10. * Find the limiting distribution under the null of $\frac{\chi^2-k}{\sqrt{k}}$, where $k = k(n) \to \infty$; does a weak limit always exist ?

Exercise 27.11. * With $h = N(0,1), g = DoubleExp(0,1)$, in the notation of section 27.6, does $\beta(g_{\theta_n}, n, k, \alpha)$ converge to 1,0, or something in between ?

Exercise 27.12. * With $h = Gamma(2,1), g = lognormal(0,1)$, in the notation of section 27.6, does $\beta(g_{\theta_n}, n, k, \alpha)$ converge to 1,0, or something in between ?

27.8 References

Cox,D.R.(2000).Karl pearson and the chisquared test,Goodness of Fit Tests and Model Validity,3-8,Birkhauser,Boston.

Ferguson,T.S.(1996).A Course in Large Sample Theory,Chapman and Hall,London.

Greenwood,P.E. and Nikulin,M.S.(1996).A Guide to Chi-squared Testing,John Wiley,New York.

Kallenberg,W.C.M.,Oosterhoff,J. and Schriever,B.F.(1985). The number of classes in chi-squared goodness of fit tests,JASA,80,392,959-968.

Mann,H.B. and Wald,A.(1942).On the choice of the number of class intervals in the application of the chisquare test,Ann.Math.Statist.,13,306-317.

Oosterhoff,J.(1985).Choice of cells in chisquare tests,Stat.Neerlandica,39,2,115-128.

Pearson,K.(1900).On a criterion that a given system of deviations 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 5,50,157-175.

Rao,C.R.(2000).Pearson chisquare test, the dawn of statistical inference,Goodness of Fit Tests and Model Validity,9-24,Birkhauser,Boston.

Rao,K.C. and Dobson,D.S.(1974).A chi-square statistic for goodness of fit tests within the Exponential family,Comm.Stat.,3,1139-1153.

Serfling,R.(1980).Approximation Theorems of Mathematical Statistics,John Wiley,New York.

Stuart,A. and Ord,K.(1991).Kendall's Theory of Statistics,Fourth edition,Vol II,Clarendon Press,New York.