Powerful Goodness-of-Fit and Multi-Sample Tests

By

Jin Zhang

A thesis submitted to the Faculty of Graduate Studies in

partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Programme in Statistics

York University

Toronto, Ontario

July 2001

Powerful Goodness-of-Fit and Multi-Sample Tests

by **Jin Zhang**

a dissertation submitted to the Faculty of Graduate Studies of York University in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Abstract

There are a large number of goodness-of-fit tests in the literature. The most common used tests are Pearson's chi-squared tests and EDF (empirical distribution function) tests, such as Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling tests. The chi-squared tests are easy to use, but they are generally less powerful than EDF tests.

A parameterization approach is proposed to construct a general goodness-of-fit test for a specified distribution F_0. It includes traditional EDF tests, as well as new likelihood-ratio tests, which are the analogues of the old tests in representation but are generally much more powerful.

If F_0 has some unknown parameters, we need to estimate the parameters first and then apply the tests. Thus, we can test the goodness of fit for a family of distributions. To test normality, for example, suppose F_0 is a normal distribution with unknown mean and variance, we can estimate them by the sample mean and variance. Then the new tests can be applied to test the goodness of fit for normality. In such a case, they outperform the best tests of normality in the literature according to our simulation.

The methodology developed for goodness-of-fit tests is applied to the general two-sample problems. Similarly, we can not only generate classical two-sample tests, but also produce new powerful tests, which are sensitive to the difference in location, scale and shape between the distributions of the two-sampled populations. Conventional tests, however, are location-sensitive only.

Besides, the new two-sample tests are generalized to multi-sample tests, and parallel results have been obtained.

Since the exact sampling distributions of the EDF test statistics are intractable, a simple distribution family is introduced to approximate their sampling distributions in the end.

Acknowledgments

I take this opportunity to express my sincerest gratitude to Professor Yuehua Wu, my supervisor for Ph.D. thesis. I have benefited greatly from her for academic guidance, financial support and all kinds of helps, which led to my successful Ph.D. study at York University. From the bottom of my heart, I thank her for her kindness and friendship. I also thank Professors Georges Monette and Jianhong Wu, two other members of the Supervisory Committee, for their precious suggestions and helps.

I am grateful to my teachers for their kindly help and encouragement, especially to Professors Stephen Chamberlin, Peter Song and Masoud Asgharian, whom I had good fortune to learn from. My thanks are extended to the faculty and staff at the Department of Mathematics and Statistics, York University.

I would like to thank the Department of Mathematics & Statistics and the Faculty of Graduate Studies of York University, for the financial support of George and Frances Denzel Award for Excellence in Statistics and the Dean's Academic Excellence Scholarships,

Finally, I am indebted tremendously to my wife Jikun Yi and our daughter Yili Zhang, who sacrificed much and supported me to complete my Ph.D. study.

Table of Contents

List of Tables

List of Illustrations

One-Sample Tests

Two-Sample Tests

k-Sample Tests

To my mother Zhenying Zhu,
my wife Jikun Yi,
and our daughter Yili Zhang

In the memory of my father Junmo Zhang

1. Introduction

There are many kinds of goodness-of-fit tests in the literature. Some of them are special purpose tests, so that they are suitable and perform well only for some special situations. Others are omnibus tests that are applicable to general cases.

The most common used omnibus tests are Pearson's chi-squared tests and the tests based on EDF (empirical distribution function), such as traditional Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling tests. Chi-square tests are easy to use, but they are generally less powerful than EDF tests (D'Agostino and Stephens, 1986).

We now introduce a new method based on parameterization, which can not only generate traditional EDF tests, but also produce new powerful omnibus tests.

Let X be a continuous random variable with distribution function $F(x)$, and X_1, X_2, ..., X_n be a random sample from X with order statistics $X_{(1)}$, $X_{(2)}$, ..., $X_{(n)}$. We wish to test the nuil hypothesis

$$H: \ F(x) = F_0(x), \quad \texttt{for all } x \in (-\infty, \infty)$$

against the general alternative

$$\bar{H}: \ F(x) \neq F_0(x), \quad \texttt{for some } x \in (-\infty, \infty)$$

where $F_0(x)$ is a hypothesized distribution function to be tested. Here we discuss only the basic situation where $F_0(x)$ is completely known. For other cases, see

Section 6. Note that

$$H = \bigcap_{t\in(-\infty,\ \infty)} H_t \quad \text{and} \quad \bar{H} = \bigcup_{t\in(-\infty,\ \infty)} \bar{H}_t\ ,$$

with $H_t:\ F(t) = F_0(t)$ and $\bar{H}_t:\ F(t) \neq F_0(t)$. Then testing H vs. $\bar{H}$ is equivalent to testing H_t vs. $\bar{H}_t$ for every $t \in (-\infty,\ \infty)$.

To test H_t vs. $\bar{H}_t$ with t fixed, we have a binary random sample based on the indicator function: $X_{it} = I(X_i \leq t)$ $(i = 1, 2, ..., n)$ satisfying $P(X_{it} = 1) = F(t)$ and $P(X_{it} = 0) = 1 - F(t)$.

Note that $F(x)$ is an unknown distribution function arbitrary, while $F(t)$ with t fixed is nothing but a unknown parameter. Through introducing the new binary sample, the nonparametric test for H vs. $\bar{H}$ is simplified to a family of parametric tests for H_t vs. $\bar{H}_t$, $t \in (-\infty,\ \infty)$. The simplification is a process of parameterization, through which parametric approaches can be applied to nonparametric tests.

For each fixed $t \in (-\infty,\ \infty)$ and the corresponding random sample $X_{1t},\ X_{2t},\ ...,\ X_{nt}$, let Z_t be a statistic for testing H_t vs. $\bar{H}_t$ such that its large values reject H_t. Then two types of statistics for testing H vs. $\bar{H}$ can be defined by

$$Z = \int_{-\infty}^{\infty} Z_t\ dw(t) \qquad \text{and} \qquad Z_{max} = \sup_{t\in(-\infty,\ \infty)} [\ Z_t w(t)\ ]\ , \tag{1.1}$$

where $w(t)$ is some weight function and large values of Z or Z_{max} reject the null hypothesis H.

The power of Z or Z_{max} depends on Z_t and $w(t)$. Two natural candidates for Z_t are Pearson's chi-squared test statistic and the likelihood-ratio test statistic, which

are respectively (after simplification)

$$X_t^2 = \frac{n[F_n(t) - F_0(t)]^2}{F_0(t)[1 - F_0(t)]} \tag{1.2}$$

and

$$G_t^2 = 2n\{\ F_n(t) \log \frac{F_n(t)}{F_0(t)} + [1 - F_n(t)] \log \frac{1 - F_n(t)}{1 - F_0(t)}\ \}\ , \tag{1.3}$$

where $F_n(t)$ is the empirical distribution function of the original sample X_1, X_2, ..., X_n.

A large family of Z_t which embeds X_t^2 and G_t^2 can be obtained by using the Cressie and Read (1984) family of divergence statistics $2nI^\lambda$ for testing the goodness of fit of a multinomial distribution. In fact, for the above binary sample X_{1t}, X_{2t}, ..., X_{nt} with t fixed, the Cressie-Read family of divergence statistics for testing H_t vs. $\bar{H}_t$ is

$$2nI_t^\lambda = \frac{2n}{\lambda(\lambda+1)}\{\ F_n(t)[\frac{F_n(t)}{F_0(t)}]^\lambda + [1 - F_n(t)][\frac{1 - F_n(t)}{1 - F_0(t)}]^\lambda\ -\ 1\ \}\ , \tag{1.4}$$

which includes X_t^2 (λ=1) and G_t^2 (λ=0), as well as other important statistics (Cressie and Read 1984; Read and Cressie 1988).

We focus on X_t^2 and G_t^2 because X_t^2 is associated with classical tests while G_t^2 is the best choice of Z_t in (1.1) among the family (1.4) according to our simulation.

By choosing different weight functions, we will show in Sections 2-4 by simulation that (a) using X_t^2 as Z_t generates traditional EDF test statistics; (b) using G_t^2 as Z_t produces new EDF tests; (c) the new EDF tests are generally more powerful than traditional goodness-of-fit tests.

The sampling distributions of the new test statistics are intractable. Their em-

pirical percentage points are given in Section 5. In Sections 6-7, we will consider the case when F_0 has some parameters unknown. As a typical example, we discuss the goodness-of-fit test for normality, which is a fundamental issue in statistics and has always been a hot topic in the literature. Our simulation indicates that the new EDF tests outperform the best tests of normality in the literature.

In Sections 9-11, the idea of parameterization for one-sample tests is developed and applied to the general two-sample problem. Similarly, we can not only generate traditional two-sample tests, but also produce new powerful distribution-free tests. Furthermore, the results for two-sample tests are generalized to k-sample tests (see Sections 12-15). Monte Carlo simulation shows that conventional multi-sample tests are sensitive to location difference among distributions, but are dull to detect the variation in shape. However, the new tests are both location- and shape-sensitive.

A simple distribution family is introduced in Section 16-19 to approximate the distribution functions of goodness-of-fit test statistics, but the approach is applicable to general continuous random variables. Some traditional EDF test statistics are used as illustrative examples. Finally, concluding remarks are given in Section 20.

2. Traditional Goodness-of-Fit Tests Based on EDF

Using X_t^2 in (1.2) as Z_t in (1.1) but choosing different weight functions, we can derive traditional goodness-of-fit tests based on EDF. Below are three examples.

1. $w(t) = n^{-1}F_0(t)[1 - F_0(t)]$

Replacing Z_t of statistic Z_{max} in (1.1) with X_t^2 in (1.2) generates

$$\begin{aligned} K_S^2 &= [\sup_{t\in(-\infty,\,\infty)} \mid F_n(t) - F_0(t) \mid]^2 \\ &= [\max_{1\le i\le n}\{\max[\frac{i}{n} - F_0(X_{(i)}),\ F_0(X_{(i)})\frac{i-1}{n}]\}]^2, \end{aligned}$$

where K_S is the Kolmogorov-Smirnov statistic, the most well-known statistic for goodness-of-fit tests (Kolmogorov, 1933; Smirnov, 1939; Massey, 1951; Stephens, 1970, 1974; Conover, 1980; Pratt and Gibbons, 1981; D'Agostino and Stephens, 1986; Gibbons, 1992; Cabaña, 1996).

2. $w(t) = F_0(t)$

Replacing Z_t of statistic Z in (1.1) with X_t^2 in (1.2) generates the Anderson-Darling statistic

$$\begin{aligned} A^2 &= n\int_{-\infty}^{\infty} [F_n(t) - F_0(t)]^2\ F_0(t)^{-1}[1 - F_0(t)]^{-1}\ dF_0(t) \\ &= -\frac{2}{n}\sum_{i=1}^{n}\left[(i - 0.5)\log F_0(X_{(i)}) + (n - i + 0.5)\log[1 - F_0(X_{(i)})]\right] - n\ , \end{aligned}$$

one of the most powerful and important goodness-of-fit tests in the literature (Anderson and Darling, 1952, 1954; Stephens, 1970, 1974; D'Agostino and Stephens, 1986; Sinclair and Spurr, 1988). The last equality here (as well as that below) can be obtained by evaluating the associated integral.

3. $dw(t) = F_0(t)[1 - F_0(t)]dF_0(t)$

Replacing Z_t of statistic Z in (1.1) with X_t^2 in (1.2) generates the famous Cramér-von Mises statistic (Cramér, 1928; von Mises, 1931; Smirnov, 1936, 1937; Stephens, 1970, 1974; Knott, 1974; Conover, 1980; D'Agostino and Stephens, 1986; Csörgő

and Faraway, 1996; Spinelli and Stephens, 1997)

$$W^2 = n\int_{-\infty}^{\infty} [F_n(t) - F_0(t)]^2 \, dF_0(t)$$
$$= \sum_{i=1}^{n} [\, F_0(X_{(i)}) - \frac{i-0.5}{n} \,]^2 + \frac{1}{12n} \, .$$

3. New Powerful EDF Tests

As a goodness-of-fit test for a multinomial distribution, the Pearson's chi-squared statistic is asymptotically equivalent to the likelihood-ratio statistic. Therefore, under the null hypothesis H_t in Section 1, the chi-squared statistic X_t^2 in (1.2) and the likelihood- ratio statistic G_t^2 in (1.3) are equivalent in large sample situations, but they are different under the alternative $\bar{H}_t$. We have seen in Section 2 that traditional tests can be generated by using X_t^2 as Z_t in (1.1). In this section we shall use G_t^2 to produce new powerful tests by choosing proper weight functions.

For any continuous hypothetical distribution function F_0, let $U_i = F_0(X_i)$ (i=1, 2, ..., n) so that $U_{(i)} = F_0(X_{(i)})$. Note that X_1, X_2, ..., X_n are i.i.d. from F_0 if and only if U_1, U_2, ..., U_n are i.i.d. from $U(0, 1)$, the standard uniform distribution. To test H vs. $\bar{H}$, consider a statistic with form $T = T(U_1, U_2, ..., U_n)$, where $T(\cdots)$ is a given function independent of F_0. Since U_i and $1 - U_i$ are identically distributed under H, a reasonable T should satisfy

$$T(U_1, U_2, ..., U_n) = T(1 - U_1, 1 - U_2, ..., 1 - U_n) \, .$$

In such a case, we say that T is *distribution-symmetric* about the median.

It is obvious that the traditional statistics K_S, W^2 and A^2 in Section 2 are

functions of U_1, U_2, ..., U_n, and they are distribution-symmetric. In order to generate new distribution-symmetric tests, we have to choose proper weight functions. Moreover, we sometimes need to make modifications to $F_n(t)$ at its discontinuous points $X_{(i)}$ $(i = 1, 2, ..., n)$ by defining $F_n(X_{(i)}) = (i - c)/(n + 1 - 2c)$, where c is a constant between 0 and 1. The natural and intuitive choice of c is 0.5 so that $F_n(X_{(i)}) = (i - 0.5)/n$ or $[F_n(X_{(i)} - 0) + F_n(X_{(i)} + 0)]/2$, which is a common way (something like continuity correction) to modify the empirical distribution function. In fact, we can imagine that at point $x = X_{(i)}$, there are $i - 0.5$ or $n - i + 0.5$ observations among X_1, X_2, ..., X_n which are less or greater than the x. Indeed, our simulation shows that $c = 0.5$ seems to be the best choice in terms of power. Finally, the traditional test statistics in Section 2 also suggest that $F_n(X_{(i)})$ should be $(i - 0.5)/n$ instead of i/n.

When necessary, we always define $F_n(X_{(i)}) = (i-0.5)/n$. Then new distribution-symmetric tests can be generated by choosing proper weight functions as follows.

1. $w(t) = 1$

Let $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$. Replacing Z_t of statistic Z_{max} in (1.1) with G_t^2 in (1.3) produces

$$\sup_{t\in(-\infty,\,\infty)} G_t^2 = \max_{0\le i\le n}\{\sup_{X_{(i)}\le t<X_{(i+1)}} G_t^2\} = \max_{1\le i\le n} G_{X_{(i)}}^2 ,$$

which is equivalent to

$$Z_K = \max_{1\le i\le n}\left[(i-0.5)\log\frac{i-0.5}{nF_0(X_{(i)})} + (n-i+0.5)\log\frac{n-i+0.5}{n[1-F_0(X_{(i)})]}\right] . \quad (3.1)$$

2. $dw(t) = F_n(t)^{-1}[1 - F_n(t)]^{-1}dF_n(t)$

Replacing Z_t of statistic Z in (1.1) with G_t^2 in (1.3) produces

$$2\sum_{i=1}^{n}\left[\frac{n}{n-i+0.5}\log\frac{i-0.5}{nF_0(X_{(i)})}+\frac{n}{i-0.5}\log\frac{n-i+0.5}{n[1-F_0(X_{(i)})]}\right],$$

which is equivalent to

$$Z_A=-\sum_{i=1}^{n}\left[\frac{\log F_0(X_{(i)})}{n-i+0.5}+\frac{\log[1-F_0(X_{(i)})]}{i-0.5}\right]. \tag{3.2}$$

3. $dw(t)=F_0(t)^{-1}[1-F_0(t)]^{-1}dF_0(t)$

Replacing Z_t of statistic Z in (1.1) with G_t^2 in (1.3) produces

$$\sum_{i=1}^{n}\left[\log[F_0(X_{(i)})^{-1}-1]-b_{i-1}+b_i\right]^2+C_n,$$

where C_n is a constant and $b_i=i\log(i/n)+(n-i)\log(1-i/n)$.

Since $b_{i-1}-b_i\approx\log[(n-0.5)/(i-0.75)-1]$, the above test statistic is approximately equivalent to

$$Z_C=\sum_{i=1}^{n}\left[\log\frac{F_0(X_{(i)})^{-1}-1}{(n-0.5)/(i-0.75)-1}\right]^2. \tag{3.3}$$

The new statistics Z_K, Z_A and Z_C are distribution-symmetric. They look like traditional K_S, A^2 and W^2 respectively, but they are generally much more powerful. According to our simulation, they are sensitive not only to the location or scale, but also to the shape of the alternative distribution.

4. Power Comparison by Simulation

In this section we will use the Monte Carlo approach to simulate the powers of the new statistics Z_A, Z_C, Z_K and the traditional Kolmogorov-Smirnov statistic

K_S, Cramér-von Mises statistic W^2, Anderson-Darling statistic A^2 and Pearson's chi-squared statistic X^2. For the chi-squared test, the sample observations need be grouped. Here we use the associated function in Splus with default values (See, e.g., *S-Plus 2000 Guide to Statistics, Volume 1*, Data Analysis Products Division, MathSoft, Seattle, WA, p. 100).

The simulation size is 10,000, and the significance level or the probability of type I error for testing goodness of fit is $\alpha = 0.05$. For various situations about the null hypothesis H and the alternative $\bar{H}$, all simulated powers for the seven statistics are illustrated with graphs, where the powers are plotted against the sample size n for selected values of n=10, 20, 30, 50, 70, 100, 150, 200 and 300.

Example 4.1:

$$H: X_1, ..., X_n \overset{iid}{\sim} U(0, 1) \text{ vs. } \bar{H}: X_1, ..., X_n \overset{iid}{\sim} Beta(p, q)$$

Without loss of generality, we can assume that the underlying distribution F is the standard uniform $U(0, 1)$ under the null hypothesis H. Then the natural candidate for F under the alternative $\bar{H}$ is the beta distribution $Beta(p, q)$ with parameters p and q, which includes the uniform $U(0, 1)$ or $Beta(1, 1)$. So, this example is actually a parametric test for $H: (p, q) = (1, 1)$ vs. $\bar{H}: (p, q) \neq (1, 1)$.

For $(p, q) = (0.6, 0.8), (0.6, 0.6), (0.8, 0.8), (1.3, 1.3), (1.6, 1.6), (1.3, 1.6)$, the powers of Z_A, Z_C, Z_K, K_S, W^2, A^2, X^2 under $\bar{H}$ are plotted in Fig. 4.1 respectively. We see that Z_A or Z_C has the highest power in the cases where $p, q > 1$ or $p, q < 1$, and they dominate all others. Although Z_K is not as powerful as Z_A and Z_C, it is still overwhelmingly powerful compared to its analogue K_S.

We also consider the power of the entropy-based test of uniformity proposed by Dudewicz and van der Meulen (1981). Their method involves choosing the best integer m which depends on the sample size n, but their power results in Table 3 are obtained from choosing the best m not only for different n but also for different alternative distributions. Of course, different tests fit different models. However, when performing a nonparametric test, we have no idea about the alternative distribution. Therefore, if a fixed m is used for the same n but different alternative distributions, the power of such a test is generally lower than that of Anderson-Darling test A^2.

Example 4.2:

$$H: X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2) \ vs. \ \bar{H}: X_1, ..., X_n \overset{iid}{\sim} t(k)$$

Because of the importance of the normal distribution, F is assumed to be a normal distribution $N(\mu, \sigma^2)$ under the null hypothesis H. It is interesting to consider that (a) F also has a symmetric distribution under the alternative $\bar{H}$, say $t(k)$, the t distribution with k degrees of freedom; (b) both distributions have the same mean and variance, i.e. $\mu = 0$ and $\sigma^2 = k/(k-2)$.

Since $N(0, 1) = t(\infty)$, testing $H: F = N(\mu, \sigma^2)$ $vs.$ $\bar{H}: F = t(k)$ is equivalent to testing $H: k = \infty$ $vs.$ $\bar{H}: k \neq \infty$. Fig. 4.2 compares the powers of the seven statistics Z_A, Z_C, Z_K, K_S, W^2, A^2, X^2 for k=3, 5, 10. Obviously Z_C is the best and Z_A, Z_C, Z_K dominate the others (sometimes they are much more powerful).

The Cauchy and logistic distributions are also typical examples of symmetric distributions, which can be considered as the underlying distribution under $\bar{H}$. The

power comparison for logistic distribution is just like that for $t(9)$, while the Cauchy distribution is $t(1)$.

Example 4.3:

$$H: X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2) \ vs. \ \bar{H}: X_1, ..., X_n \overset{iid}{\sim} Gamma(r, 1)$$

In this example F is also assumed to be $N(\mu, \sigma^2)$ under H, but it has an asymmetric distribution under $\bar{H}$, such as $Gamma(r, 1)$, the gamma distribution with shape parameter r and scale parameter 1, which includes exponential and chi-squared distributions. We also assume that both distributions have the same mean and variance, i.e. $\mu = r$ and $\sigma^2 = r$.

Similarly, since the asymptotic distribution of $Gamma(r, 1)$ is normal when $r \longrightarrow \infty$, testing $H: F = N(\mu, \sigma^2) \ vs. \ \bar{H}: F = Gamma(r, 1)$ is equivalent to testing $H: r = \infty \ vs. \ \bar{H}: r \neq \infty$.

Simulated powers of the seven statistics for r=5, 10 and 20 are also plotted in Fig. 4.2, which shows that (a) Z_A, Z_C, Z_K dominate the others; (b) the powers of Z_A and Z_C are sometimes substantially higher than those of the traditional statistics.

Other asymmetric distributions, such as the log-normal, Weibull, F and Beta were also considered as alternative distributions against the normal. These situations are similar to that of gamma.

Example 4.4:

$$H: X_1, ..., X_n \overset{iid}{\sim} N(0, 1) \ vs. \ \bar{H}: X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

In the last example, F is assumed to be normal under both H and $\bar{H}$. Without loss of generality, we need to consider only the test for $H: F = N(0, 1)$ *vs.* $\bar{H}: F = N(\mu, \sigma^2)$, or equivalently, $H: (\mu, \sigma^2) = (0, 1)$ *vs.* $\bar{H}: (\mu, \sigma^2) \neq (0, 1)$.

Six cases are considered with alternatives (1) $N(0.1, 1)$, (2) $N(0.4, 1)$, (3) $N(0, 1.5)$, (4) $N(0, 2)$, (5) $N(0.1, 2)$ and (6) $N(0.4, 1.5)$. Note that in cases 1 and 2, the two distributions have the same variance but different means, and in cases 3 and 4, they have the same mean but different variances. In cases 5 and 6, means and variances are both different.

For normal models, the distributions differ in mean and variance only. There is no shape difference in terms of skewness and kurtosis. For each case Fig. 4.3 compares the powers of the seven tests, as well as the optimal parametric t- test and the χ^2-test for normal mean and variance.

It is clear that for cases 1, 2, 6 where the major difference between the two distributions arises from means rather than variances, there is no significant difference in power between the new tests and their analogues A^2, W^2 and K_S. Conversely, for the other three cases, the advantage of the new tests is obvious. When the difference in distribution arises from means only, such as cases 1 and 2, the six tests are almost as powerful as the optimal t-test. In cases 3 and 4 where the only difference comes from variances, the power lost by using the new tests over the χ^2-test is much less than that by using their analogues. Among A^2, W^2 and K_S, A^2 is almost the best in all cases, which is also true for other examples.

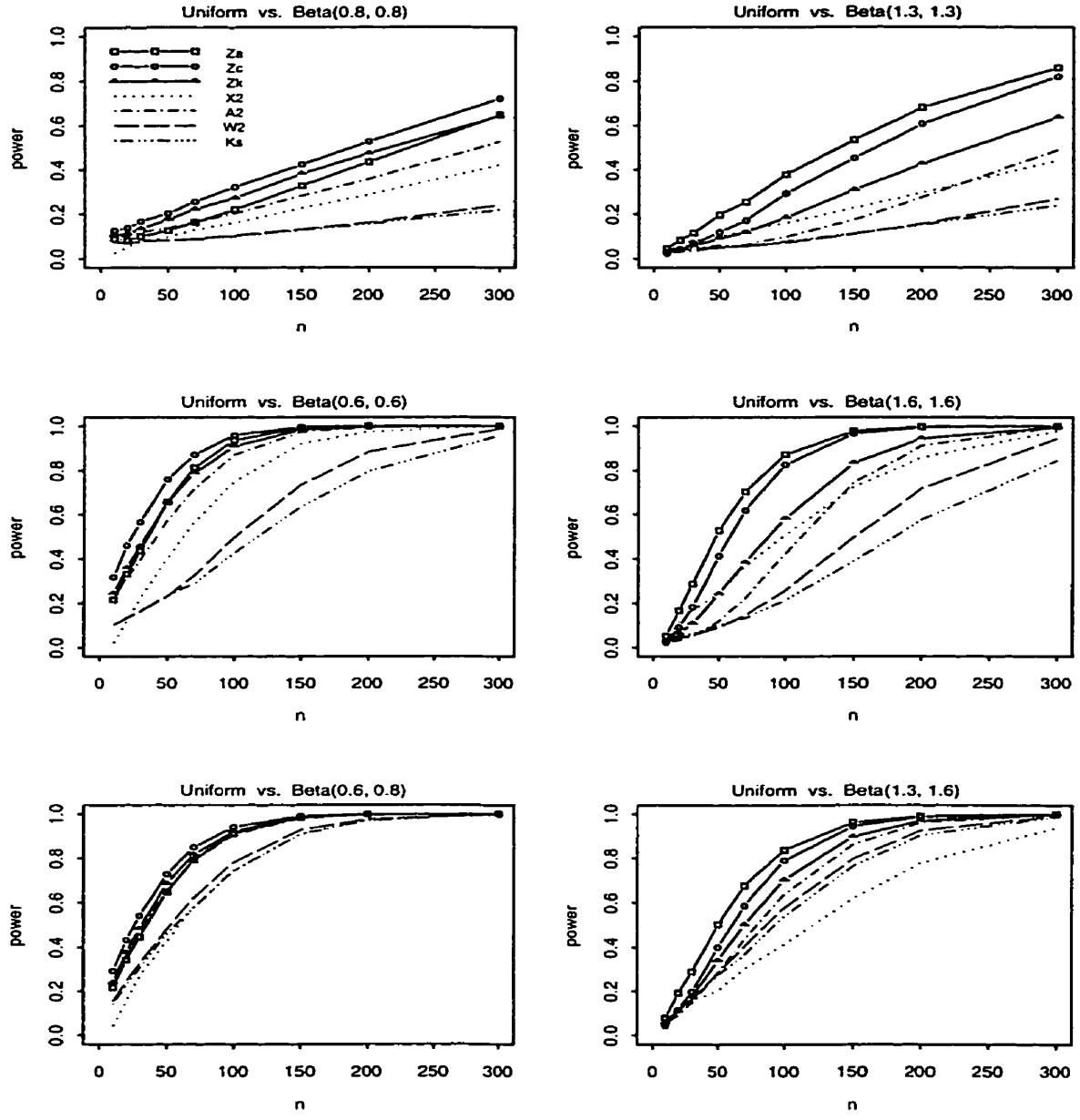

Fig. 4.1. Power comparison when testing $H : F = U(0,\ 1)$ vs $\bar{H} : F = Beta(p,\ q)$ at level $\alpha = 0.05$

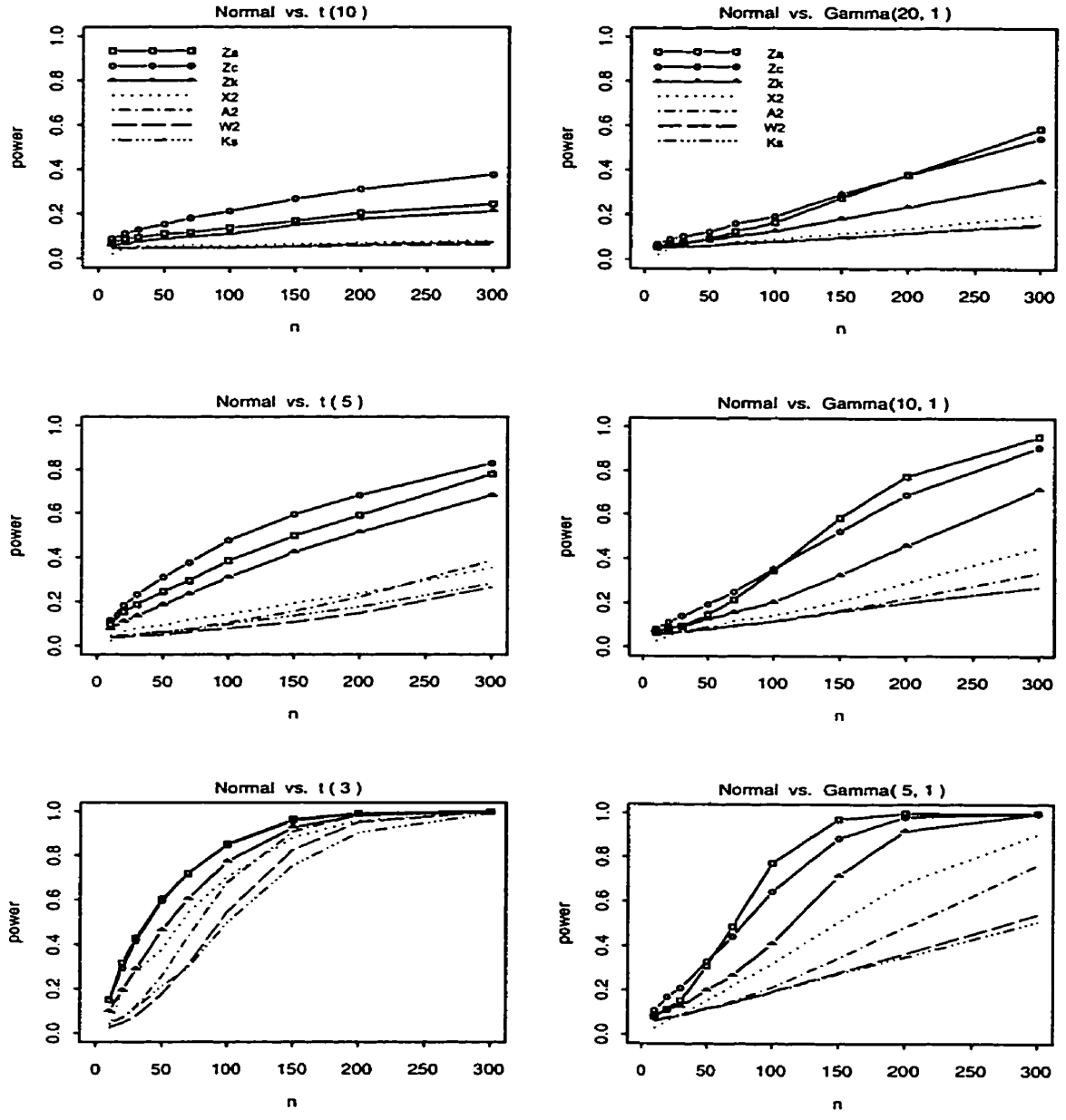

Fig. 4.2. Power comparison when testing (a) $H : F = N(\mu,\ \sigma^2)$ vs $\bar{H} : F = t(k)$ and (b) $H : F = N(\mu,\ \sigma^2)$ vs $\bar{H} : F = Gamma(r,\ 1)$ at level $\alpha = 0.05$

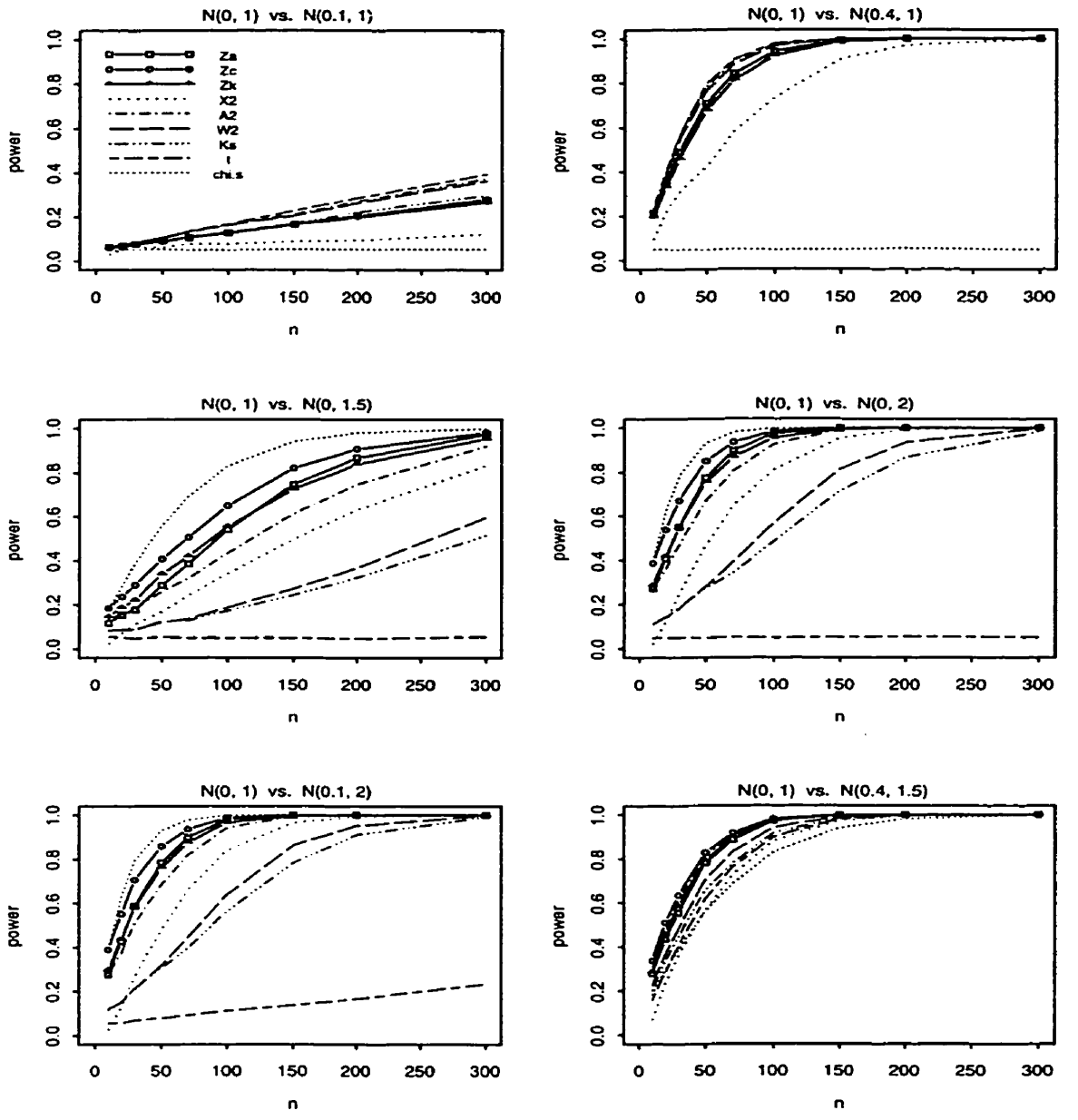

Fig. 4.3. Power comparison when testing $H : F = N(0,\ 1)$ *vs* $\tilde{H} : F = N(\mu,\ \sigma^2)$ at level $\alpha = 0.05$

5. The Distributions of Z_A, Z_C and Z_K

Like the Anderson-Darling A^2, Cramér-von Mises W^2 and Kolmogorov- Smirnov K_S, the new statistics Z_A, Z_C and Z_K are distribution- free. Our simulation on skewness and kurtosis shows that the sampling distributions of Z_A, Z_C and Z_K converge very slowly. Therefore, it is of limited practical value to study their asymptotic distributions. Just as for A^2, W^2 and K_S, it is difficult to find their exact null distributions for finite sample cases except for small sample sizes.

Again Monte Carlo simulation is used to approximate the percentage points of Z_A, Z_C and Z_K for some selected sample sizes. Tables 5.1-5.3 respectively give their approximate percentage points, which are based on a simulation of size one million. The p-quantiles or percentage points of these tables are provided for use at $p =$ 0.001, 0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.95, 0.99, 0.999.

The simulation error in Tables 5.1-5.3 can be estimated in terms of confidence intervals for the true percentage levels rather percentage points. For each simulated percentage point at level p, the 99.73% confidence interval for the true percentage level is $p \pm 3\sqrt{p(1-p)/N}$, where N=1,000,000 is the replicates of simulation. Thus, the 99.73% confidence intervals for the true percentage levels in Tables 5.1-5.3 are respectively 0.001 ± 0.0001, 0.01 ± 0.0003, 0.05 ± 0.0007, 0.10 ± 0.0009, 0.20 ± 0.0012, 0.30 ± 0.0014, 0.4 ± 0.0015, 0.50 ± 0.0015, 0.60 ± 0.0015, 0.70 ± 0.0014, 0.80 ± 0.0012, 0.90 ± 0.0009, 0.95 ± 0.0007, 0.99 ± 0.0003, 0.999 ± 0.0001.

TABLE 5.1. (a) Percentage points for $10Z_A$ - 32 (Lower Tail)

n	0.001	0.01	0.05	0.10	0.20	0.30	0.40	0.50
5	-0.1639	0.140	0.666	1.094	1.824	2.549	3.334	4.235
6	0.0718	0.358	0.836	1.224	1.875	2.514	3.200	3.989
7	0.2386	0.506	0.942	1.292	1.879	2.452	3.064	3.763
8	0.3532	0.613	1.015	1.335	1.872	2.392	2.946	3.574
9	0.4454	0.683	1.061	1.359	1.854	2.327	2.834	3.405
10	0.5140	0.742	1.095	1.374	1.831	2.270	2.736	3.261
12	0.6158	0.823	1.137	1.381	1.782	2.164	2.566	3.017
14	0.6861	0.872	1.156	1.375	1.732	2.071	2.427	2.824
16	0.7350	0.906	1.164	1.364	1.686	1.989	2.309	2.665
18	0.7659	0.929	1.168	1.351	1.645	1.923	2.214	2.536
20	0.7964	0.945	1.167	1.336	1.607	1.862	2.127	2.421
25	0.8427	0.972	1.159	1.301	1.528	1.741	1.961	2.204
30	0.8668	0.982	1.147	1.271	1.467	1.650	1.838	2.047
40	0.8980	0.990	1.122	1.220	1.377	1.521	1.668	1.831
50	0.9120	0.990	1.102	1.184	1.314	1.433	1.555	1.689
70	0.9233	0.983	1.070	1.132	1.230	1.319	1.410	1.510
100	0.9270	0.974	1.038	1.085	1.157	1.222	1.289	1.361
150	0.9272	0.960	1.007	1.040	1.092	1.138	1.184	1.234
200	0.9246	0.951	0.988	1.014	1.054	1.089	1.125	1.164
300	0.9207	0.940	0.966	0.985	1.013	1.037	1.062	1.089
500	0.9156	0.928	0.945	0.957	0.975	0.991	1.006	1.023
1000	0.9097	0.917	0.926	0.933	0.942	0.951	0.959	0.968

TABLE 5.1. (b) Percentage points for $10Z_A$ - 32 (Upper Tail)

n	0.50	0.60	0.70	0.80	0.90	0.95	0.99	0.999
5	4.235	5.317	6.71	8.70	12.24	15.98	25.43	39.77
6	3.989	4.929	6.14	7.86	10.88	14.04	22.10	34.55
7	3.763	4.599	5.67	7.18	9.83	12.62	19.68	30.39
8	3.574	4.322	5.28	6.63	9.01	11.49	17.69	27.46
9	3.405	4.084	4.95	6.18	8.32	10.54	16.17	24.74
10	3.261	3.887	4.69	5.80	7.75	9.79	14.89	22.74
12	3.017	3.554	4.23	5.19	6.86	8.60	12.89	19.64
14	2.824	3.294	3.89	4.73	6.19	7.70	11.46	17.32
16	2.665	3.087	3.62	4.37	5.67	7.01	10.35	15.47
18	2.536	2.917	3.40	4.07	5.24	6.46	9.47	14.11
20	2.421	2.769	3.21	3.82	4.89	5.99	8.69	12.84
25	2.204	2.490	2.85	3.35	4.23	5.13	7.32	10.71
30	2.047	2.291	2.60	3.03	3.76	4.53	6.39	9.22
40	1.831	2.022	2.26	2.59	3.16	3.75	5.17	7.37
50	1.689	1.845	2.04	2.31	2.77	3.25	4.41	6.16
70	1.510	1.626	1.77	1.97	2.31	2.65	3.49	4.75
100	1.361	1.445	1.55	1.69	1.93	2.18	2.78	3.70
150	1.234	1.292	1.36	1.46	1.63	1.80	2.20	2.81
200	1.164	1.209	1.26	1.34	1.47	1.59	1.90	2.35
300	1.089	1.120	1.16	1.21	1.30	1.38	1.59	1.90
500	1.023	1.042	1.07	1.10	1.15	1.20	1.33	1.52
1000	0.968	0.978	0.99	1.01	1.03	1.06	1.13	1.22

TABLE 5.2. (a) Percentage points for Z_C (Lower Tail)

n	0.001	0.01	0.05	0.10	0.20	0.30	0.40	0.50
5	0.319	0.801	1.63	2.26	3.24	4.13	5.04	6.03
6	0.437	0.977	1.86	2.51	3.54	4.46	5.41	6.44
7	0.536	1.118	2.04	2.72	3.78	4.74	5.71	6.76
8	0.621	1.241	2.20	2.91	3.99	4.97	5.97	7.05
9	0.703	1.352	2.34	3.06	4.18	5.18	6.19	7.29
10	0.762	1.444	2.46	3.20	4.34	5.36	6.40	7.52
12	0.892	1.625	2.68	3.45	4.63	5.68	6.75	7.90
14	1.002	1.763	2.86	3.65	4.87	5.95	7.04	8.21
16	1.087	1.893	3.02	3.83	5.07	6.18	7.29	8.48
18	1.168	1.999	3.15	3.99	5.26	6.38	7.51	8.73
20	1.232	2.087	3.27	4.12	5.41	6.55	7.69	8.93
25	1.408	2.286	3.53	4.41	5.75	6.93	8.11	9.38
30	1.520	2.472	3.74	4.65	6.02	7.24	8.44	9.74
40	1.720	2.714	4.07	5.03	6.46	7.72	8.97	10.32
50	1.907	2.928	4.33	5.32	6.79	8.09	9.39	10.76
70	2.130	3.244	4.72	5.76	7.31	8.66	9.99	11.42
100	2.390	3.583	5.15	6.23	7.84	9.25	10.64	12.12
150	2.701	3.981	5.63	6.78	8.47	9.95	11.38	12.93
200	2.939	4.256	5.98	7.16	8.90	10.41	11.90	13.48
300	3.269	4.662	6.48	7.71	9.54	11.11	12.66	14.29
500	3.630	5.178	7.11	8.41	10.33	11.97	13.60	15.30
1000	4.217	5.867	7.96	9.37	11.42	13.17	14.86	16.65

TABLE 5.2. (b) Percentage points for Z_C (Upper Tail)

n	0.50	0.60	0.70	0.80	0.90	0.95	0.99	0.999
5	6.03	7.18	8.61	10.6	14.2	18.3	30.6	54.6
6	6.44	7.63	9.11	11.2	14.8	18.9	31.1	54.9
7	6.76	7.98	9.50	11.6	15.3	19.4	31.5	54.4
8	7.05	8.29	9.84	12.0	15.7	19.9	31.8	54.8
9	7.29	8.56	10.14	12.3	16.1	20.2	32.1	55.6
10	7.52	8.81	10.42	12.6	16.5	20.6	32.4	55.1
12	7.90	9.22	10.85	13.1	17.0	21.2	32.9	56.2
14	8.21	9.56	11.24	13.5	17.5	21.7	33.4	56.2
16	8.48	9.86	11.57	13.9	17.9	22.2	33.8	56.3
18	8.73	10.13	11.85	14.2	18.3	22.6	34.3	56.8
20	8.93	10.35	12.10	14.5	18.6	22.9	34.5	57.1
25	9.38	10.82	12.62	15.1	19.3	23.6	35.4	57.6
30	9.74	11.22	13.05	15.5	19.8	24.2	35.8	57.4
40	10.32	11.86	13.75	16.3	20.7	25.2	36.9	59.1
50	10.76	12.34	14.26	16.9	21.3	25.9	37.5	59.4
70	11.42	13.05	15.03	17.7	22.3	26.9	38.6	60.2
100	12.12	13.79	15.84	18.6	23.3	28.0	39.8	61.4
150	12.93	14.67	16.79	19.6	24.4	29.2	41.1	62.2
200	13.48	15.26	17.43	20.3	25.2	30.1	42.2	63.5
300	14.29	16.13	18.36	21.4	26.3	31.3	43.5	64.7
500	15.30	17.21	19.52	22.6	27.7	32.8	45.1	66.1
1000	16.65	18.66	21.06	24.3	29.6	34.8	47.3	68.5

TABLE 5.3. (a) Percentage points for Z_K (Lower Tail)

n	0.001	0.01	0.05	0.10	0.20	0.30	0.40	0.50
5	0.0385	0.0952	0.190	0.268	0.402	0.532	0.672	0.829
6	0.0572	0.1225	0.229	0.316	0.461	0.601	0.749	0.916
7	0.0740	0.1462	0.263	0.356	0.510	0.658	0.812	0.985
8	0.0870	0.1670	0.293	0.392	0.554	0.708	0.868	1.048
9	0.1014	0.1865	0.319	0.423	0.592	0.751	0.916	1.100
10	0.1130	0.2041	0.344	0.451	0.626	0.790	0.960	1.148
12	0.1380	0.2370	0.387	0.502	0.686	0.857	1.034	1.228
14	0.1575	0.2634	0.422	0.542	0.734	0.911	1.095	1.295
16	0.1740	0.2873	0.453	0.578	0.775	0.957	1.145	1.350
18	0.1900	0.3094	0.480	0.610	0.814	1.000	1.191	1.400
20	0.2039	0.3268	0.504	0.637	0.844	1.034	1.229	1.441
25	0.2367	0.3683	0.555	0.694	0.910	1.108	1.309	1.527
30	0.2589	0.3986	0.596	0.741	0.964	1.167	1.373	1.597
40	0.3001	0.4492	0.660	0.811	1.047	1.259	1.473	1.703
50	0.3298	0.4879	0.707	0.866	1.108	1.326	1.544	1.781
70	0.3756	0.5460	0.778	0.943	1.195	1.420	1.645	1.887
100	0.4208	0.6040	0.850	1.022	1.285	1.517	1.750	1.998
150	0.4728	0.6665	0.926	1.108	1.381	1.623	1.861	2.117
200	0.5075	0.7091	0.977	1.164	1.444	1.689	1.933	2.191
300	0.5564	0.7702	1.051	1.243	1.531	1.782	2.031	2.296
500	0.6174	0.8414	1.134	1.335	1.633	1.891	2.145	2.415
1000	0.6859	0.9332	1.239	1.449	1.757	2.024	2.284	2.557

TABLE 5.3. (b) Percentage points for Z_K (Upper Tail)

n	0.50	0.60	0.70	0.80	0.90	0.95	0.99	0.999
5	0.829	1.02	1.25	1.59	2.16	2.74	4.10	6.01
6	0.916	1.11	1.36	1.71	2.30	2.90	4.30	6.29
7	0.985	1.19	1.45	1.81	2.42	3.03	4.46	6.45
8	1.048	1.26	1.52	1.89	2.51	3.13	4.57	6.59
9	1.100	1.32	1.59	1.96	2.59	3.22	4.69	6.70
10	1.148	1.37	1.64	2.02	2.66	3.30	4.77	6.83
12	1.228	1.46	1.74	2.13	2.78	3.43	4.94	7.08
14	1.295	1.53	1.82	2.22	2.88	3.54	5.06	7.20
16	1.350	1.59	1.88	2.29	2.96	3.63	5.16	7.31
18	1.400	1.64	1.94	2.35	3.04	3.71	5.25	7.41
20	1.441	1.69	1.99	2.41	3.10	3.78	5.32	7.48
25	1.527	1.78	2.09	2.51	3.21	3.90	5.47	7.67
30	1.597	1.85	2.17	2.60	3.31	4.01	5.58	7.78
40	1.703	1.97	2.29	2.73	3.46	4.17	5.78	7.97
50	1.781	2.05	2.38	2.82	3.55	4.27	5.89	8.13
70	1.887	2.16	2.50	2.96	3.70	4.42	6.05	8.31
100	1.998	2.28	2.62	3.09	3.84	4.57	6.23	8.52
150	2.117	2.41	2.76	3.22	3.99	4.73	6.40	8.67
200	2.191	2.48	2.84	3.31	4.08	4.83	6.51	8.84
300	2.296	2.59	2.95	3.43	4.21	4.96	6.65	8.97
500	2.415	2.72	3.08	3.56	4.35	5.10	6.80	9.16
1000	2.557	2.87	3.23	3.72	4.52	5.28	6.99	9.35

6. Tests of Normality

We discuss the goodness-of-fit tests for a specified distribution in Sections 1-5 where the underlying distribution function $F_0(x)$ is assumed to be completely known. To discuss the goodness-of-fit tests for a family of distributions, Suppose now that $F_0(x)$ has some unknown parameters. In such a case, we need to estimate the parameters first and then apply the test statistics Z_K, Z_C and Z_A in (3.1)-(3.3). However, the statistics are no longer distribution-free because we are testing the goodness of fit for a family of distributions instead of a specific one. For different families, the sampling distributions of the statistics are different.

Since normal distributions are the most important distributions in statistics, we now consider the normal distribution family $N(\mu,\ \sigma^2)$ with mean μ and variance σ^2. To test if X has a normal distribution $N(\mu,\ \sigma^2)$ with μ and σ^2 given, Z_K, Z_C and Z_A can be used as goodness-of-fit test statistics with $F_0(x) = \Phi(\frac{x-\mu}{\sigma})$, where $\Phi(x)$ denotes the distribution function of $N(0,\ 1)$, the standard normal distribution.

Usually, μ and $\sigma > 0$ are unknown parameters. Then tests based on Z_K, Z_C and Z_A are not applicable because $F_0(x)$ is unknown. In such a case, we estimate μ and σ by the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ respectively and then apply the goodness-of-fit tests. It is recommended to always estimate the parameters no matter whether they are known or not. Estimating the parameters from the data can improve the power of the tests when they are actually known (see the end of Section 7).

Comparison of the powers for testing normality will be given in the next sec-

tion, where Z_K, Z_C and Z_A are compared with the best existing test statistics, including the Shapiro-Wilk statistic W (Shapiro and Wilk, 1965, 1968), Anderson-Darling statistic A^2 (Anderson and Darling, 1952, 1954) and D'Agostino's statictic D (D'Agostino, 1971, 1973).

The sampling distributions of Z_A, Z_C and Z_K are intractable. Table 6.1, 6.2, and 6.3 respectively give their approximate percentage points for testing normality, which are based on Monte Carlo simulation of size 1,000,000. The p-quantiles or percentage points in these tables are provided for use at $p = 0.001$, 0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.95, 0.99, 0.999.

The simulation error in Tables 6.1-6.3 can be estimated in terms of confidence intervals for the true percentage levels rather percentage points. For each simulated percentage point at level p, the 99.73% confidence interval for the true percentage level is $p \pm 3\sqrt{p(1-p)/N}$, where N=1,000,000 is the replicates of simulation. Thus, the 99.73% confidence intervals for the true percentage levels in Tables 5.1-5.3 are respectively 0.001 ± 0.0001, 0.01 ± 0.0003, 0.05 ± 0.0007, 0.10 ± 0.0009, 0.20 ± 0.0012, 0.30 ± 0.0014, 0.4 ± 0.0015, 0.50 ± 0.0015, 0.60 ± 0.0015, 0.70 ± 0.0014, 0.80 ± 0.0012, 0.90 ± 0.0009, 0.95 ± 0.0007, 0.99 ± 0.0003, 0.999 ± 0.0001.

TABLE 6.1. (a) Percentage points for Z_A when testing normality (Lower Tail)

n	0.001	0.01	0.05	0.10	0.20	0.30	0.40	0.50
5	3.165	3.170	3.181	3.192	3.209	3.226	3.244	3.261
6	3.187	3.194	3.207	3.217	3.234	3.250	3.266	3.283
7	3.203	3.211	3.224	3.234	3.251	3.266	3.281	3.298
8	3.216	3.224	3.237	3.247	3.263	3.277	3.291	3.307
9	3.226	3.234	3.246	3.256	3.271	3.284	3.298	3.314
10	3.233	3.242	3.254	3.263	3.277	3.290	3.304	3.318
12	3.245	3.253	3.264	3.273	3.286	3.298	3.311	3.324
14	3.253	3.261	3.272	3.280	3.292	3.303	3.314	3.327
16	3.259	3.266	3.277	3.284	3.295	3.306	3.317	3.329
18	3.264	3.271	3.280	3.287	3.298	3.308	3.318	3.329
20	3.267	3.274	3.283	3.290	3.300	3.309	3.319	3.329
25	3.274	3.280	3.288	3.293	3.302	3.311	3.319	3.328
30	3.278	3.283	3.290	3.295	3.304	3.311	3.318	3.327
40	3.282	3.287	3.293	3.297	3.304	3.310	3.316	3.323
50	3.285	3.289	3.294	3.298	3.304	3.309	3.314	3.320
70	3.287	3.291	3.295	3.298	3.303	3.307	3.311	3.316
100	3.289	3.292	3.295	3.297	3.301	3.304	3.307	3.311
150	3.290	3.292	3.294	3.296	3.299	3.301	3.304	3.306
200	3.290	3.292	3.294	3.295	3.298	3.299	3.301	3.304
300	3.291	3.292	3.293	3.294	3.296	3.297	3.299	3.300
500	3.291	3.291	3.292	3.293	3.294	3.295	3.296	3.297
1000	3.290	3.291	3.291	3.292	3.292	3.293	3.294	3.294

TABLE 6.1. (b) Percentage points for Z_A when testing normality (Upper Tail)

n	0.50	0.60	0.70	0.80	0.90	0.95	0.99	0.999
5	3.261	3.282	3.308	3.346	3.409	3.472	3.619	3.757
6	3.283	3.304	3.330	3.366	3.428	3.492	3.643	3.823
7	3.298	3.317	3.342	3.376	3.437	3.499	3.650	3.847
8	3.307	3.326	3.350	3.383	3.441	3.502	3.648	3.854
9	3.314	3.332	3.355	3.386	3.442	3.499	3.641	3.844
10	3.318	3.336	3.357	3.388	3.441	3.497	3.632	3.826
12	3.324	3.340	3.360	3.388	3.438	3.488	3.612	3.798
14	3.327	3.342	3.361	3.387	3.432	3.478	3.594	3.764
16	3.329	3.342	3.360	3.384	3.426	3.469	3.574	3.729
18	3.329	3.342	3.358	3.381	3.420	3.460	3.557	3.703
20	3.329	3.341	3.357	3.378	3.415	3.452	3.543	3.678
25	3.328	3.339	3.352	3.371	3.403	3.435	3.513	3.625
30	3.327	3.336	3.348	3.365	3.393	3.422	3.490	3.591
40	3.323	3.331	3.341	3.355	3.378	3.402	3.456	3.538
50	3.320	3.327	3.336	3.348	3.367	3.387	3.434	3.503
70	3.316	3.321	3.328	3.337	3.353	3.368	3.405	3.457
100	3.311	3.315	3.320	3.328	3.339	3.351	3.379	3.419
150	3.306	3.309	3.313	3.318	3.327	3.336	3.356	3.386
200	3.304	3.306	3.309	3.313	3.320	3.327	3.343	3.367
300	3.300	3.302	3.304	3.307	3.312	3.317	3.329	3.347
500	3.297	3.298	3.300	3.302	3.305	3.308	3.316	3.327
1000	3.294	3.295	3.296	3.297	3.298	3.300	3.305	3.311

TABLE 6.2. (a) Percentage points for Z_C when testing normality (Lower Tail)

n	0.001	0.01	0.05	0.10	0.20	0.30	0.40	0.50
5	0.664	0.724	0.874	1.007	1.252	1.501	1.749	2.002
6	0.704	0.808	1.003	1.176	1.475	1.756	2.028	2.298
7	0.745	0.881	1.120	1.322	1.662	1.966	2.255	2.555
8	0.781	0.945	1.218	1.446	1.819	2.144	2.453	2.778
9	0.813	0.997	1.306	1.556	1.955	2.299	2.631	2.979
10	0.842	1.049	1.388	1.658	2.079	2.438	2.789	3.155
12	0.895	1.138	1.526	1.831	2.290	2.682	3.065	3.463
14	0.935	1.212	1.645	1.972	2.464	2.888	3.298	3.723
16	0.976	1.277	1.746	2.096	2.616	3.064	3.495	3.944
18	1.014	1.334	1.838	2.207	2.754	3.222	3.675	4.147
20	1.046	1.396	1.924	2.309	2.875	3.361	3.835	4.328
25	1.120	1.519	2.103	2.519	3.137	3.664	4.176	4.707
30	1.170	1.618	2.246	2.693	3.349	3.910	4.456	5.023
40	1.285	1.783	2.483	2.972	3.693	4.307	4.901	5.521
50	1.366	1.912	2.674	3.193	3.957	4.612	5.248	5.913
70	1.512	2.131	2.963	3.535	4.367	5.079	5.771	6.499
100	1.693	2.369	3.279	3.902	4.810	5.590	6.344	7.132
150	1.891	2.653	3.655	4.339	5.327	6.175	6.999	7.862
200	2.043	2.867	3.923	4.649	5.696	6.593	7.464	8.376
300	2.298	3.196	4.338	5.118	6.245	7.209	8.149	9.123
500	2.609	3.596	4.861	5.702	6.932	7.977	8.990	10.055
1000	3.072	4.191	5.588	6.526	7.885	9.045	10.169	11.346

TABLE 6.2. (b) Percentage points for Z_C when testing normality (Upper Tail)

n	0.50	0.60	0.70	0.80	0.90	0.95	0.99	0.999
5	2.002	2.261	2.577	3.002	3.639	4.213	5.460	6.757
6	2.298	2.602	2.967	3.433	4.158	4.849	6.382	8.280
7	2.555	2.891	3.288	3.797	4.610	5.385	7.196	9.635
8	2.778	3.141	3.567	4.123	5.007	5.864	7.908	10.954
9	2.979	3.362	3.810	4.404	5.350	6.267	8.545	12.102
10	3.155	3.559	4.035	4.659	5.663	6.650	9.138	13.115
12	3.463	3.902	4.424	5.105	6.209	7.304	10.156	15.147
14	3.723	4.195	4.755	5.486	6.669	7.862	11.103	16.969
16	3.944	4.443	5.036	5.808	7.065	8.352	11.838	18.474
18	4.147	4.669	5.287	6.099	7.422	8.767	12.493	19.899
20	4.328	4.869	5.511	6.362	7.752	9.157	13.150	21.149
25	4.707	5.298	5.994	6.918	8.438	9.984	14.432	23.753
30	5.023	5.649	6.391	7.375	8.998	10.662	15.580	26.091
40	5.521	6.209	7.031	8.109	9.888	11.733	17.223	29.333
50	5.913	6.648	7.522	8.683	10.594	12.583	18.480	31.707
70	6.499	7.302	8.262	9.540	11.640	13.835	20.399	35.532
100	7.132	8.011	9.059	10.452	12.758	15.171	22.242	39.126
150	7.862	8.818	9.970	11.488	14.027	16.628	24.405	42.354
200	8.376	9.391	10.613	12.244	14.934	17.714	25.839	44.611
300	9.123	10.220	11.530	13.276	16.179	19.139	27.523	46.663
500	10.055	11.246	12.674	14.567	17.717	20.927	29.760	49.888
1000	11.346	12.654	14.224	16.322	19.796	23.301	32.811	53.458

TABLE 6.3. (a) Percentage points for Z_K when testing normality (Lower Tail)

n	0.001	0.01	0.05	0.10	0.20	0.30	0.40	0.50
5	0.004	0.015	0.042	0.067	0.107	0.142	0.176	0.213
6	0.010	0.030	0.068	0.096	0.139	0.178	0.218	0.261
7	0.019	0.046	0.088	0.119	0.167	0.211	0.255	0.303
8	0.029	0.061	0.107	0.141	0.193	0.240	0.288	0.340
9	0.038	0.073	0.124	0.160	0.216	0.267	0.319	0.373
10	0.047	0.085	0.139	0.177	0.237	0.291	0.345	0.403
12	0.063	0.106	0.165	0.209	0.275	0.334	0.392	0.455
14	0.076	0.124	0.189	0.236	0.307	0.370	0.432	0.498
16	0.088	0.140	0.209	0.260	0.335	0.401	0.466	0.536
18	0.099	0.154	0.228	0.281	0.360	0.429	0.497	0.570
20	0.110	0.168	0.245	0.300	0.382	0.454	0.525	0.600
25	0.133	0.196	0.281	0.340	0.429	0.506	0.582	0.663
30	0.149	0.219	0.310	0.373	0.467	0.549	0.629	0.714
40	0.177	0.255	0.355	0.424	0.527	0.615	0.703	0.795
50	0.202	0.284	0.391	0.464	0.573	0.667	0.759	0.857
70	0.237	0.328	0.444	0.524	0.641	0.743	0.843	0.949
100	0.272	0.373	0.500	0.586	0.714	0.824	0.932	1.046
150	0.314	0.424	0.562	0.656	0.795	0.915	1.031	1.155
200	0.342	0.457	0.604	0.703	0.849	0.976	1.100	1.231
300	0.383	0.509	0.667	0.774	0.930	1.065	1.198	1.339
500	0.432	0.569	0.741	0.857	1.027	1.174	1.318	1.470
1000	0.498	0.651	0.840	0.968	1.156	1.318	1.476	1.645

TABLE 6.3. (b) Percentage points for Z_K when testing normality (Upper Tail)

n	0.50	0.60	0.70	0.80	0.90	0.95	0.99	0.999
5	0.213	0.257	0.310	0.383	0.512	0.630	0.899	1.175
6	0.261	0.312	0.373	0.456	0.596	0.731	1.034	1.413
7	0.303	0.359	0.425	0.515	0.666	0.814	1.149	1.584
8	0.340	0.399	0.471	0.567	0.726	0.882	1.240	1.724
9	0.373	0.435	0.511	0.612	0.777	0.940	1.313	1.838
10	0.403	0.468	0.547	0.652	0.824	0.992	1.379	1.924
12	0.455	0.525	0.610	0.723	0.906	1.083	1.493	2.063
14	0.498	0.573	0.663	0.783	0.975	1.162	1.583	2.191
16	0.536	0.614	0.709	0.834	1.035	1.229	1.661	2.258
18	0.570	0.651	0.750	0.878	1.087	1.286	1.732	2.365
20	0.600	0.684	0.786	0.920	1.135	1.339	1.795	2.406
25	0.663	0.753	0.862	1.006	1.238	1.458	1.944	2.601
30	0.714	0.810	0.925	1.076	1.319	1.553	2.072	2.757
40	0.795	0.899	1.023	1.188	1.453	1.708	2.269	3.039
50	0.857	0.967	1.098	1.273	1.555	1.826	2.440	3.291
70	0.949	1.067	1.210	1.399	1.708	2.005	2.685	3.660
100	1.046	1.174	1.328	1.535	1.874	2.202	2.969	4.087
150	1.155	1.295	1.464	1.690	2.061	2.429	3.289	4.529
200	1.231	1.379	1.558	1.799	2.195	2.591	3.507	4.785
300	1.339	1.497	1.690	1.950	2.380	2.803	3.793	5.228
500	1.470	1.643	1.854	2.137	2.606	3.068	4.160	5.756
1000	1.645	1.836	2.069	2.385	2.904	3.422	4.620	6.300

7. Comparison of Power for Testing Normality

There are a large number of tests for normality in the literature (e.g., D'Agostino and Stephens, 1986) since normal distributions are the most important ones in statistics. Some of these tests are only sensitive to certain kinds of departures from normality, such as directional tests based on skewness and kurtosis. Thus they are suitable and perform well only for some special situations. Others are omnibus tests, which are applicable to general cases.

It is well known that the most powerful omnibus test of normality in the literature is the Shapiro-Wilk statistic W (Shapiro and Wilk, 1965, 1968; Shapiro, Wilk and Chen, 1968), which is essentially the squared ratio of the best linear unbiased estimator for scale to the standard deviation. Unfortunately, W is applicable to small sample sizes only. In fact, W is computable exactly up to n=20, and valid approximations exist up to n=50. Therefore, various kinds of modifications to the W have been made, including the Shapiro and Francia's (1972) W' and the well-known D'Agostino's (1971, 1973) statistic D which we will discuss soon, but more or less they loose some powers (e.g., D'Agostino and Stephens, 1986).

Another important omnibus test of normality is the Anderson-Darling statistic A^2 (Anderson and Darling, 1952, 1954; Stephens, 1974; Sinclair and Spurr, 1988), which is slightly less powerful than W but is the best existing EDF test (D'Agostino and Stephens, 1986, p. 404).

As general goodness-of-fit tests, the new EDF statistics Z_K, Z_C and Z_A in Section 3 are much more powerful than traditional ones. When used as omnibus tests of

normality, they are expected to perform well. In the following, we will compare the powers of the six statistics Z_K, Z_C Z_A, W, A^2 and D for various kinds of situations about alternative distributions for testing normality. Monte Carlo approach is used with simulation size 10,000, and the significance level of the test is $\alpha = 0.05$. For each situation, the simulated powers of the six statistics for n=10, 20, 30, 40, and 50 are plotted for comparison (the W test requires $n \leq 50$).

Example 7.1: Normal vs Beta(p, q)

In the first example of testing normality, the alternative distribution is the beta distribution Beta(p, q) with parameters p and q, which is symmetric when $p = q$ and includes the standard uniform distribution $U(0, 1)$=Beta(1, 1) and an asymptoticly normal distribution, i.e, Beta(∞, ∞).

The powers of Z_K, Z_C Z_A, W, A^2 and D are plotted in Fig. 7.1 for (p, q)=(2, 2), (2, 1), (1, 1.5), (1, 1), (0.5, 1) and (0.5, 0.5), which correspond to different beta distributions with various departures from normality.

We can see that the performance of the D'Agotino's statistic D is extremely poor with its power almost equaling the significance level $\alpha = 0.05$ for some cases. The differences in power among Z_C, Z_A and W are not so obvious for all cases. None of these three statistics is always the best, but overall, they are the best among the six statistics. It seems that A^2 generally performs a little bit better than Z_K (refer to other examples).

Example 7.2: Normal vs $t(k)$

In the second example, the alternative distribution is $t(k)$, the t distribution with k degrees of freedom, which is symmetric and includes Cauchy distribution, i.e., $t(1)$ and the standard normal distribution $N(0,\ 1) = t(\infty)$. For different values of k, the powers of the six statistics are plotted in Fig. 7.2. In all cases, there is no major difference of powers among them with D being the best and W the worst.

Example 7.3: Normal vs $Gamma(a,\ b)$

In the third example, the alternative distribution is Gamma$(a,\ b)$, the gamma distribution with shape parameter a and scale parameter b, which includes exponential and chi-squared distributions.

Since all six statistics Z_K, Z_C Z_A, W, A^2 and D are invariant under any affine transformation $Y = (X - c)/d$ with $d > 0$, we need just consider Gamma$(a,\ 1)$ without loss of generality, which is a non-symmetric distribution but is asymptoticly normal as a goes into infinity.

The powers of these statistics for different values of a, say 1, 3 and 7, are also exhibited in Fig. 7.2. It can be seen that Z_A is the best and dominates the others. There is almost no difference in power between Z_C and W, which are slightly less powerful than Z_A. Also, there are no major differences between Z_K and A^2. Finally, D behaves poorly and is dominated by the others.

Example 7.4: Normal vs Weibull$(a,\ b)$

In the fourth example, the alternative distribution is Weibull$(a,\ b)$, the Weibull distribution with shape parameter a and scale parameter b.

As in Example 7.3, we just need to consider Weibull(a, 1) without loss of generality. Fig. 7.3 compares the powers of the six statistics for different values of a. The results are much similar to those in Example 7.3, but the advantage of Z_A is more obvious.

Example 7.5: Normal vs Lognormal(μ, σ)

In the last example, the alternative distribution is Lognormal(μ, σ), the lognormal distribution with parameters μ and σ. We just consider the case of Lognormal(0, σ) since it has the same skewness and kurtosis as Lognormal(μ, σ).

Powers of the six tests at different σ are given in Fig. 7.3, which exhibits a clear pattern of domination with the following ranks:

$$Z_A \succ Z_C \succ W \succ A^2 \succ Z_K \succ D \,. \tag{7.1}$$

We can also consider other alternative distributions, such as logistic and F distributions, but the situations are similar. It can be seen from all the examples that as omnibus tests of normality,

(a) the new EDF test statistics Z_A, Z_C and Z_K are very powerful and robust for various kinds of departures from normality;

(b) Z_A and Z_C are generally outperform W and dominate A^2 and Z_K;

(c) Z_K is almost as powerful as A^2, while its analogue, the Kolomorov-Smirnov statistic, is well known to be very poor for testing normality (e.g., D'Agostino and Stephens, 1986);

(d) D is not a very good omnibus test statistic, but it performs very well in some situations.

(e) (7.1) is also the overall ranks of the six statistics in terms of power performance.

Finally, it is important that when applying EDF tests of normality, we had better estimate the parameters in $F_0(x) = \Phi(\frac{x-\mu}{\sigma})$ no matter whether they are known or not. Estimating the parameters from the data can significantly improve the power of the test when they are actually known. This phenomenon has been observed by Stephens (1974) and Dyer (1974), and it can also be demonstrated by the examples here together with those in Section 4.

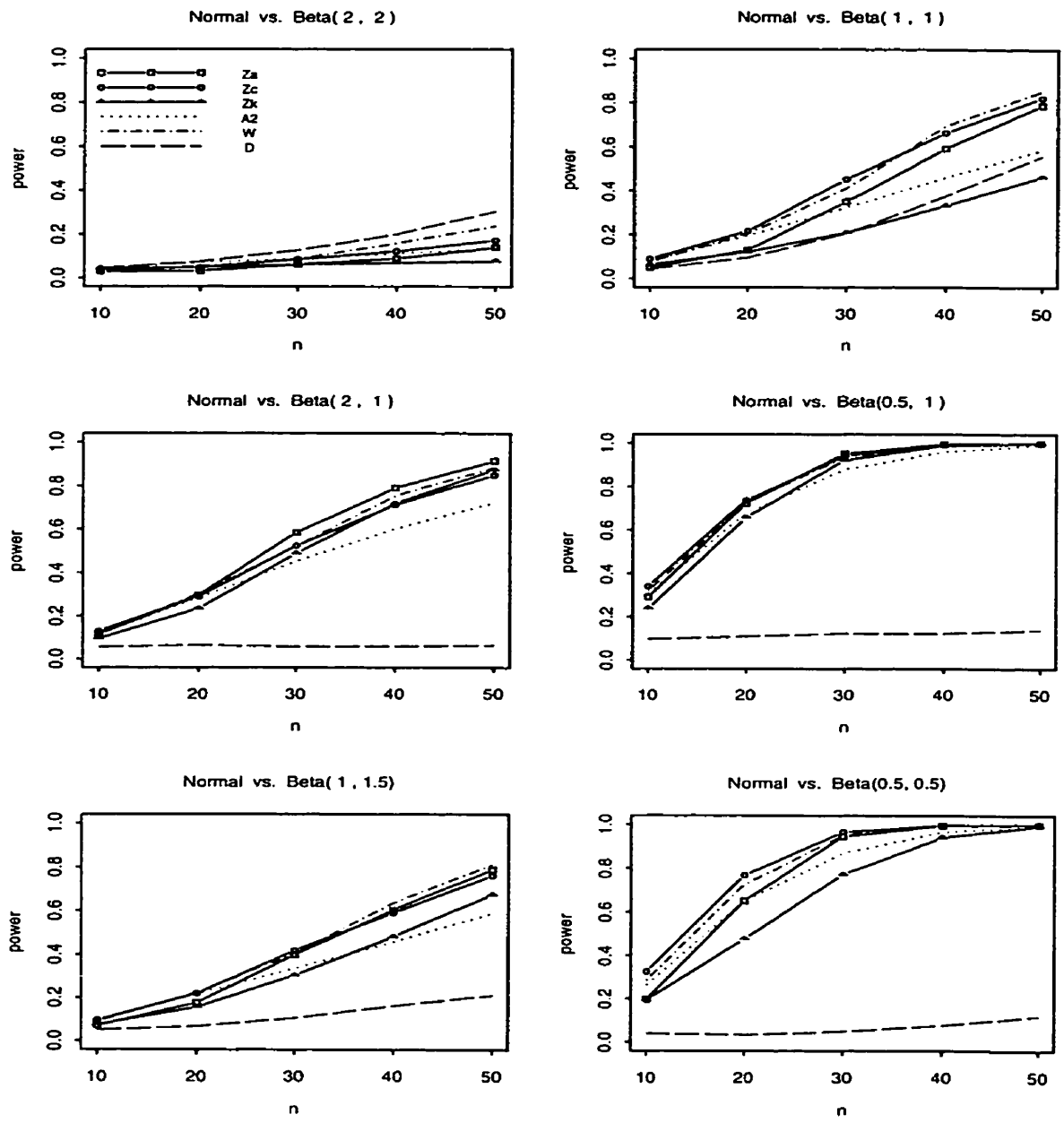

Fig. 7.1. Power comparison when testing Normal vs $Beta(p,\ q)$ at level $\alpha = 0.05$

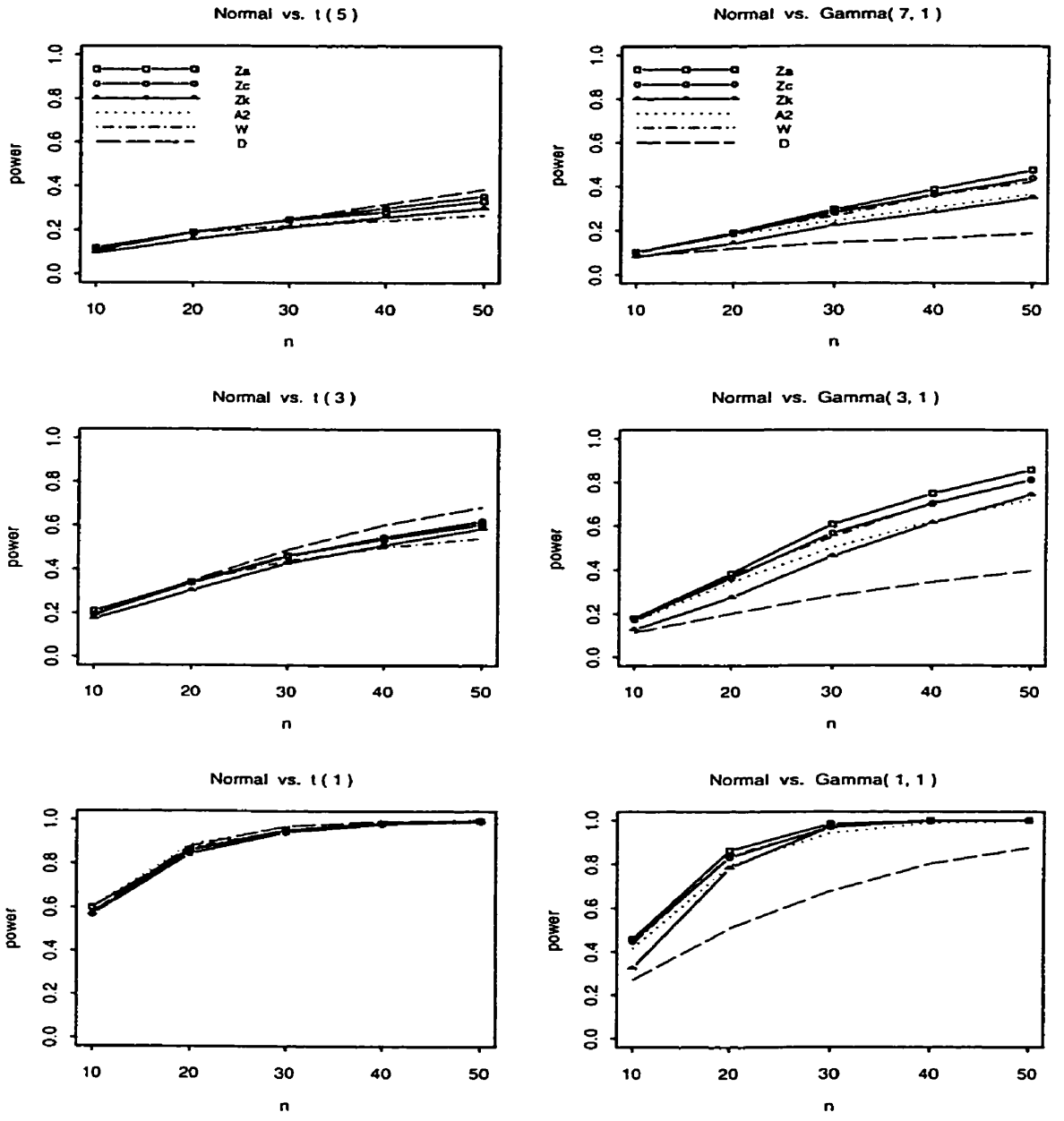

Fig. 7.2. Power comparison when testing Normal vs $t(k)$ and Normal vs $Gamma(a,\ b)$ at level $\alpha = 0.05$

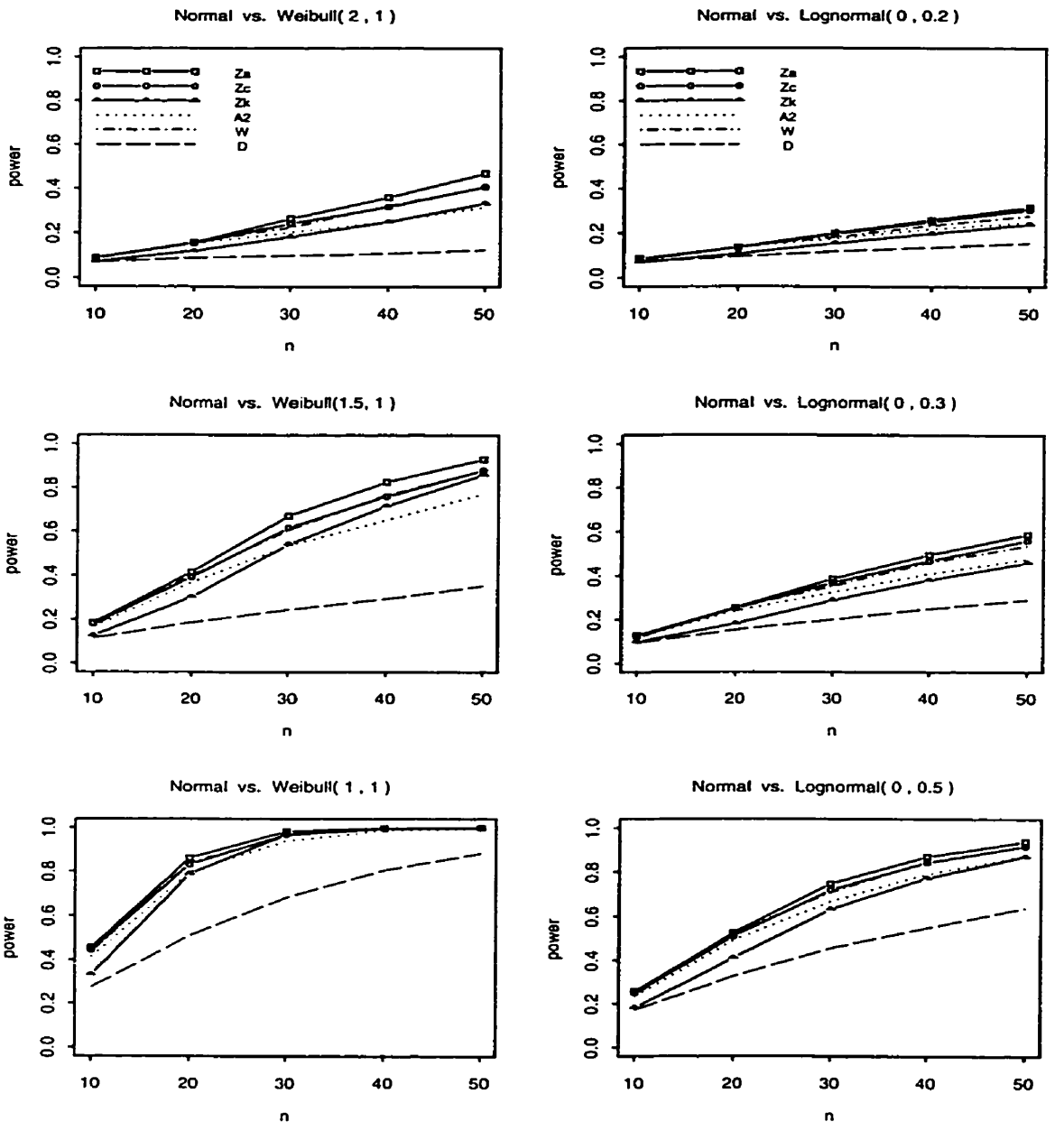

Fig. 7.3. Power comparison when testing Normal vs $Weibull(a,\ b)$ and Normal vs $Lognormal(\mu,\ \sigma^2)$ at level $\alpha = 0.05$

8. General Two-Sample Problem

In this and next section, we will use the method of parameterization introduced in Section 1 to study the two-sample tests.

Let X_{i1}, X_{i2}, ..., X_{in_i} be a random sample from a continuous population with distribution function $F_i(x)$ (i=1, 2), and let X_1, X_2, ..., X_n ($n = n_1 + n_2$) be the pooled sample with order statistics $X_{(1)}$, $X_{(2)}$, ..., $X_{(n)}$. Denote R_{ij} the rank of the j-th ordered observation $X_{i(j)}$ in the pooled sample. We wish to test the null hypothesis

$$H: \ F_1(x) = F_2(x), \quad \texttt{for all } x \in (-\infty, \ \infty)$$

against the alternative

$$\bar{H}: \ F_1(x) \neq F_2(x), \quad \texttt{for some } x \in (-\infty, \ \infty).$$

Since

$$H = \bigcap_{t\in(-\infty, \ \infty)} H_t \quad \text{and} \quad \bar{H} = \bigcup_{t\in(-\infty, \ \infty)} \bar{H}_t \ ,$$

where

$$H_t: \ F_1(t) = F_2(t) \quad \text{and} \quad \bar{H}_t: \ F_1(t) \neq F_2(t)$$

testing H vs. $\bar{H}$ is equivalent to testing H_t vs. $\bar{H}_t$ for every $t \in (-\infty, \ \infty)$.

To test H_t vs. $\bar{H}_t$ with t fixed, we define new samples based on index function: $X_{ijt} = I(X_{ij} \leq t)$ ($i = 1, 2$; $j = 1, \ 2, \ ..., \ n_i$) satisfying $P(X_{ijt} = 1) = F_i(t)$ and $P(X_{ijt} = 0) = 1 - F_i(t)$.

For each fixed $t \in (-\infty, \ \infty)$ and the corresponding random samples X_{i1t}, X_{i2t}, ..., X_{in_it} (i=1, 2), let Z_t be a statistic for testing H_t vs. $\bar{H}_t$ such that large values

reject H_t. Then two types of statistics for testing H vs. $\bar{H}$ can be defined by

$$Z = \int_{-\infty}^{\infty} Z_t \, dw(t) \qquad \text{and} \qquad Z_{max} = \sup_{t \in (-\infty, \infty)} [\, Z_t w(t) \,] \,, \tag{8.1}$$

where $w(t)$ is some weight function and the large value of Z or Z_{max} rejects the null hypothesis H.

(8.1) is the same as (1.1) in the one-sample case, but the Pearson's chi- squared test statistic in (1.2) and the likelihood-ratio test statistic in (1.3) now become (after simplification)

$$X_t^2 = \frac{n_1 n_2 [\hat{F}_1(t) - \hat{F}_2(t)]^2}{n \hat{F}(t)[1 - \hat{F}(t)]} \tag{8.2}$$

and

$$G_t^2 = 2 \sum_{i=1}^{2} n_i \{ \, \hat{F}_i(t) \log \frac{\hat{F}_i(t)}{\hat{F}(t)} + [1 - \hat{F}_i(t)] \log \frac{1 - \hat{F}_i(t)}{1 - \hat{F}(t)} \, \} \,, \tag{8.3}$$

where $\hat{F}(t)$ and $\hat{F}_i(t)$ are respectively the empirical distribution functions of the pooled sample and sub-sample $X_{i1}, X_{i2}, \ldots, X_{in_i}$ (i=1, 2).

9. New Powerful Two-Sample Tests

Using (8.2) as Z_t in (8.1) with proper weight function, we can derive traditional nonparametric two-sample tests. For example, with $w(t) = \hat{F}(t)[1 - \hat{F}(t)]$, $dw(t) = \hat{F}(t)[1 - \hat{F}(t)] d\hat{F}(t)$ and $w(t) = \hat{F}(t)$ respectively, the first or second statistic in (8.1) generates the two-sample Kolmogorov-Smirnov statistic K_S, Cramér-von Mises statistic W^2 and Anderson-Darling statistic A^2 (Smirnov 1939; Massey 1951, 1952; Gnedenko 1954; Darling 1957; Hodges 1958; Anderson 1962; Burr 1963, 1964; Pittitt 1976; Cononver 1980; Epps and Singleton 1986; Scholz and Stephens 1987; Gibbons 1992; Baumgartner *et al.* 1998; Ferger 2000).

Using (8.3) as Z_t, on the other hand, we can produce new types of omnibus two-sample tests as follows. Just as the one-sample case, modifications are made to empirical distribution functions when necessary,. For instance, modifications to $\hat{F}(t)$ are made at its discontinuous points $X_{(k)}$ ($k = 1, 2, ..., n$) by defining $\hat{F}(X_{(k)}) = (k - 0.5)/n$.

1. $w(t) = 1$

Let $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$. Replacing Z_t of the second statistic in (8.1) with G_t^2 in (8.3) produces

$$\sup_{t\in(-\infty,\ \infty)} G_t^2 = \max_{0\le k\le n}\{\ \sup_{X_{(k)}\le t<X_{(k+1)}} G_t^2\ \} = \max_{1\le k\le n} G^2_{X_{(k)}}\ ,$$

which is equivalent to

$$Z_K = \max_{1\le k\le n}\left\{\ \sum_{i=1}^{2} n_i[\ F_{ik}\log\frac{F_{ik}}{F_k} + (1-F_{ik})\log\frac{1-F_{ik}}{1-F_k}\ ]\ \right\}\ , \tag{9.1}$$

where $F_k = \hat{F}(X_{(k)})$ and $F_{ik} = \hat{F}_i(X_{(k)})$ so that $F_k = (k - 0.5)/n$ and $F_{ik} = (j - 0.5)/n_i$ if $k = R_{ij}$ for some j, or $F_{ik} = j/n_i$ if $R_{ij} < k < R_{ij+1}$ ($R_{i0} = 1$, $R_{in_i+1} = n + 1$).

The large value of Z_K rejects the null hypothesis H.

2. $dw(t) = \hat{F}(t)^{-1}[1 - \hat{F}(t)]^{-1}d\hat{F}(t)$

Replacing Z_t of the first statistic in (8.1) with G_t^2 in (8.3) produces

$$\frac{2}{n}\sum_{k=1}^{n}\sum_{i=1}^{2}\frac{n_i}{F_k(1-F_k)}[\ F_{ik}\log\frac{F_{ik}}{F_k} + (1-F_{ik})\log\frac{1-F_{ik}}{1-F_k}\ ]\ ,$$

which is a decreasing function of

$$Z_A = -\sum_{k=1}^{n}\sum_{i=1}^{2} n_i\frac{F_{ik}\log F_{ik} + (1-F_{ik})\log(1-F_{ik})}{(k-0.5)(n-k+0.5)} \tag{9.2}$$

because $F_k = (k - 0.5)/n$ and $\sum_{i=1}^{2} n_i F_{ik} = nF_k$.

Hence, the small value of Z_A rejects the null hypothesis H.

3. $dw(t) = F(t)^{-1}[1 - F(t)]^{-1}dF(t)$

Here $F(t)$ is the common underlying distribution under the null hypothesis H. Replacing Z_t of the first statistic in (8.1) with G_t^2 in (8.3) produces

$$2\sum_{k=1}^{n}(b_{k-1} - b_k)\log[F(X_{(k)})^{-1} - 1] - 2\sum_{i=1}^{2}\sum_{j=1}^{n_i}(b_{ij-1} - b_{ij})\log[F(X_{i(j)})^{-1} - 1],$$

where $b_k = k\log(k/n) + (n-k)\log(1-k/n)$ and $b_{ij} = j\log(j/n_i) + (n_i - j)\log(1 - j/n_i)$.

Since $F(X_{(k)}) \approx \hat{F}(X_{(k)}) = (k - 0.5)/n$, $F(X_{i(j)}) \approx \hat{F}(X_{i(j)}) = (R_{ij} - 0.5)/n$ and $b_{ij-1} - b_{ij} \approx \log[n_i/(j - 0.5) - 1]$, the above statistic is (approximately) a decreasing function of

$$Z_C = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\log(\frac{n_i}{j - 0.5} - 1)\log(\frac{n}{R_{ij} - 0.5} - 1)\ , \tag{9.3}$$

the small value of which rejects the null hypothesis H.

The new statistics Z_K, Z_C and Z_A in (9.1)-(9.3) are analogues of the traditional two-sample Kolmogorov-Smirnov statistic K_S, Cramér-von Mises Statistic W^2 and Anderson-Darling statistic A^2. The sampling distributions of Z_K, Z_C and Z_A will be discussed in Section 11, and power comparisons between the new and old tests are given in Section 10.

10. Power Comparison for Two-Sample Tests

In this section we will compare the powers of new two-sample statistics Z_K, Z_C and Z_A with the traditional two-sample Kolmogorov-Smirnov statistic K_S, Cramér-von Mises Statistic W^2 and Anderson-Darling statistic A^2, as well as the parametric t- and F-tests which are optimal for detecting the differences of normal means and variances.

Since the new tests are generated by the likelihood-ratio statistic G_t^2 in (9.2), they should generally be powerful. Unfortunately, it is difficult to give a theoretical proof of some optimality for nonparametric tests with completely unknown distributions of the two samples. The general theory of (globally or locally) optimal tests assumes that we know the densities or at least the types of the densities about the underling distributions (e.g., Hájek and Šidák 1967, p.259 or Pratt and Gibbons, 1981).

Again Monte Carlo simulation is used to approximate the powers of associated tests. The approximations will approach the true values if the replicates of simulation can be sufficiently large. In the following examples, the replicates or simulation size is 10,000, and the significance level for rejecting H is $\alpha = 0.05$. For various situations about the distributions for null hypothesis H and the alternative $\bar{H}$, the simulated powers are exhibited with graphs, where the powers are plotted against the pooled sample size $n = n_1 + n_2$ for selected values of $(n_1,\ n_2)$: (10, 10), (10, 20), (20, 20), (20, 50), (50, 50), (50, 100), (100, 100), (100, 200), (200, 200).

Example 10.1: $U(0,1)$ vs. $Beta(p,q)$

Without loss of generality, we assume that the underlying distribution $F_1(x)$ of the first sample is a standard uniform $U(0,\ 1)$. Then the natural candidate for

$F_2(x)$, the underlying distribution of the second sample, is $Beta(p,\ q)$ which includes $U(0,\ 1)$. So, the two-sample test for $H:\ F_1 = F_2$ vs. $\hat{H}:\ F_1 \neq F_2$ is actually a parametric test for $H:\ (p,\ q) = (1,\ 1)$ vs. $\bar{H}:\ (p,\ q) \neq (1,\ 1)$.

For $(p,\ q)$=(0.5, 0.5), (0.5, 0.7), (0.7, 0.7), (1.5, 1.5), (1.5, 2) and (2, 2), the powers of statistics Z_A, Z_C, Z_K, A^2, W^2, K_S, t and F under alternative $\bar{H}$ are plotted in Fig. 10.1 respectively. We can see that Z_A and Z_C have the highest powers and dominate all others. Although Z_K is not as powerful as Z_A and Z_C, its power is still much higher than its analogue K_S. A^2 is the best among the three conventional nonparametric tests (this is also true for the next two examples). As anticipated, t-test can not detect the distribution difference in variances in the cases of $(p,\ q)$=(0.5, 0.5), (0.7, 0.7), (1.5, 1.5) and (2, 2). In fact, its power approximately equals the significance level $\alpha = 0.05$.

Since the difference in distribution (in terms of mean, variance, skewness and kurtosis, for example) between $U(0,\ 1)$ and $Beta(0.7,\ 0.7)$ or between $U(0,\ 1)$ and $Beta(1.5,\ 1.5)$ is less than it is for the other four cases, the overall power of all nonparametric statistics at $(p,\ q)$ =(0.5, 0.5) and (1.5, 1.5) is much lower compared to the other cases. Generally speaking, the larger the difference, the higher the power.

Example 10.2: $N(0,\ 1)$ vs. $N(\mu,\ \sigma^2)$

Because of the importance of normal distribution, $F_1(x)$ is assumed to be the standard normal distribution $N(0,\ 1)$ and $F_2(x)$ has a general normal distribution $N(\mu,\ \sigma^2)$. The two-sample test in this case is equivaient to testing $H:\ (\mu,\ \sigma^2) =$

$(0,\ 1)$ vs. $\bar{H}:\ (\mu,\ \sigma^2) \neq (0,\ 1)$.

Under normal assumptions, the distributions differ in mean and variance only. There is no shape difference in terms of skewness and kurtosis. That is why t- and F-tests are optimal for mean and variance shift models. However, they can not detect other forms of differences between distributions.

Fig. 10.2 compares the powers of the eight statistics at $(\mu,\ \sigma^2) = (0.3,\ 1)$, $(0.6,\ 1)$, $(0,\ 2)$, $(0,\ 3)$, $(0.3,\ 3)$, $(0.6,\ 2)$. It is obvious that when the main difference between F_1 and F_2 arises from their means rather than variances, say $(\mu,\ \sigma^2) = (0.3,\ 1)$, $(0.6,\ 1)$ or $(0.6,\ 2)$, there is little difference in power between the new nonparametric tests and the old ones. Conversely, for the other three cases, the advantage of the new tests over the old ones is obvious. In fact, the traditional tests are sensitive only to the difference in location or mean, but are dull to detect the variation in scale or shape (see also Examples 10.1 and 10.3). When the difference in distribution arises from locations only, such as the cases of $(\mu,\ \sigma^2) = (0.3,\ 1)$ and $(0.6,\ 1)$, all the nonparametric tests are almost as powerful as the optimal t-test. In the cases of $(\mu,\ \sigma^2) = (0,\ 2)$ and $(0,\ 3)$ when the only difference in distribution comes from scales, the power lost by using the new tests over the optimal F-test is much less than that by using the old ones.

For cases $N(0,\ 1)$ vs. $N(0,\ 2)$ and $N(0,\ 1)$ vs. $N(0,\ 3)$, the difference between F_1 and F_2 arises from their variances only. Since the former case has less difference, it has lower power and lower speed of convergence for the all statistics except t. This is also true for cases $N(0,\ 1)$ vs. $N(0.3,\ 1)$ and $N(0,\ 1)$ vs. $N(0.6,\ 1)$ where the difference between F_1 and F_2 comes from their means only. For the other two

cases, the difference between F_1 and F_2 arises from both their means and variances, so the power of the tests is higher and it converges faster.

We can also consider a more general case where F_2 has a symmetric distribution, such as Cauchy, logistic and t distributions, but the results are similar.

Example 10.3: $N(\mu,\ \sigma^2)$ vs. $Gamma(r,\ 1)$

In this example $F_1(x)$ is also assumed to be the normal distribution $N(\mu,\ \sigma^2)$, but $F_2(x)$ has an asymmetric distribution, say $Gamma(r,\ 1)$, the gamma distribution with shape parameter r and scale parameter 1. Six cases are considered: (1) $N(3,\ 2)$ vs. $Gamma(3,\ 1)$; (2) $N(2,\ 3)$ vs. $Gamma(3,\ 1)$; (3) $N(3,\ 3)$ vs. $Gamma(3,\ 1)$; (4) $N(7,\ 5)$ vs. $Gamma(7,\ 1)$; (5) $N(5,\ 7)$ vs. $Gamma(7,\ 1)$; (6) $N(5,\ 5)$ vs. $Gamma(7,\ 1)$. Plots of the power using for the eight statistics are shown in Fig. 10.3. Note that in cases 1 and 4, F_1 and F_2 have the same mean but different variances, and in cases 2 and 5, they have different means but the same variance. On the other hand, means and variances are the same in case 3 but both different in case 6.

Again, when the major difference between F_1 and F_2 arises from their means, such as cases 5 and 6, the difference in power between the new and old nonparametric tests is not so significant. Otherwise, the power improvement of the new tests on the old ones is tremendous, especially for case 3, where the difference between F_1 and F_2 comes purely from the their shapes so that the improvement can be significant. Just as other examples, the t-test (F-test) fails to detect the difference in variances (means). Moreover, both t- and F-tests are failed in case 3.

Note that case 1 has higher power for the all tests (excluding t-test) than case 4 because it has larger difference in shape between F_1 and F_2 in terms of variance, skewness and kurtosis.

Other asymmetric distributions, such as log-normal, Weibull, F and Beta, are also considered as the distribution of F_2. The situations are much similar to that of gamma.

It can be seen from Examples 10.1-10.3 that the traditional two-sample tests are only sensitive to the difference in locations or means between the underlying distributions of the two samples, but are dull to detect the variation in their shapes. This fact is well known for conventional rank tests, such as the Wilcoxon test or Mann-Whitney test (see Conover, 1980; Pratt and Gibbons, 1981; Gibbons, 1992), but there is no a major breakthrough yet in finding an omnibus test which is very sensitive to both location and shape differences. For example, a new test given by Baumgartner *et al.* (1998) is actually the Anderson-Darling test A^2, which is the best existing test but is still poor compared with Z_A and Z_C. Ferger (2000) introduced some new tow-sample tests based on the so-called change-point model. The tests are applicable to multivariate data, but the simulation results reported (Ferger 2000, p.28-30) show that compared with the Kolmogorov-Smirnov test K_S, which is the least powerful among all tests we discussed, Ferger's tests are less powerful when detecting location difference even though they are more sensitive to scale variation.

In this aspect, the new tests Z_A, Z_C and Z_K make great contributions. In fact, if the two samples have the same shape but different locations or means, the new tests are as powerful as the old ones. Otherwise, they are much better in terms of

power.

Z_K is not so powerful as Z_A and Z_C, but it is much better than its analogue K_S. Finally, Z_A and Z_C are almost equivalent, but Z_C is most recommended because it has a simple and elegant representation in (9.3).

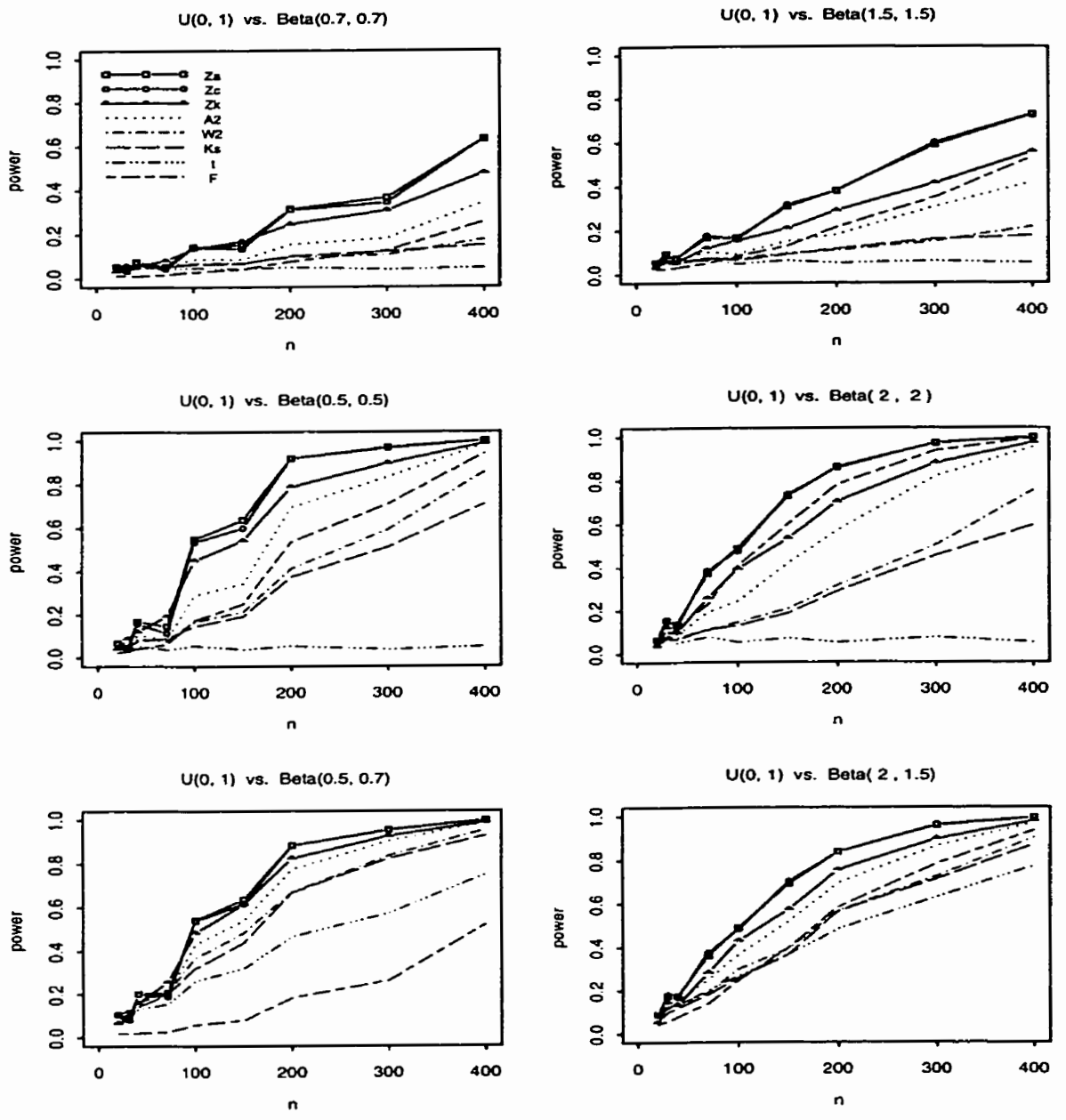

Fig. 10.1. Power comparison for testing $U(0,1)$ vs $Beta(p,q)$ at level $\alpha = 0.05$

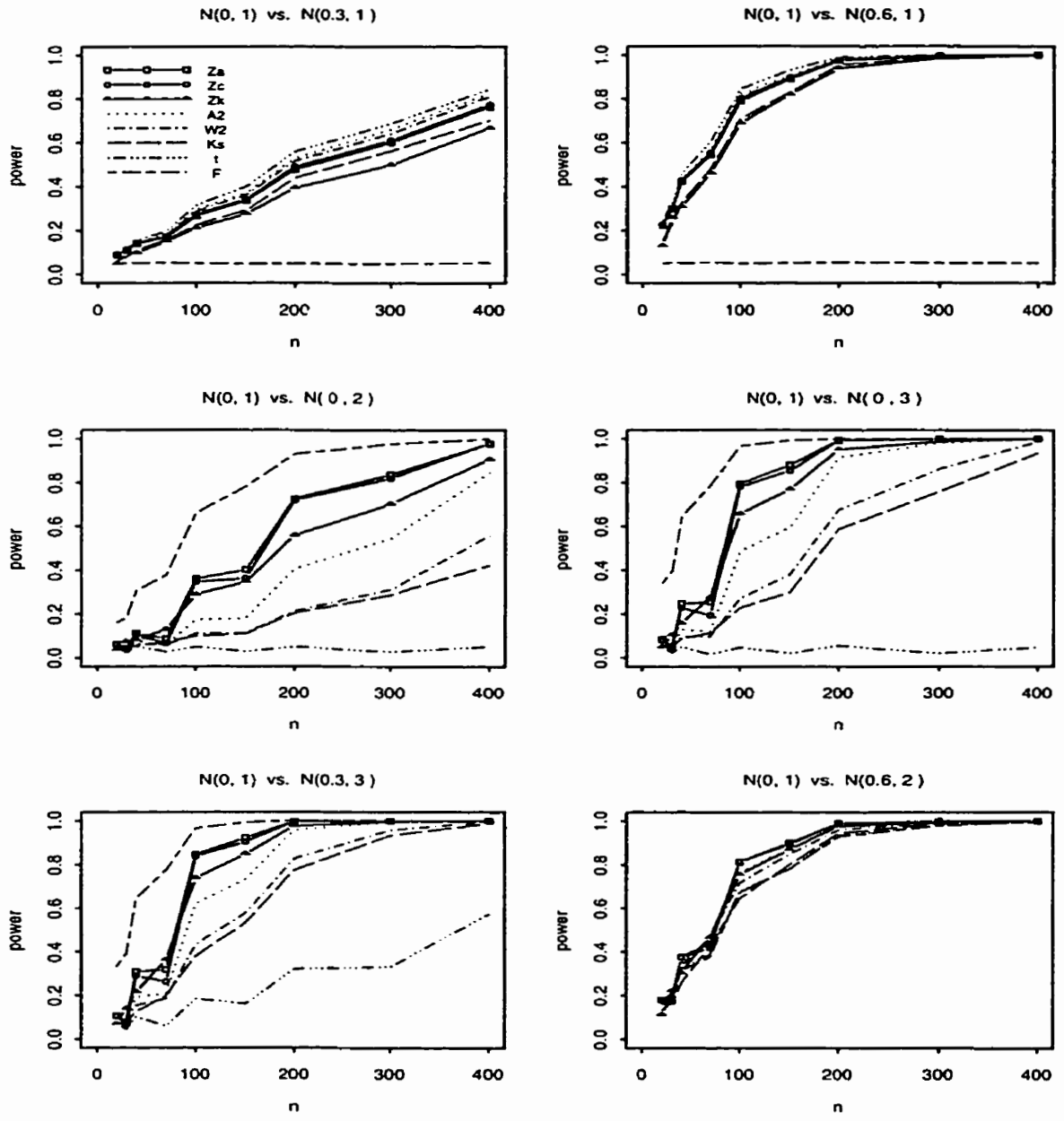

Fig. 10.2. Power comparison for testing $N(0,\ 1)$ vs $N(\mu,\ \sigma^2)$ at level $\alpha = 0.05$

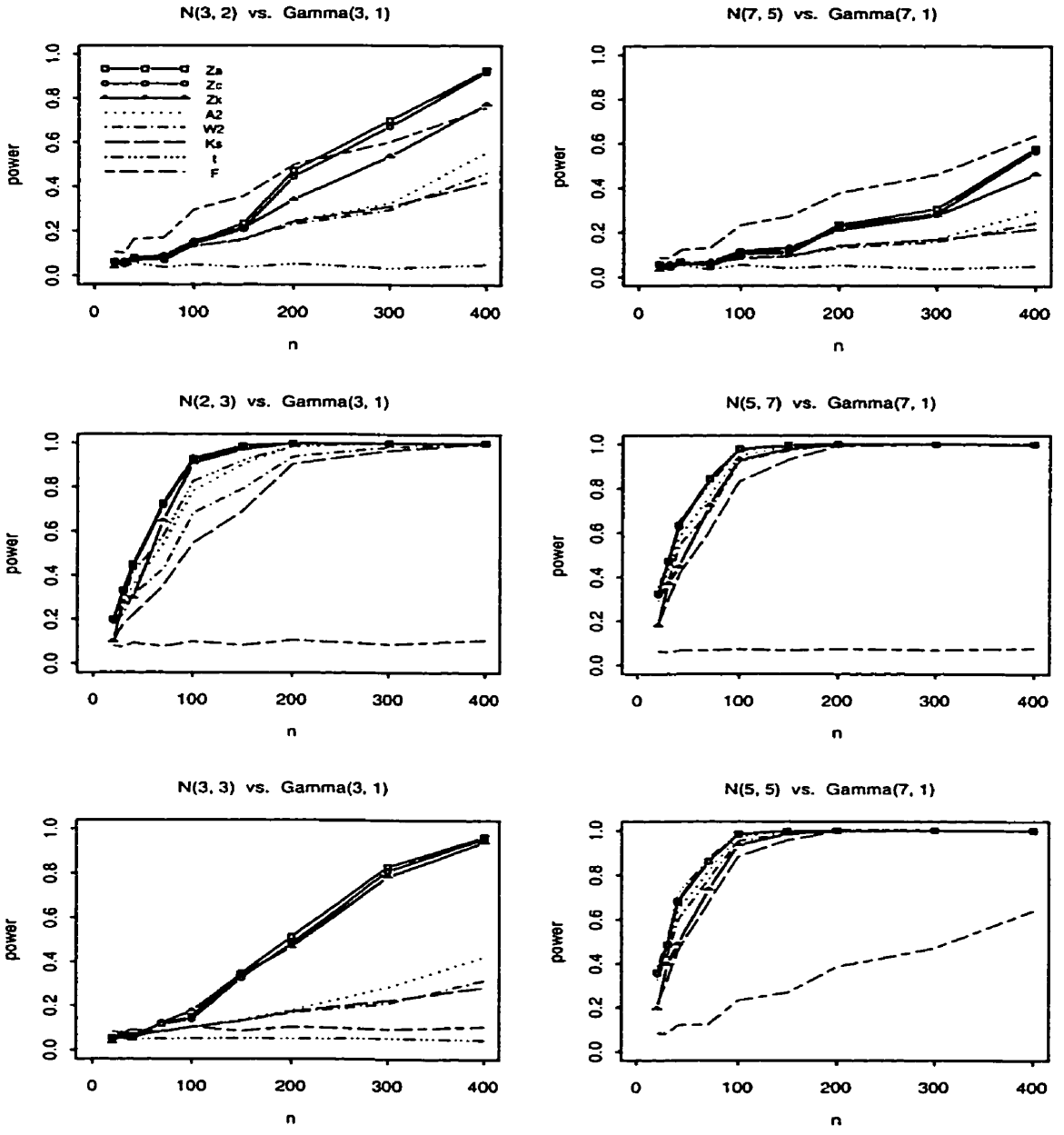

Fig. 10.3. Power comparison for testing $N(\mu,\ \sigma^2)$ vs $Gamma(r,\ 1)$ at level $\alpha = 0.05$

11. The Distributions of Two-Sample Z_A, Z_C and Z_K

Like traditional two-sample test statistics A^2, W^2 and K_S, the statistics Z_A, Z_C and Z_K are distribution-free. Their null distributions are discrete and uniform. Therefore, they can be obtained by enumeration of all possible values of the statistics by considering the $n!/(n_1!n_2!)$ combinations of the ranks for the first sample. In fact, Z_A, Z_C and Z_K are only the functions of R_{11}, R_{12}, ..., R_{1n_1}. Under the null hypothesis H, every combination of n_1 integers from 1, 2, ..., n is equally likely to be the ranks of the first sample.

Since the number of values that each of the statistics can take on increases very rapidly with n_1 and n_2, it is not feasible to give the full distribution unless both n_1 and n_2 are small. Moreover, extensive tables have to be used for tabulating their percentage points with different n_1 and n_2. Thus, the exact quantiles of W^2 and A^2 are available only for small sample sizes (see Anderson 1962; Burr 1963, 1964; Pettitt 1976).

Instead of tabulating limited percentage points for the new statistics, we can use Monte Carlo approach to get their approximate p-values for the two- sample test. In fact, whenever we do a significance test, we do not need the exact p-value of the test (Actually, it is impossible to find the exact p-value for a real test unless all assumptions about the model are 100 percent true and there is no rounding error). An approximate but reasonably accurate p-value is often sufficient for the purpose of statistical inference. With today's computing facilities and software, it is easy to approximate the p-value using Monte Carlo simulation with 5,000 or 10,000

replicates.

Computer programs in Splus code (Programs 1-3) are given below to calculate each new statistic and its simulated p-value for the two-sample test. In Program 1-3, N is simulation size, while $X1$ and $X2$ are vectors of data for the first and second samples, i.e., $X1 = (x_{11},\ x_{12},\ ...,\ x_{1n_1})$ and $X2 = (x_{21},\ x_{22},\ ...,\ x_{1n_2})$.

Program 1. Calculating Z_C and its p-value (two-sample case)

```
 f <- function(X1,X2,N) {
n1 <- length(X1)
n2 <- length(X2)
 R <- rank(c(X1, X2))
 n <- n1+n2
 S <- 0
 g <- function(m,r,M) sum(log(m/(1:m-.5)-1)*log(M/(r-.5)-1))
Zc <- (g(n1, sort(R[1:n1]), n)+g(n2, sort(R[(n1+1):n]), n))/n
      for (j in 1:N) {
       R <- sample(n)
      zc <- (g(n1, sort(R[1:n1]), n)+g(n2, sort(R[(n1+1):n]), n))/n
       S <- S + (zc < Zc) }
 p.value <- S/N
      return(Zc, p.value)  }
```

Program 2. Calculating Z_A and its p-value (two-sample case)

```
 f <- function(X1,X2,N) {
n1 <- length(X1)
n2 <- length(X2)
 R <- ceiling(rank(c(X1,X2)))
 n <- n1+n2
 w <- (1:n-.5)*(n:1-.5)
 g <- function(m,r,M) {
      d <- sort(r)
      D <- c(1,d,M+1)
      p <- rep(0:m, D[2:(m+2)]-D[1:(m+1)])
   p[d] <- p[d]-.5
      p <- p/m
      m*(p*log(p+.0000000001)+(1-p)*log(1-p+.0000000001)) }
Za <- -sum((g(n1, R[1:n1], n)+g(n2, R[(n1+1):n], n))/w)
 S <- 0
      for (j in 1:N) {
       R <- sample(n)
      za <- -sum((g(n1, R[1:n1], n)+g(n2, R[(n1+1):n], n))/w)
       S <- S + (za < Za) }
 p.value <- S/N
```

```
      return( Za, p.value)  }
```

Program 3. Calculating Z_K and its p-value (two-sample case)

```
 f <- function(X1,X2,N) {
n1 <- length(X1)
n2 <- length(X2)
 R <- ceiling(rank(c(X1,X2)))
 n <- n1+n2
 P <- (1:n-.5)/n
 w <- n*(P*log(P)+(1-P)*log(1-P))
 g <- function(m,r,M) {
      d <- sort(r)
      D <- c(1,d,M+1)
      p <- rep(0:m, D[2:(m+2)]-D[1:(m+1)])
      p[d] <- p[d]-.5 ; p <- p/m
      m*(p*log(p+.0000000001)+(1-p)*log(1-p+.0000000001)) }
Zk <- max( g(n1, R[1:n1], n) + g(n2, R[(n1+1):n], n) - w )
 S <- 0
      for (j in 1:N) {
       R <- sample(n)
      zk <- max( g(n1, R[1:n1], n) + g(n2, R[(n1+1):n], n) - w )
```

```
      S <- S + (zk > Zk) }
p.value <- S/N
     return( Zk, p.value)  }
```

These programs are easy to run even on PC. Program 1 requires only two minutes to run on PC (Pentium II-MMX CPU at 300MHz) for a 10,000-sized simulation of test Z_C with n_1=200 and n_2=300.

An Illustration

Consider an example from an accelerated life test experiment. The data from Table 11.1 (Nair 1984, p.824) are times to breakdown of an insulating fluid under two elevated voltage stresses of 32 Kv and 36 Kv. Hall and Padmanabhan (1997) use the data as an illustrative example for the two-sample problem. We wish to test H: the two sampled populations have the same probability distribution.

Using Programs 1-3, we apply the new two-sample tests of Z_A, Z_C and Z_K to the data respectively. We find that Z_A=2.9048, Z_C=2.6245 and Z_K=4.4349 with corresponding p-values (simulated with 10,000 replicates): 0.0318, 0.0305, 0.0227.

On the other hand, if we use the classical Kolmogorov-Smirnov two-sample test K_S (the associated function in Splus is ks.gof), we have $K_S = 0.4667$ with p-value = 0.0755. Obviously, the new tests give smaller p-values to reject the null hypothesis H.

Table 11.1. Times (in Minutes) to Breakdown of an Insulating Fluid

32 Kv	.27	.40	.69	.79	2.75	3.91	9.88	13.95
	15.93	27.80	53.24	82.85	89.29	100.58	215.10	
36 Kv	.35	.59	.96	.99	1.69	1.97	2.07	2.58
	2.71	2.90	3.67	3.99	5.35	13.77	25.50	

12. General k-Sample Problem

We now generalize the powerful two-sample tests Z_K, Z_A and Z_C in (9.1)-(9.3) into multi-sample cases.

Let X_{i1}, X_{i2}, ..., X_{in_i} be a random sample from a continuous population with distribution function $F_i(x)$, i=1, 2, ..., k ($k \geq 2$), and let X_1, X_2, ..., X_n ($n = n_1 + \cdots + n_k$) be the pooled sample with order statistics $X_{(1)}$, $X_{(2)}$, ..., $X_{(n)}$. Denote R_{ij} the rank of the j-th ordered observation $X_{i(j)}$ in the pooled sample. We wish to test the null hypothesis

$$H: \ F_1(x) = F_2(x) = \cdots = F_k(x), \quad \texttt{for all } x \in (-\infty, \infty)$$

without specifying the common distribution function $F(x)$. Since

$$H = \bigcap_{t \in (-\infty, \infty)} H_t \quad \texttt{where} \quad H_t: \ F_1(t) = F_2(t) = \cdots = F_k(t) \ ,$$

testing H is equivalent to testing H_t for every $t \in (-\infty, \infty)$.

To test H_t with t fixed, we define new samples based on index function: $X_{ijt} = I(X_{ij} \leq t)$ ($i = 1, 2, ..., k$; $j = 1, 2, ..., n_i$) satisfying $P(X_{ijt} = 1) = F_i(t)$ and $P(X_{ijt} = 0) = 1 - F_i(t)$.

For each fixed $t \in (-\infty, \infty)$ and the corresponding random samples X_{i1t}, X_{i2t}, ..., X_{in_it} (i=1, 2, ..., k), let Z_t be a statistic for testing H_t such that its large values reject H_t. Similarly, two types of statistics for testing H can be defined (8.1), but the Pearson's chi-squared test statistic in (8.2) and the likelihood-ratio test statistic in (8.3) are generalized as

$$X_t^2 = \sum_{i=1}^{k} n_i \frac{[\hat{F}_i(t) - \hat{F}(t)]^2}{\hat{F}(t)[1 - \hat{F}(t)]}$$

and

$$G_t^2 = 2\sum_{i=1}^{k} n_i \{ \hat{F}_i(t)\log\frac{\hat{F}_i(t)}{\hat{F}(t)} + [1 - \hat{F}_i(t)]\log\frac{1 - \hat{F}_i(t)}{1 - \hat{F}(t)} \} ,$$

where $\hat{F}(t)$ and $\hat{F}_i(t)$ are respectively the empirical distribution functions of the pooled sample and sub-sample X_{i1}, X_{i2}, ..., X_{in_i} (i=1, 2, ..., k).

Using the X_t^2 as Z_t in (8.1) but choosing different weight functions, we can derive the following traditional k-sample tests.

1. $w(t) = \hat{F}(t)[1 - \hat{F}(t)]$

Replacing Z_t of the second statistic in (8.1) with X_t^2 generates

$$K_S = \sup_{-\infty<t<\infty} \sum_{i=1}^{k} n_i[\hat{F}_i(t) - \hat{F}(t)]^2 , \qquad (12.1)$$

which is a k-sample version of the traditional Kolmogorov-Smirnov statistic (Kiefer 1959). For other k-sample versions of Kolmogorov-Smirnov tests, see Conover (1965, 1980) and Wolf and Naus (1973).

2. $w(t) = \hat{F}(t)$

Replacing Z_t of the first statistic in (8.1) with X_t^2 generates the k-sample Anderson-

Darling statistic (Scholz and Stephens 1987)

$$A^2 = \sum_{i=1}^{k} n_i \int_{-\infty}^{\infty} \frac{[\hat{F}_i(t) - \hat{F}(t)]^2}{\hat{F}(t)[1 - \hat{F}(t)]} \, d\hat{F}(t) \; . \tag{12.2}$$

3. $dw(t) = \hat{F}(t)[1 - \hat{F}(t)]d\hat{F}(t)$

Replacing Z_t of the first statistic in (8.1) with X_t^2 generates the k-sample Cramér-von Mises Statistic (Kiefer, 1959)

$$W^2 = \sum_{i=1}^{k} n_i \int_{-\infty}^{\infty} [\hat{F}_i(t) - \hat{F}(t)]^2 \, d\hat{F}(t) \; . \tag{12.3}$$

Next we will derive the new k-sample tests by using above G_t^2 as Z_t in (8.1).

13. New k-Sample Tests

The new two-sample tests in (9.1)-(9.3) can be generalized as follows.

1. $w(t) = 1$

Replacing Z_t of the second statistic in (8.1) with G_t^2 in Section 12 produces

$$\sup_{-\infty<t<\infty} G_t^2 = \max_{1\le m\le n} G^2_{X_{(m)}} \; ,$$

which is equivalent to

$$Z_K = \max_{1\le m\le n} \left\{ \sum_{i=1}^{k} n_i [\, F_{im} \log \frac{F_{im}}{F_m} + (1 - F_{im}) \log \frac{1 - F_{im}}{1 - F_m} \,] \right\} \; , \tag{13.1}$$

where $F_m = \hat{F}(X_{(m)})$ and $F_{im} = \hat{F}_i(X_{(m)})$ so that $F_m = (m - 0.5)/n$ and $F_{im} = (j - 0.5)/n_i$ if $m = R_{ij}$ for some j, or $F_{im} = j/n_i$ if $R_{ij} < m < R_{ij+1}$ ($R_{i0} = 1$, $R_{in_i+1} = n + 1$).

The large value of Z_K rejects the null hypothesis H.

2. $dw(t) = \hat{F}(t)^{-1}[1 - \hat{F}(t)]^{-1}d\hat{F}(t)$

Replacing Z_t of the first statistic in (8.1) with G_t^2 in Section 12 produces

$$\frac{2}{n}\sum_{m=1}^{n}\sum_{i=1}^{k}\frac{n_i}{F_m(1-F_m)}[\ F_{im}\log\frac{F_{im}}{F_m} + (1-F_{im})\log\frac{1-F_{im}}{1-F_m}\ ]\ ,$$

which is a decreasing function of

$$Z_A = -\sum_{m=1}^{n}\sum_{i=1}^{k} n_i \frac{F_{im}\log F_{im} + (1-F_{im})\log(1-F_{im})}{(m-0.5)(n-m+0.5)}\ , \tag{13.2}$$

because $F_m = (m-0.5)/n$ and $\sum_{i=1}^{k} n_i F_{im} = nF_m$.

Small values of Z_A reject the null hypothesis H.

3. $dw(t) = F(t)^{-1}[1 - F(t)]^{-1}dF(t)$

Replacing Z_t of the first statistic in (8.1) with G_t^2 in Section 12 produces

$$2\sum_{m=1}^{n}(b_{m-1} - b_m)\log[F(X_{(m)})^{-1} - 1] - 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(b_{ij-1} - b_{ij})\log[F(X_{i(j)})^{-1} - 1],$$

where $b_m = m\log(m/n) + (n-m)\log(1-m/n)$ and $b_{ij} = j\log(j/n_i) + (n_i - j)\log(1-j/n_i)$.

Since $F(X_{(m)}) \approx \hat{F}(X_{(m)}) = (m-0.5)/n$, $F(X_{i(j)}) \approx \hat{F}(X_{i(j)}) = (R_{ij}-0.5)/n$ and $b_{ij-1} - b_{ij} \approx \log[n_i/(j-0.5) - 1]$, the above statistic is (approximately) a decreasing function of

$$Z_C = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\log(\frac{n_i}{j-0.5} - 1)\log(\frac{n}{R_{ij}-0.5} - 1)\ , \tag{13.3}$$

small values of which reject the null hypothesis H.

14. Power Comparison for k-Sample Tests

For two-sample tests (k=2), our simulation show that the powers of Z_K, Z_A and Z_C in (9.1)-(9.3) are generally much higher than those of the traditional tests, which are location-sensitive only.

In this section we will give some examples of three-sample tests (k=3), which, together with the examples of two-sample tests in Section 10, are illustrations for general k-sample tests. The powers of Z_K, Z_C and Z_A in (13.1)-(13.3) are compared with the conventional k-sample tests K_S, A^2 and W^2 in (12.1)-(12.3), as well as the Kruskal-Wallis test K_W (Kruskal and Wallis, 1952), which is a common used non-parametric k- sample test. The asymptotic relative efficiency of K_W to the one-way F test is $3/\pi$=0.955 under the condition of normality, but it is usually greater than one in the case of non-normality, say 1.5 for exponential distributions (see Conover, 1980; Gibbons, 1992).

For the following situations about the underlying distributions under the null and alternative hypotheses, the powers of the seven statistics are approximated by Monte Carlo simulation, where the number of replicates is 10,000, and the significance level for testing H is $\alpha = 0.05$.

The simulated powers are plotted against the total sample size $n = n_1 + n_2 + n_3$ for selected values of (n_1, n_2, n_3): (10, 10, 10), (10, 10, 20), (10, 20, 20), (20, 20, 20), (20, 20, 50), (20, 50, 50), (50, 50, 50), (50, 50, 100), (50, 100, 100), (100, 100, 100).

Example 14.1: $F_1 = U(0,\ 1)$ and $F_i = Beta(p_i,\ q_i)$ (i=2, 3)

Without loss of generality, we assume that the underlying distribution F_1 of the first sample is the standard uniform $U(0,\ 1)$. Then the natural candidates for F_2 and F_3, the underlying distributions of the second and third samples, belong to the family of beta distribution $Beta(p,\ q)$, which include the uniform one because of $Beta(1,\ 1) = U(0,\ 1)$. Assuming $F_i = Beta(p_i,\ q_i)$ (i=2, 3), we can see that the three-sample test for $H :\ F_1 = F_2 = F_3$ is actually a parametric test for $H :\ (p_i,\ q_i) = (1,\ 1)$ (i=2, 3).

For the following cases about the alternative hypothesis:

(1) $F_1 = U(0,\ 1)$, $F_2 = Beta(0.7,\ 0.7)$ and $F_3 = Beta(0.5,\ 0.5)$

(2) $F_1 = U(0,\ 1)$, $F_2 = Beta(\ 1\ ,\ 0.7)$ and $F_3 = Beta(0.7,\ 0.5)$

(3) $F_1 = U(0,\ 1)$, $F_2 = Beta(\ 1\ ,\ 0.7)$ and $F_3 = Beta(0.7,\ 1\ )$

(4) $F_1 = U(0,\ 1)$, $F_2 = Beta(\ 1\ ,\ 1.5)$ and $F_3 = Beta(1.5,\ 1\ )$

(5) $F_1 = U(0,\ 1)$, $F_2 = Beta(\ 1\ ,\ 1.5)$ and $F_3 = Beta(1.5,\ 2\ )$

(6) $F_1 = U(0,\ 1)$, $F_2 = Beta(1.5,\ 1.5)$ and $F_3 = Beta(\ 2\ ,\ 2\ )$

the powers of Z_A, Z_C, Z_K, A^2, W^2, K_S and K_W are plotted in Fig. 14.1 respectively. We can see that Z_A and Z_C have the highest powers and dominate the others. Although not as powerful as Z_A and Z_C, Z_K is overwhelming compared to its analogue K_S. Besides, among A^2, W^2 and K_S, A^2 is the best and K_S is the worst (this is

also true for other examples). Finally, in cases 1 or 6 where F_i (i=1, 2, 3) have the same location (mean) but different shapes, the new statistics are much powerful than the old ones. In such a case, the conventional Kruskal-Wallis test K_W totally fails because its power is almost equal to $\alpha = 0.05$, the significance level of the test.

Example 14.2: $F_1 = N(0,\ 1)$ and $F_i = N(\mu_i,\ \sigma_i^2)$ (i=2, 3)

Because of the importance of normal distribution, F_i (i=1, 2, 3) are assumed to be normal distributions $N(\mu_i,\ \sigma_i^2)$. We can assume that $F_1 = N(0,\ 1)$ without loss of generality. Then testing $H:\ F_1 = F_2 = F_3$ is equivalent to testing $H:\ (\mu_i,\ \sigma_i^2) = (0,\ 1)$ (i=2, 3).

Fig. 14.2 compares the powers of the seven statistics for the following situations about the alternative hypothesis:

(1) $F_1 = N(0,\ 1)$, $F_2 = N(\ 0\ ,\ \ 2\ )$ and $F_3 = N(\ 0\ ,\ \ 4\ )$

(2) $F_1 = N(0,\ 1)$, $F_2 = N(\ 0\ ,\ \ 2\ )$ and $F_3 = N(\ 0\ ,\ 0.5)$

(3) $F_1 = N(0,\ 1)$, $F_2 = N(0.3,\ \ 1\ )$ and $F_3 = N(0.6,\ \ 1\ )$

(4) $F_1 = N(0,\ 1)$, $F_2 = N(0.6,\ \ 1\ )$ and $F_3 = N(\ 1\ ,\ \ 1\ )$

(5) $F_1 = N(0,\ 1)$, $F_2 = N(0.3,\ 0.5)$ and $F_3 = N(0.6,\ \ 2\ )$

(6) $F_1 = N(0,\ 1)$, $F_2 = N(0.6,\ 0.5)$ and $F_3 = N(0.3,\ \ 2\ )$.

It is obvious that in cases 3-4 where F_i (i=1, 2, 3) have different locations (means)

but the same dispersion (variance), there is little difference of powers between the new statistics and the old ones. Conversely, in cases 1-2 where F_i have the same location but different dispersions, the power improvements of the new statistics on the old ones are tremendous. In such cases, the Kruskal-Wallis test K_W has almost 'no power' (see also other examples). Finally, in cases 5-6 where F_i have different locations and dispersions, the advantage of the new statistics is still significant.

We can also consider a more general case where $F_1 = N(0, 1)$ and F_i (i=2, 3) have symmetric distributions, such as Cauchy, logistic and t distributions, but the results are similar according to our simulation.

Example 14.3: $F_1 = N(\mu, \sigma^2)$ and $F_i = Gamma(a_i, b_i)$ (i=2, 3)

In this example $F_1(x)$ is also assumed to be a normal distribution $N(\mu, \sigma^2)$, but $F_i(x)$ (i=2, 3) have non-symmetric distributions, say $Gamma(a_i, b_i)$, the gamma distributions with shape parameter a_i and scale parameter b_i.

Six cases about the alternative hypothesis are considered as follows:

(1) $F_1 = N(3, 1)$, $F_2 = Gamma(3, 1)$ and $F_3 = Gamma(6, 2)$

(2) $F_1 = N(5, 1)$, $F_2 = Gamma(5, 1)$ and $F_3 = Gamma(10, 2)$

(3) $F_1 = N(2, 3)$, $F_2 = Gamma(3, 1)$ and $F_3 = Gamma(5, 1.3)$

(4) $F_1 = N(6, 5)$, $F_2 = Gamma(5, 1)$ and $F_3 = Gamma(10, 1.4)$

(5) $F_1 = N(2, 2)$, $F_2 = Gamma(3, 1)$ and $F_3 = Gamma(5, 2)$

(6) $F_1 = N(3, 3)$, $F_2 = Gamma(5, 1)$ and $F_3 = Gamma(8, 2)$.

Note that F_i (i=1, 2, 3) have the same mean but different variances in cases 1-2, different means but approximately the same variance in cases 3-4, different means and variances in cases 5-6. Power comparisons are given in Fig. 14.3, the results are almost the same as those in Example 14.2.

Similar results can be obtained if we consider such cases where $F_1 = N(0, 1)$ and F_i (i=2, 3) have other non-symmetric distributions, such as log- normal, Weibull and F distributions.

Example 14.4: $F_1 = N(0.5, 0.1)$ and $F_i = Beta(p_i, q_i)$ (i=2, 3)

In the last example, $F_1 = N(0.5, 0.1)$ but $F_i = Beta(p_i, q_i)$ (i=2, 3), where F_2 is symmetric ($p_2 = q_2$) while F_3 is non-symmetric ($p_3 \neq q_3$). The following situations are considered:

(1) $F_1 = N(0.5, 0.1)$, $F_2 = Beta(2, 2)$ and $F_3 = Beta(2, 2.5)$

(2) $F_1 = N(0.5, 0.1)$, $F_2 = Beta(2, 2)$ and $F_3 = Beta(2, 1.5)$

(3) $F_1 = N(0.5, 0.1)$, $F_2 = Beta(1.5, 1.5)$ and $F_3 = Beta(1.5, 2)$

(4) $F_1 = N(0.5, 0.1)$, $F_2 = Beta(1.5, 1.5)$ and $F_3 = Beta(1.5, 1)$

(5) $F_1 = N(0.5, 0.1)$, $F_2 = Beta(1, 1)$ and $F_3 = Beta(1, 1.5)$

(6) $F_1 = N(0.5, 0.1)$, $F_2 = Beta(1, 1)$ and $F_3 = Beta(1, 0.5)$.

The powers of the seven statistics are exhibited in Fig. 14.4, which shows that the new statistics are much more powerful than the old ones.

We can see from these examples that the traditional k-sample tests are sensitive to location differences among F_i (i=1, 2, ..., k), but are dull to detect the variations of their shapes. As a result, it seems that

(a) when the differences among F_i arrive from their locations or means only, the new statistics are as powerful as the traditional ones with Z_A, Z_C, A^2 and K_W being better;

(b) when the differences among F_i come not only from their locations or means but also from their shapes (including dispersions), the new statistics are much more powerful than the old ones; In such a case, Z_A and Z_C dominate the others, Z_K or A^2 is the second best, and K_W or K_S is the worst;

(c) Z_K is overwhelmingly powerful compared to its analogue K_S;

(d) Among A^2, W^2 and K_S, the traditional k-sample tests based on EDF (empirical distribution function), A^2 and K_S are respectively the best and worst in all situations;

(e) the Kruskal-Wallis test K_W fails to detect the difference in shapes among F_i.

(f) Z_A and Z_C are almost equivalent, but Z_C is the best because of its simple representation in (13.3).

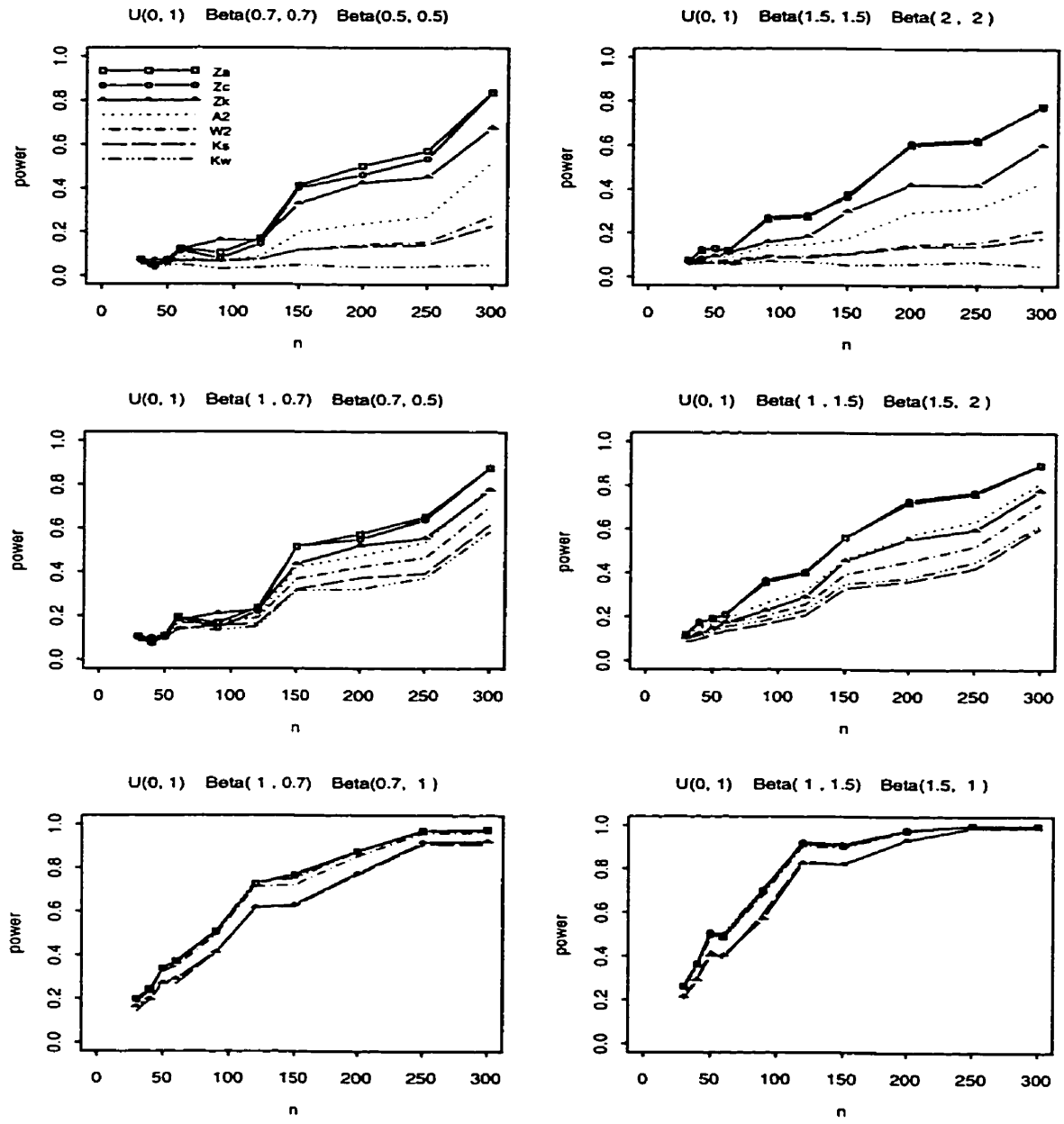

Fig. 14.1. Power comparison for testing $F_1 = U(0,\ 1)$ and $F_i = Beta(p_i,\ q_i)$ (i=2, 3) at level $\alpha = 0.05$

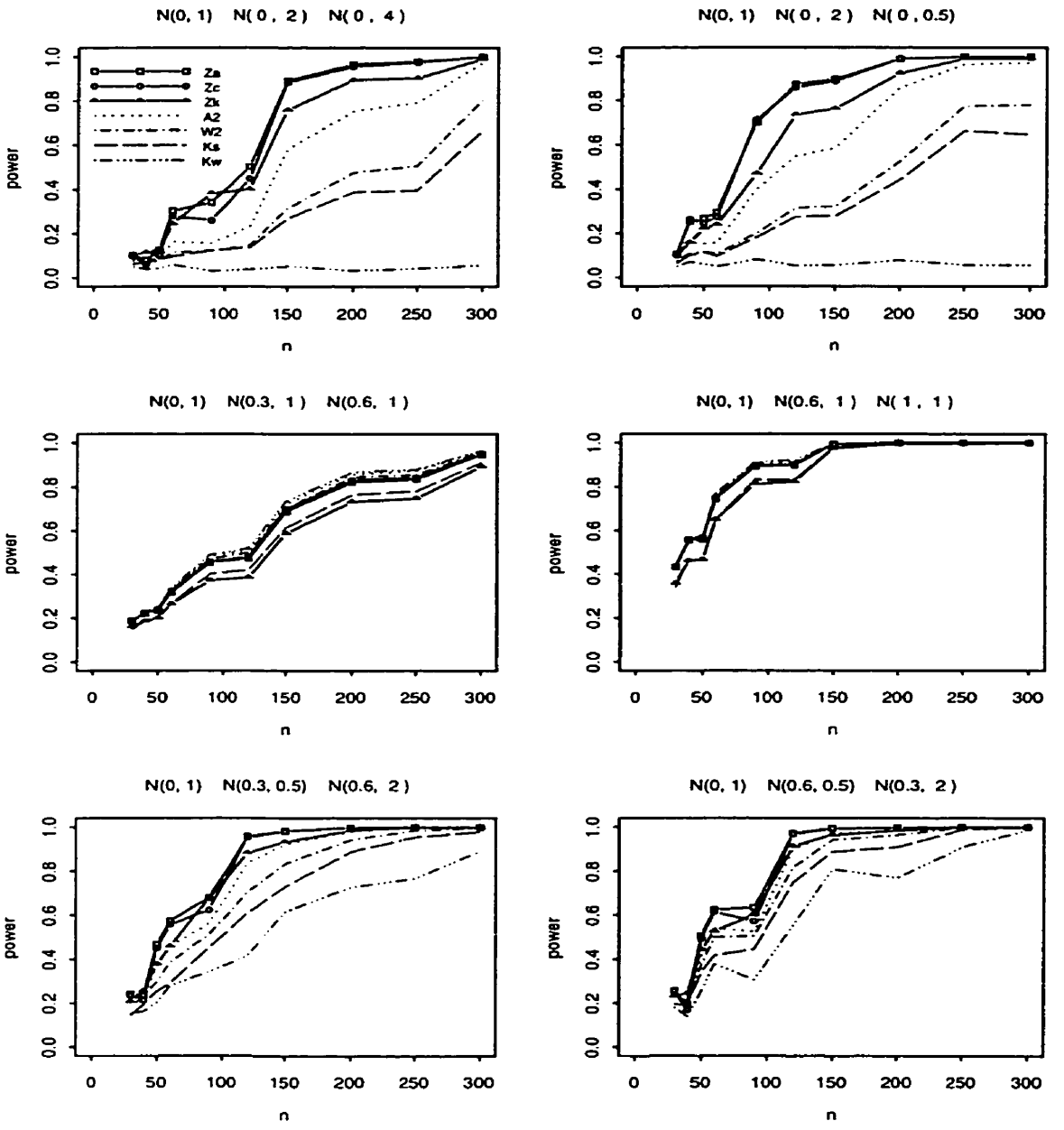

Fig. 14.2. Power comparison for testing $F_1 = N(0,\ 1)$ and $F_i = N(\mu_i,\ \sigma_i^2)$ (i=2, 3) at level $\alpha = 0.05$

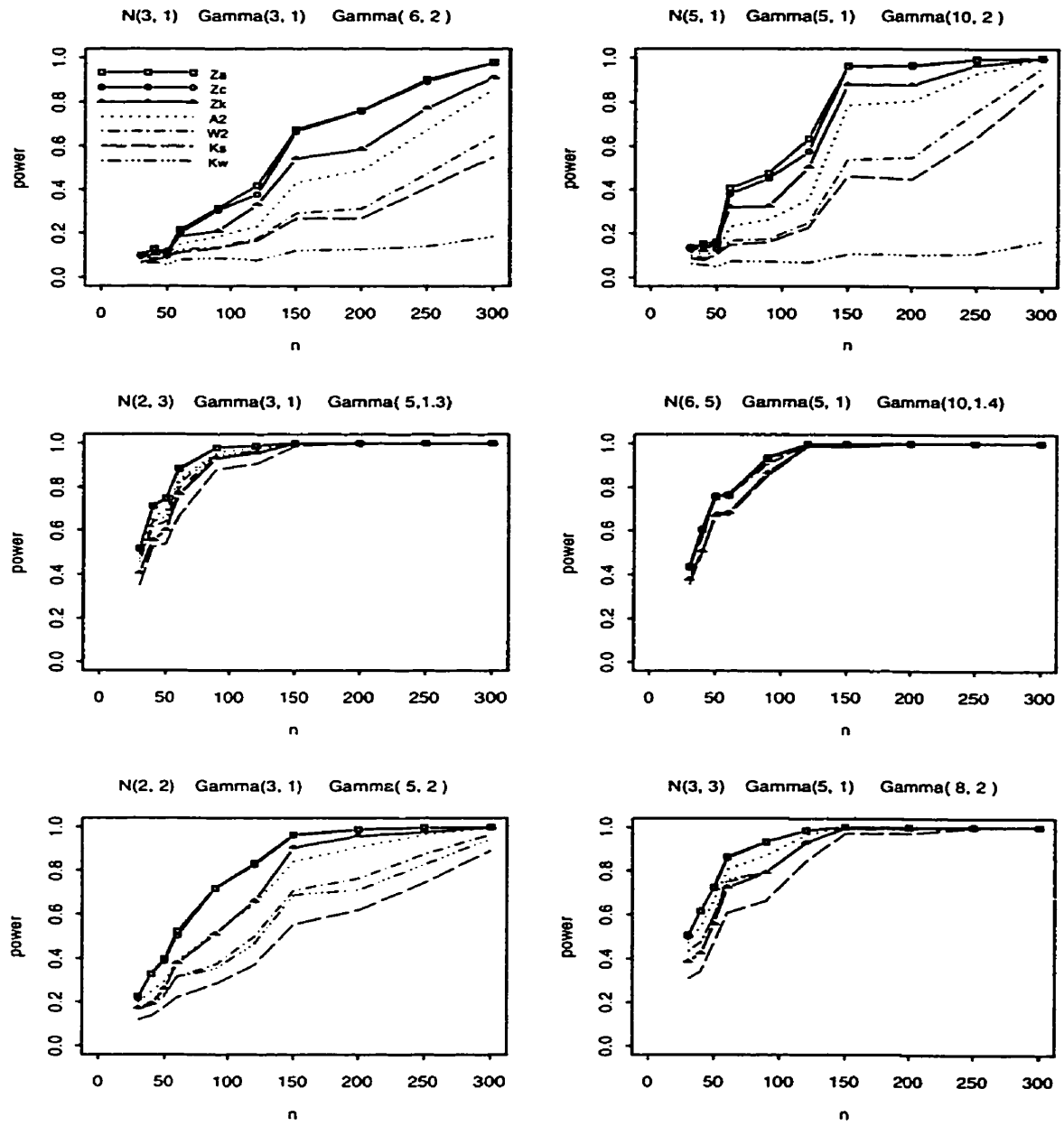

Fig. 14.3. Power comparison for testing $F_1 = N(\mu,\ \sigma^2)$ and $F_i = Gamma(a_i,\ b_i)$ (i=2, 3) at level $\alpha = 0.05$

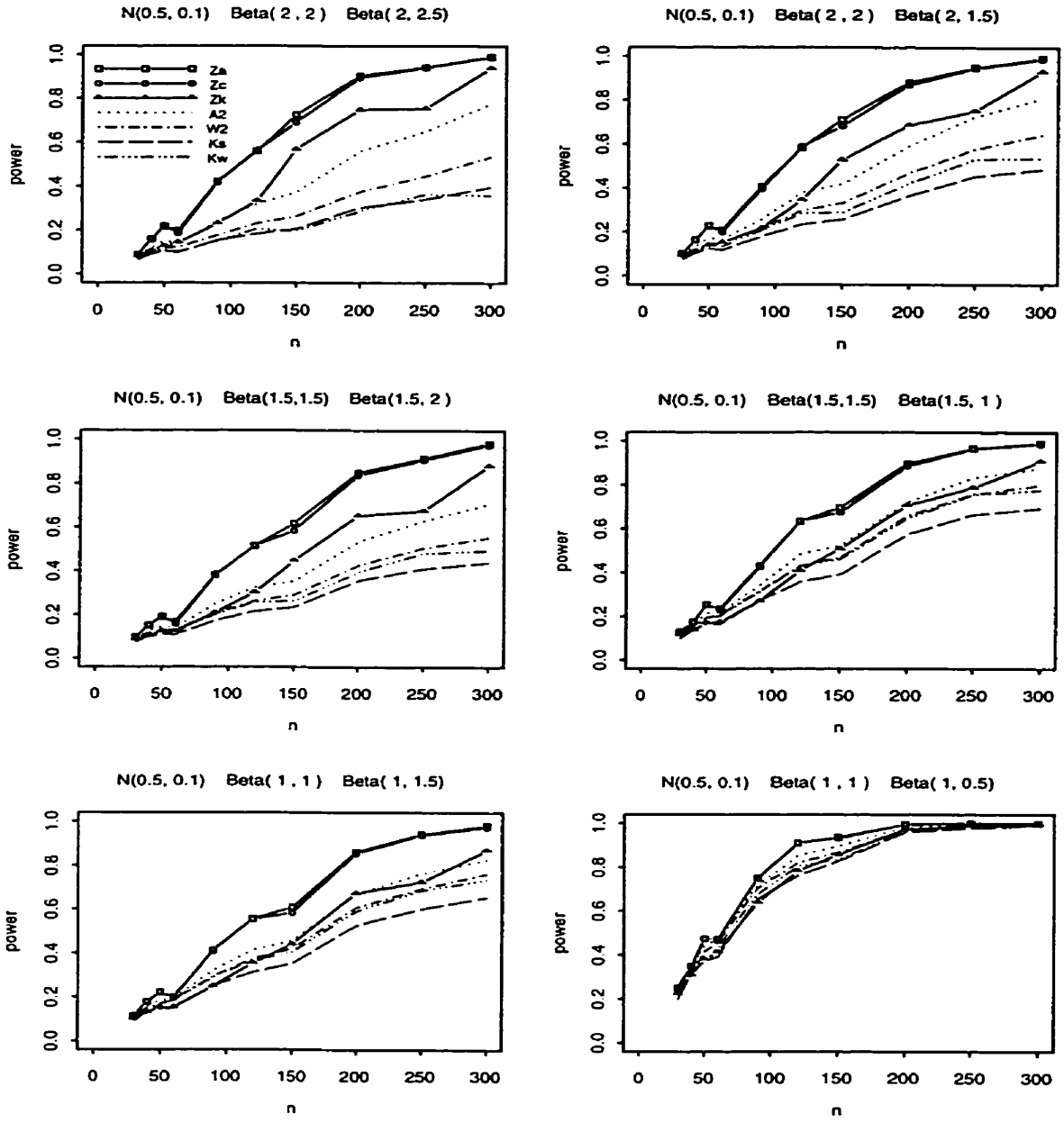

Fig. 14.4. Power comparison for testing $F_1 = N(0.5,\ 0.1)$ and $F_i = Beta(p_i,\ q_i)$ (i=2, 3) at level $\alpha = 0.05$

15. The Distributions of k-Sample Z_A, Z_C and Z_K

As the two-sample case, the k-sample statistics Z_A, Z_C and Z_K in (13.1)-(13.3) are distribution-free. Their null distributions are uniformly discrete, and thus can be obtained by enumeration of all possible values of the statistics by considering $n!/(n_1!n_2!\cdots n_k!)$ combinations of the ranks R_{ij} ($i = 1, 2, ..., k; j = 1, 2, ..., n_i$). Note that Z_A, Z_C and Z_K are functions of R_{ij}. Under the null hypothesis H, every combination of 1, 2, ..., n into k groups of sizes n_1, n_2 ..., n_k is equally likely to be the ranks of the k samples.

Since the number of values that each of the statistics can take on increases very rapidly with n_i and k, it is not feasible to give the full distribution unless n_i and k are small. Moreover, extensive tables have to be used for tabulating their percentage points with different n_i and k (even if they are small).

Fortunately, with today's computing facilities and software, it is easy to get the approximate p-value of such a k-sample test by Monte Carlo simulation. The accuracy is good enough in practice if the number of replicates (N) is sufficiently large. In fact, the standard error of simulated p-value is $\sqrt{p(1-p)/N}$.

Computer programs in Splus code (Program 1-3) are given below to calculate each new statistic and its simulated p-value for the k-sample test. These programs are easy to run even on PC. For example, for a three-sample test with sizes n_1=100 and n_2=n_3=200, Program 1 requires about three minutes to run on PC (Pentium II-MMX CPU at 300 MHz) for a 10,000-sized simulation.

In Program 1-3, M is simulation size, $n = (n_1,\ n_2,\ ...,\ n_k)$ and X is the vector of data for the k samples, i.e.,

$$X = (x_{11},\ x_{12},\ ...,\ x_{1n_1}, x_{21},\ x_{22},\ ...,\ x_{1n_2},\ \ .....\ \ , x_{k1},\ x_{k2},\ ...,\ x_{1n_k}).$$

Program 1. Calculating Z_C and its p-value (k-sample case)

```
 f <- function(n, X, M) {
 N <- sum(n)
 r <- ceiling(rank(X))
 P <- (1:N-.5)/N
 W <- N*(P*log(P)+(1-P)*log(1-P))
 g <- function(n,r,N,W) {
 u <- 0
 k <- length(n)
 m <- 0:k
      for (i in 1:k) {
      m[i+1] <- sum(n[1:i])
           d <- (m[i]+1):m[i+1]
 u <- u+sum(log(n[i]/(1:n[i]-.5)-1)*log(N/(sort(r[d])-.5)-1)) }
      return(u/N) }
Zc <- g(n,r,N,W)
 z <- 0
      for (m in 1:M) z <- z + (g(n,rank(sample(N)),N,W) < Zc)
```

```
            p.value <- z/M
    return(Zc, p.value) }
f( c(10, 20, 20), runif(50), 100 )
```

Program 2. Calculating Z_A and its p-value (k-sample case)

```
f <- function(n, X, M) {
N <- sum(n)
r <- ceiling(rank(X))
P <- (1:N-.5)/N
W <- N*(P*log(P)+(1-P)*log(1-P))
g <- function(n,r,N,W) {
u <- rep(0,N)
k <- length(n)
m <- 0:k
for (i in 1:k) { m[i+1] <- sum(n[1:i])
d <- (m[i]+1):m[i+1]
d <- sort(r[d])
D <- c(1, d, N+1)
p <- rep(0:n[i], D[2:(n[i]+2)]-D[1:(n[i]+1)])
    p[d] <- p[d]-.5
p <- p/n[i]
```

```
 u <- u-n[i]*(p*log(p+.000000001)+(1-p)*log(1-p+.000000001)) }
      return( sum(u/(1:N-.5)/(N:1-.5)) ) }
Za <- g(n,r,N,W)
 z <- 0
      for (m in 1:M) z <- z + (g(n,rank(sample(N)),N,W) < Za)
               p.value <- z/M
      return(Za, p.value) }
  f( c(10, 20, 20), runif(50), 100 )
```

Program 3. Calculating Z_K and its p-value (k-sample case)

```
 f <- function(n, X, M) {
 N <- sum(n)
 r <- ceiling(rank(X))
 P <- (1:N-.5)/N
 W <- N*(P*log(P)+(1-P)*log(1-P))
 g <- function(n,r,N,W) {
 u <- rep(0,N)
 k <- length(n)
 m <- 0:k
      for (i in 1:k) { m[i+1] <- sum(n[1:i])
 d <- (m[i]+1):m[i+1]
```

```
 d <- sort(r[d])
 D <- c(1, d, N+1)
      p <- rep(0:n[i], D[2:(n[i]+2)]-D[1:(n[i]+1)])
      p[d] <- p[d]-.5
 p <- p/n[i]
 u <- u+n[i]*(p*log(p+.000000001)+(1-p)*log(1-p+.000000001)) }
      return( max(u-W) ) }
Zk <- g(n,r,N,W)
 z <- 0
      for (m in 1:M) z <- z + (g(n,rank(sample(N)),N,W) > Zk)
               p.value <- z/M
      return(Zk, p.value) }
  f( c(10, 20, 20), runif(50), 100 )
```

16. Beta Approximation to the Distribution of K_S

Although the sampling distributions of EDF tests are intractable, sometimes they can be approximated by a simple distribution family. As a simple example, we now consider the Kolmogorov-Smirnov statistic K_S in Section 2.

The asymptotic distribution of K_S under null hypothesis was derived by Kolmogorov (1933), and Smirnov (1939) gave a simpler proof. However, the exact null distribution for finite-sample case is complicated to express. Kolmogorov (1933) and Massey (1950) established recursive formulas for calculating the null probability $P(K_S < k/n)$ for integer values of k. Then Birnbaum (1952) tabulated these

values for n=1, 2, ..., 100 and k=1, 2, ..., 15.

Since the exact null distribution of K_S is only available at k/n for limited integer values of k, approximate methods have been explored. For example, some critical values of K_S based on interpolation were given by Massey (1951) and Birnbaum (1952), and the most common-used approximate critical values in statistical tables and literature were from Miller (1956). However, the approximation is only valid for the upper tail of the distribution, since the critical values (with level α) are approximated by the exact ones (with level $\alpha/2$) for one-sided test. See Conover (1980) and Gibbons (1992).

Research on the Kolmogorov-Smirnov statistics and their sampling distributions remains very active. See, for instance, Cabaña (1996), Cabaña and Cabaña (1994,1997), Friedrich and Schellhaas (1998), Justel, Peña and Zamar (1997), Kim (1999), Kulinskaya (1995), Paramasamy (1992) and Rama(1993) among others.

In the following we will show that the distribution of the K_S can be globally approximated by a general beta distribution. The approximation is very simple and reliable. Therefore, we may use a beta distribution to find the practical p-value of the K_S test.

On the other hand, traditional methods of approximating the p-value are more complicated and less accurate. For example, the current approximation method used in S-Plus is based on interpolation for small sample ($n \leq 50$) or the limiting distribution for $n > 50$, which is not be accurate enough.

Let $B_{p,\,q}$ denote a random variable having standard beta distribution $Beta(p,\,q)$

with density

$$b_{p,\,q}(x) = x^{p-1}(1-x)^{q-1}/B(p,\ q), \quad 0 < x < 1,$$

and distribution function

$$B_{p,\,q}(x) = \int_{-\infty}^{x} b_{p,\,q}(t)\ dt, \quad -\infty < x < \infty,$$

where $B(p,\ q)$ is beta function with $p,\ q > 0$.

Our simulation study shows that the distribution of Kolmogrov-Smirnov statistic K_S approximately equals that of a general beta variable $aB_{p,\,q} + b$, where constants $a,\ b,\ p,\ q$ are chosen such that K_S and $aB_{p,\,q} + b$ have the same first four moments, or equivalently have the same mean μ, standard deviation σ, skewness $r_1 = \bar{\mu}_3/\sigma^3$ and kurtosis $r_2 = \bar{\mu}_4/\sigma^4$, where $\bar{\mu}_k$ denotes the k-th central moment.

Let μ_n, σ_n, r_{n1} and r_{n2} be, respectively, the mean, standard deviation, skewness and kurtosis of K_S. It is easy to prove that K_S and $aB_{p,\,q}+b$ $(a > 0)$ have the same mean, standard deviation, skewness and kurtosis (or equivalently have the same first four moments) if and only if

$$\begin{cases} \mu_n = \dfrac{ap}{p+q} + b\ , & r_{n1} = \dfrac{2(q-p)}{p+q+2}\sqrt{\dfrac{p+q+1}{pq}}\ , \\ \sigma_n = \dfrac{a}{p+q}\sqrt{\dfrac{pq}{p+q+1}}\ , & r_{n2} = \dfrac{3(p+q+1)}{p+q+3}\left[\dfrac{2(q-p)^2}{pq(p+q+2)} + 1\right]\ . \end{cases}$$

Note that

$$\begin{cases} p+q = P(r_{n1},\ r_{n2}) & \text{with } P = P(x,\ y) = 6(y - x^2 - 1)/(3x^2 - 2y + 6), \\ p\,q = Q(r_{n1},\ r_{n2}) & \text{with } Q(x,\ y) = 4P^2/[16 + x^2(P+2)^2/(P+1)]. \end{cases}$$

Hence, a, b, p, q ($a > 0$) are uniquely decided by

$$\begin{cases} p,\ q \ = \ \left[\ P(r_{n1},\ r_{n2}) \pm \sqrt{P(r_{n1},\ r_{n2})^2 - 4Q(r_{n1},\ r_{n2})}\ \right]/2\ , \\ a \ = \ \sigma_n(p+q)\sqrt{(p+q+1)/(pq)}\ , \qquad b \ = \ \mu_n - ap/(p+q)\ , \end{cases} \tag{16.1}$$

with $q > p$ ($q \leq p$) if $r_{n1} > 0$ ($r_{n1} \leq 0$).

Then K_S and $aB_{p,\,q} + b$ have the exactly same first four moments, as well as the approximately same moments of higher order based on our simulation (see below). Therefore, they have approximately the same moment generating function or characteristic function, and thus they have approximately the same distribution.

As a result, $F_{K_S}(x)$ and $f_{K_S}(x)$, the distribution and density functions of K_S, can be simply approximated by those of $aB_{p,\,q} + b$, i.e.

$$F_{K_S}(x) \ \approx \ B_{p,\,q}(\frac{x-b}{a}) \qquad \text{and} \qquad f_{K_S}(x) \ \approx \ b_{p,\,q}(\frac{x-b}{a})/a\ , \tag{16.2}$$

where a, b, p, q are given by (16.1) and will be approximated by (16.3).

Usually, having the same first four moments is not enough to guarantee a very good approximation, but (16.2) is a special case where the two distributions also have sufficiently close moments of higher orders and thus have approximately the same moment generating or characteristic function. For n=10, 100 and 1000, for example, the first ten standard moments of the two variables are listed in Table 16.1, where the upper numbers in double entries are the moments of K_S based on simulation with size of one million, and the lower numbers correspond to $aB_{p,\,q} + b$. The specific values of a, b, p, q are given in Table 16.2 below.

It can be seen from Table 16.1 that they do have the same first four moments and similar moments of higher orders. Unfortunately, the exact first four moments

of K_S are not available to determine a, b, p, q, so we use Monte Carlo approach. For some selected values of n, Table 16.2 lists the first four standard moments of K_S obtained by a one-million-size simulation together with the corresponding values of a, b, p, q calculated from (16.1).

TABLE 16.1. The first ten standard moments of K_S and $aB_{p,\,q}+b$

n	μ	σ	$\bar{\mu}_3/\sigma^3$	$\bar{\mu}_4/\sigma^4$	$\bar{\mu}_5/\sigma^5$	$\bar{\mu}_6/\sigma^6$	$\bar{\mu}_7/\sigma^7$	$\bar{\mu}_8/\sigma^8$	$\bar{\mu}_9/\sigma^9$	$\bar{\mu}_{10}/\sigma^{10}$
10	.2592	.07983	.8180	3.697	8.237	29.78	98.51	377.7	1500	6329
	.2592	.07983	.8180	3.697	8.399	30.60	104.1	408.7	1672	7315
100	.0852	.02592	.8561	3.869	9.209	34.70	124.2	516.0	2235	10533
	.0852	.02592	.8561	3.869	9.311	35.10	127.1	529.6	2329	10999
1000	.0273	.00823	.8616	3.884	9.285	34.93	125.2	517.0	2248	10403
	.0273	.00823	.8616	3.884	9.399	35.47	129.0	539.2	2380	11286

TABLE 16.2. The first four standard moments of K_S and corresponding a, b, p, q

n	μ_n	σ_n	r_{n1}	r_{n2}	a	b	p	q
5	0.35826	0.109496	0.7583	3.495	1.1571	0.14674	2.867	12.82
10	0.25916	0.079832	0.8180	3.697	1.0186	0.10590	2.980	16.83
20	0.18636	0.057362	0.8389	3.784	0.8215	0.07555	3.093	19.84
30	0.15331	0.047046	0.8457	3.818	0.7153	0.06190	3.165	21.60
50	0.11967	0.036602	0.8559	3.862	0.5974	0.04830	3.229	23.80
70	0.10150	0.031013	0.8562	3.863	0.5053	0.04106	3.224	23.73
100	0.08519	0.025916	0.8561	3.869	0.4327	0.03445	3.267	24.59
150	0.06985	0.021213	0.8599	3.887	0.3659	0.02823	3.298	25.70
200	0.06063	0.018404	0.8607	3.878	0.3051	0.02481	3.225	24.25
300	0.04960	0.015020	0.8594	3.864	0.2401	0.02053	3.171	23.02
500	0.03851	0.011636	0.8574	3.863	0.1878	0.01590	3.201	23.38
1000	0.02730	0.008229	0.8616	3.884	0.1385	0.01125	3.245	24.76

Since K_S is distribution-free, its moments depend only on n, so do a, b, p, q in (16.1). For simplicity, linear functions of n^{-1} and n^d (d is fixed) are used to

approximate them. Of course, better approximations may be made by using more complicated functions at the price of the simplicity.

Directly fitting the data of a, b, p, q in Table 16.2 does not work well and could destroy their structure in (16.1) which enable the approximate distribution to have correct first moments. Instead we fit the moments first. A linear regression model $y = \beta_0 + \beta_1 n^{-1} + \beta_2 n^d$ is used to fit (by least squares approach) the data of μ_n, σ_n, r_{n1}, r_{n2} (against n) in Table 16.2 respectively. For different values of d, we have different models to fit the data. We choose a d which roughly corresponds to the best fit by the following approach. One can choose any initial value of d, and then fit the model to the data. If the fit is satisfactory, stop. Otherwise, increase or decrease the value of d and fit the model again. Then choose the d which corresponds to a better fit. Repeat this process until a satisfactory fit is obtained. In this way, the best d can be roughly reached within a few steps by our experience. The results are as follows:

$$\begin{cases} \hat{\mu}_n = -0.00008631 - 0.1348/n + 0.8587/n^{0.498} \\ \hat{\sigma}_n = 0.0004787 - 0.09059/n + 0.296/n^{0.525} \\ \hat{r}_{n1} = 0.861 - 0.3748/n - 0.6908/n^2 \\ \hat{r}_{n2} = 3.884 - 1.815/n - 0.6549/n^2. \end{cases}$$

Then, using them as the mean, standard deviation, skewness and kurtosis of K_S, we can get a new set of data for a, b, p, q (against n) via (16.1). The new data is well fitted by

$$\begin{cases} \hat{a} = 0.003326 - 6.012/n + 5.52/n^{0.53} \\ \hat{b} = -0.0004245 - 0.003397/n + 0.3204/n^{0.48} \\ \hat{p} = 3.258 - 3.727/n + 4.607/n^{1.6} \\ \hat{q} = 25 - 161.2/n + 162.2/n^{1.3}, \end{cases} \tag{16.3}$$

which thus well keeps the original structure of (16.1).

With a, b, p and q approximated by (16.3) the distribution of Kolmogorov-Smirnov statistic K_S can be simply approximated by a completely known beta distribution in (16.2).

We now discuss the accuracy of the approximation. For n=10, 50, 100, 200 and 500, Table 16.3 lists three sets of percentage points of K_S for comparison. The first line in multiple entries is obtained from (16.2) with a, b, p and q given in (16.3); the second line is based on a Monte Carlo simulation of size 100,000; the third line lists the most common-used approximate values given by Miller (1956), which are only available for upper tail (asymptotic values are used if $n > 100$). See also Conover (1980) and Gibbons (1992).

TABLE 16.3. Percentage points for K_S

n	0.01	0.05	0.10	0.20	0.50	0.80	0.90	0.95	0.99
10	.1300	.1512	.1667	.1897	.2479	.3239	.3698	.4103	.4910
	.1273	.1518	.1673	.1896	.2468	.3222	.3691	.4099	.4885
						.3226	.3687	.4093	.4889
50	.0606	.0703	.0773	.0877	.1139	.1482	.1691	.1878	.2256
	.0596	.0706	.0778	.0881	.1140	.1482	.1693	.1883	.2265
						.1484	.1696	.1884	.2260
100	.0433	.0503	.0553	.0627	.0813	.1057	.1206	.1339	.1608
	.0426	.0504	.0556	.0630	.0812	.1055	.1207	.1341	.1608
						.1056	.1207	.1340	.1608
200	.0310	.0359	.0394	.0447	.0579	.0752	.0858	.0952	.1144
	.0305	.0361	.0396	.0447	.0577	.0750	.0856	.0950	.1145
						.0759	.0865	.0960	.1151
500	.0197	.0229	.0251	.0284	.0368	.0478	.0545	.0605	.0726
	.0194	.0229	.0252	.0285	.0366	.0477	.0543	.0603	.0723
						.0480	.0547	.0607	.0728

It can be seen from Table 16.3 that (a) compared with the simulation results, our approximate values are very accurate in the whole region (lower, central and upper parts) of the distribution, and the higher the percentage level, the more accurate the approximation; (b) at the upper tail they are consistent with Miller's approximate results for $n \leq 100$ but are better than asymptotic values, which are always a little bit larger than real ones, especially when $n \leq 200$.

The exact sampling distribution of K_S is complicated. Kolmogorov (1933) and Massey (1950) established recursive formulas for calculating the null probability

$P(K_S < k/n)$ for integer values of k. Note that the recursive formulas only apply to integer k. Birnbaum (1952) tabulated these values for n=1, 2, ..., 100 and k=1, 2, ..., 15. We can use Birnbaum's tables to check the accuracy of our approximation. Below are such exact values for n=40 compared with the values obtained by the beta approximation given by (16.2):

k	3	4	5	6	7	8	9	10	11	12
exact	.0345	.2182	.4808	.7016	.8471	.9295	.9708	.9891	.9964	.9989
appr.	.0344	.2224	.4812	.7021	.8488	.9311	.9716	.9894	.9964	.9989

It can be seen that the values corresponding to the same k are almost equal, especially for large k. The situation is similar for other sample sizes.

We conclude from above that (16.2) globally gives very simple and accurate approximations to the distribution and density functions of K_S, especially at the upper tail (the most important part). Hence, we can easily use a beta distribution to find the practical p-value of the Kolmogorov-Smirnov test, which is simpler and more accurate than existing methods in the literature. For example, the current approximation method used in S-Plus is based on interpolation on limited values of exact distribution for small sample ($n \leq 50$), or the limiting distribution for $n > 50$, which has been shown in Table 16.3 that it may not give a good approximation to the true value if $n < 200$.

17. A Simple Distribution Family

In order to generalize the approach in the previous section, we now consider a family of distributions $V(a,\ b,\ p,\ q,\ c,\ d)$ generated by the random variable

$$V\ =\ a\ (Y-\mu^Y)/Z\ +\ b\ , \tag{17.1}$$

where 1. $Y \sim\ Beta(p,\ q)$, a beta distribution with parameters p and q and $\mu^Y = E(Y)$;

2. $a > 0$ and b are constants;

3. Z has the density defined by a simple step function

$$f_Z(t)\ =\ c_i,\ \ d_{i-1} < t < d_i\ \ \ (\ d_i = d + i/I\ ;\ i = 1, 2, \cdots, I) \tag{17.2}$$

with $c_i,\ d\ >\ 0$ as constants, $i = 1, 2, \ldots, I$, satisfying $\sum_{i=1}^{I} c_i\ =\ I$;

4. Y and Z are independent.

For any function $h(y)$, denote $h(b) - h(a)$ by $[h(y)]_{y=a}^{y=b}$. After some routine calculation, we can find the distribution function of V as follows .

$$F_V(x)\ =\ \sum_{i=1}^{I} \frac{c_i}{t} \left[yB_{p,\ q}(y) - \mu^Y B_{p+1,\ q}(y)\right]_{y=t_{i-1}}^{y=t_i}\ , \tag{17.3}$$

where $t\ =\ (x-b)/a,\ \ t_i\ =\ \mu^Y + td_i$ and $B_{p,\ q}(y)$ is the distribution function of $Beta(p,\ q)$.

Suppose that a random variable X has complicated distribution but finite moments $\mu_k^X\ =\ E(X^k)$, k=1, 2, 3, 4. Let the skewness and kurtosis of X denoted by $r_1^X\ =\ \bar{\mu}_3^X/(\bar{\mu}_2^X)^{3/2}$ and $r_2^X\ =\ \bar{\mu}_4^X/(\bar{\mu}_2^X)^2$ respectively, where $\bar{\mu}_k^X$ is the k-th central moment (about the mean). We wish to approximate the distribution function of X by a member of the family $V(a,\ b,\ p,\ q,\ c,\ d)$. We do this in two steps.

STEP 1.

We choose the constants a, b, p, and q such that X has the same first four moments as those of V in (17.1).

First, we let them have the same skewness and kurtosis, i.e.

$$r_1^X = r_1^Y \mu_3^{1/Z}/(\mu_2^{1/Z})^{3/2} \quad \text{and} \quad r_2^X = r_2^Y \mu_4^{1/Z}/(\mu_2^{1/Z})^2$$

or

$$r_1^Y = r_1 = r_1^X (\mu_2^{1/Z})^{3/2}/\mu_3^{1/Z} \quad \text{and} \quad r_2^Y = r_2 = r_2^X (\mu_2^{1/Z})^2/\mu_4^{1/Z}, \tag{17.4}$$

where

$$mu_k^{1/Z} = E(1/Z)^k = \sum_{i=1}^{I} c_i (d_i^{1-k} - d_{i-1}^{1-k})/(1-k). \tag{17.5}$$

Since Y has beta distribution, we have the following:

$$\begin{cases} p+q = P(r_1^Y, r_2^Y), \quad P(x,y) = 6(y-x^2-1)/(3x^2-2y+6), \\ \\ p\,q = Q(r_1^Y, r_2^Y), \quad Q(x,y) = 4P^2/[16 + x^2(P+2)^2/(P+1)]. \end{cases} \tag{17.6}$$

Substituting r_1^Y and r_2^Y in (17.6) with r_1 and r_2 in (17.4), we can solve for p and q. Hence,

$$p,\ q \ = \ \left[\ P(r_1,\ r_2) \pm \sqrt{P(r_1,\ r_2)^2 - 4Q(r_1,\ r_2)}\ \right]/2 \tag{17.7}$$

with $q > p$ $(q \le p)$ if $r_1 > 0$ $(r_1 \le 0)$.

Second, we let X and V have the same mean and variance, which can be easily achieved by choosing

$$a \ = \ \sqrt{\bar{\mu}_2^X/(\bar{\mu}_2^Y \mu_2^{1/Z})} \qquad \text{and} \qquad b \ = \ \mu^X \tag{17.8}$$

where $\bar{\mu}_2^Y \ = \ pq/(p+q)^2/(p+q+1)$.

Now we have already fixed a, b, p, and q so that X and V in (17.1) have the same first four moments. We use (17.3) to approximate the distribution function of X.

STEP 2.

Generally speaking, having the same first four moments is still not sufficient to guarantee a very good approximation. Therefore, we use c_i, $i = 1, 2, \ldots, I$, and d of the step function (17.2) as tuning parameters for further improvement.

The construction of (17.1) is based on the following: (1) beta distributions have different (but not arbitrary) kinds of shapes so that Y plays a key role in controlling the basic shape (skewness and kurtosis) of the approximate distribution; (2) step functions can approach any density function so that Z works well as an adjustment; (3) a and b are scale and location parameters. Besides, beta distributions have the excellent property (17.6), while using step functions as densities makes (17.3) and (17.5) simple.

It is obvious that the larger the number of steps (I) we use, the more accurate the approximation can be made, but, on the other hand, the more difficult to tune the parameters. Also, small I makes both (17.3) and (17.5) simple.

If higher order moments of X are known, we can choose the tuning parameters in (17.2) to match the higher order moments of X with $V(a, b, p, q, c, d)$. If higher order moments are not known, but some quantiles of X are available through simulation or other methods, we can determine the tuning parameters to match the quantiles. Examples are presented in Sections 18 and 19, where we fit a member

of the family $V(a,\ b,\ p,\ q,\ c,\ d)$ to Cramér-von Mises and Waston's statistics for goodness of fit.

18. Approximate distribution for Cramér-von Mises Statistic

The exact null distribution of the classical Cramér-von Mises statistic W^2 in Section 2 is not known except for $n = \infty$ and n=1-7 (see, e.g., Anderson and Darling (1952), Marshall (1958), and Knott (1974)). So, some approximate methods have been explored.

Using the first four moments (about the mean) of W^2,

$$\begin{cases} \mu_1 = 1/6 \\ \bar{\mu}_2 = (4 - 3/n)/180 \\ \bar{\mu}_3 = (32 - 61/n + 30/n^2)/3780 \\ \bar{\mu}_4 = (496 - 1532/n + 1671/n^2 - 630/n^3)/75600, \end{cases} \tag{18.1}$$

Pearson and Stephens (1962) obtained some approximate percentage points for n=5, 10 and ∞ by fitting Johnson's S_B curve, which can only give approximately the same first four moments but requires extensive computation.

Tiku (1965) used $a+b\chi^2_p$ to get a simple approximation by choosing a, b, and p to give the same first three moments as above. The method is easy to apply and gives good results at the upper tail of the distribution, but it fails at the lower tail. This is not surprising because the first three (or even four) moments are not sufficient to determine a distribution.

Stephens (1970) gave an empirical way to approximate the upper (lower) $100\alpha\%$ percentage point of W^2 by $U_\alpha/(1+1/n)+0.4/n-0.6/n^2$ ($L_\alpha/(1+0.5/n)+0.03/n$), where U_α (L_α) is the corresponding point of W^2_∞. This method works very well at both upper and lower tails. However, it is not suitable for the central part of the distribution. For instance, since $U_{0.25}$=0.20939, the corresponding approximate value for W^2_7 is 0.22812, but the true value is 0.21087.

The best known approximate results are given by Csörgő and Faraway (1996). Using one-term linking approximation to the limiting distribution and combining many published results from literature, they found a sophisticated approximation with high accuracy for all (upper, lower and central) parts of the distribution. The only concern is the complexity of their formula. Moreover, it includes some special functions, such as Bessel functions. Therefore it is difficult to find the practical p-value of a goodness-of- fit test.

In contrast, the method described in Section 16 is simple and can give a very accurate approximation to whichever part (upper, lower or central) of the distribution. We now discuss it in detail as below.

Let W^2 be the random variable X in Section 17. Using its known first four moments in (18.1), we can easily find a, b, p, and q via (17.7) and (17.8) with c_i, i=1, 2, ... I, and d fixed. Then W^2 and V in (17.1) have the same first four moments, and thus (17.3) can be used to approximate the distribution of W^2.

In order to have a better approximation, we need the second step i.e. the tuning procedure for c_i, $i = 1, 2, \ldots, I$, and d. How to choose the best tuning parameters

is anther topic of undergoing research. Generally speaking, the objective of tuning parameters is to make the V and W^2 have similar moments of higher orders. Unfortunately, here we do not have other moments except the first four in (18.1). Another way of tuning parameters is to let the two variables have approximately the same percentage points. Of course, the theoretical percentage points of W^2 are unknown, so we use Monte Carlo approach to simulate their values which are almost the same as the approximations given by Csörgő and Faraway(1996). With simulated percentage points of W^2, we can perform the tuning process in the following.

Since W^2 has limiting distribution, we start with large sample size, say n=1000. First, we use one-step function (I=1) to try. With c_1=1, it is easy to tune a single parameter d (just several times) so that V and W^2 have close percentage points as possible. If the accuracy is not satisfactory, then use two-step function to try, but this time we only need to tune c_1 and d since $c_2 = 1 - c_1$. For this example, we finally use a four-step function with $(c_1,\ c_2,\ c_3,\ c_4) \propto (29,\ 3,\ 1,\ 6)$ and d=0.84. This configuration of parameters also works well for n=500, 200, 100, 50 and 30. For n=20, 10, and 5, we keep using the same c_i's but choose d=0.83, 0.80, and 0.72. For simplicity, we use the following parameters for general n.

$$I = 4,\ (c_1,\ c_2,\ c_3,\ c_4) = \frac{4}{39}(29,\ 3,\ 1,\ 6),\ d = \min(0.84,\ 0.86 - 0.6/n) \qquad (18.2)$$

We spent about two hours to get this result which can be used forever. The turning process is just like tuning a machine or equipment to fit different conditions. For different statistics we need different parameters. In fact, it is because of the turning parameters that make (17.1) suitable for general case.

Since a, b, p, and q have been decided by (17.7) and (17.8), by (18.2), we obtain

a complete known formula (17.3) to approximate the distribution function of W^2.

Table 18.1 lists some of these values for n=10, 50, 200, 1000, and Table 18.2 gives the corresponding approximate percentage points (the upper numbers in double entries) obtained from (17.3) together with those (the lower numbers) obtained by Csörgő and Faraway (1996) for comparison.

We can see from Table 18.2 that the two sets of numerical results are very close. Therefore, the two approximate distributions are almost the same although they are obtained by using different methods, and thus have totally different mathematical expressions.

Compared with Csörgő and Faraway's approach, our method is much simpler without using any special function. Therefore it can be easily used to obtain practical p-values of goodness-of-fit tests, instead of tabulating limited percentage points. Moreover, the approximate function in (17.3) is a real distribution function. Finally, our method is of generality and can be applied to approximate the distribution for a general continuous random variable with known finite first four moments.

Remark: (18.2) is just a simple solution. It is not difficult to find a better configuration of c_i and d if we let c_i vary with n (like d). Besides, increasing I, the number of steps in (17.2), can absolutely raise the degree of accuracy of approximations, but our formulas will be more complicated.

TABLE 18.1. Some values of a, b, p, q

n	a	b	p	q
10	3.8428	1/6	0.6379	20.6594
50	16.7855	1/6	0.6573	88.4046
200	31.8450	1/6	0.6589	168.2916
1000	41.4957	1/6	0.6592	219.4716

TABLE 18.2. Some percentage points for W^2

n	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.99	0.999
10	.0263	.0384	.0477	.0718	.1200	.2100	.3458	.4545	.7139	1.078
	.0265	.0384	.0478	.0721	.1207	.2104	.3450	.4542	.7147	1.082
50	.0252	.0370	.0461	.0703	.1185	.2092	.3476	.4603	.7373	1.152
	.0251	.0369	.0464	.0706	.1192	.2096	.3468	.4599	.7373	1.145
200	.0248	.0367	.0458	.0701	.1183	.2090	.3479	.4613	.7415	1.164
	.0249	.0367	.0461	.0704	.1190	.2094	.3472	.4609	.7415	1.158
1000	.0248	.0366	.0457	.0700	.1182	.2090	.3480	.4616	.7427	1.168
	.0248	.0367	.0460	.0703	.1189	.2094	.3472	.4612	.7426	1.161

19. Approximate Results for Waston's Statistic

The Waston's statistic

$$U^2 = n\int_{-\infty}^{\infty}\{\ F_n(x) - F(x) - \int_{-\infty}^{\infty}[F_n(y) - F(y)]dF(y)\ \}^2\ dF(x)$$
$$= \frac{1}{12n} + \sum_{i=1}^{n}[F_0(X_{(i)}) - \frac{i-0.5}{n}]^2 - n[\frac{1}{n}\sum_{i=1}^{n}F_0(X_i) - 0.5]^2$$

is a modification to the Cramér-von Mises statistic W^2 so that it can test the goodness of fit on a circle.

Like Cramér-von Mises statistic, U^2 is distribution-free. The exact distribution of U^2 was given by Watson (1961) for $n = \infty$ and by Stephens (1963, 1964) for n=1, 2, 3, and 4. Approximate approaches based on moments or empirical estimation have been studied by Pearson and Stephens (1962), Tiku (1965) and Stephens (1970). The results are similar to those for W^2 (see Section 17). Again, the best known approximate results were given by Csörgő and Faraway (1996).

Now, using the first four moments (about the mean) given by Stephens (1963):

$$\begin{cases} \mu_1 = 1/12 \\ \bar{\mu}_2 = (1 - 1/n)/360 \\ \bar{\mu}_3 = (2 - 5/n + 3/n^2)/7560 \\ \bar{\mu}_4 = (19 - 70/n + 87/n^2 - 36/n^3)/302400\ , \end{cases} \tag{19.1}$$

we apply the method in Section 16 to U^2, which has lower skewness and kurtosis than those of W^2, but this time we just need a three-step function (17.2) (I=3).

Results parallel to those of W^2 in Section 18 are given below. We can see again from Table 19.2 that our numerical results are very close to those given by Csörgő and Faraway (1996).

$$I = 3,\ (c_1,\ c_2,\ c_3) = (21,\ 1,\ 2)/8,\ d = \min(0.91,\ 0.96 - 1.5/n). \tag{19.2}$$

TABLE 19.1. Some values of a, b, p, q

n	a	b	p	q
10	0.9582	1/12	1.1642	18.4019
50	4.5683	1/12	1.2504	86.3966
200	8.6382	1/12	1.2576	164.6947
1000	11.3197	1/12	1.2593	214.3387

TABLE 19.2. Some percentage points for U^2

n	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.99	0.999
10	.0210	.0291	.0345	.0477	.0702	.1054	.1499	.1824	.2545	.3496
	.0212	.0288	.0344	.0476	.0704	.1054	.1498	.1824	.2548	.3507
50	.0199	.0279	.0332	.0467	.0695	.1053	.1514	.1860	.2656	.3779
	.0200	.0276	.0333	.0466	.0696	.1053	.1514	.1860	.2658	.3782
200	.0197	.0277	.0330	.0466	.0693	.1053	.1517	.1867	.2676	.3838
	.0198	.0274	.0331	.0464	.0694	.1053	.1517	.1867	.2678	.3834
1000	.0196	.0276	.0330	.0465	.0693	.1053	.1518	.1868	.2682	.3848
	.0197	.0274	.0331	.0464	.0694	.1053	.1517	.1868	.2683	.3847

20. Concluding Remarks

Through the parameterization introduced in Section 1, a nonparametric goodness-

of-fit test is simplified to a family of parametric tests. As a result, more general types of EDF tests for goodness of fit can be defined, which include traditional EDF tests, as well as new EDF tests based on the likelihood ratio. This methodology for goodness-of-fit tests in Sections 1-5 is summarized in Zhang (2001a).

Besides, instead of testing for a specific distribution, the new statistics can be applied to test the goodness of fit for a family of distributions, such as the families of normal, exponential, gamma, and Weibull distributions. For instance, the new tests outperform the best tests of normality in literature by simulation (see Sections 5-6). Another interesting topic is to test multivariate normality.

The methodology of goodness-of-fit tests has been developed and applied to general two-sample and multi-sample problems, and parallel results have been obtained. The simulations in Sections 8-15 show that the new tests are sensitive to the difference in location, scale and shape among distributions, while traditional tests are dull to detect the variation in shape or scale. The major results of the two-sample tests in Sections 8-10 are given in Zhang (2001b).

Since the exact sampling distributions of the EDF test statistics are intractable, we have to investigate some approximate approaches. In Sections 16-19, a simple distribution family is introduced to approximate the distribution function for a general continuous random variable. When applied to some EDF test statistics, it gives similar numerical results as the best known, but it is much simpler and can be directly used to obtain practical p-values of goodness-of-fit tests (Zhang and Wu, 2001a, 2001b). The key issue in using simple distribution family is choosing the tuning parameters involved. This is still an area of on-going research.

References

Anderson, T. W. and Darling, D. A. (1952) Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Statist.*, **23**, 193-212.

Anderson, T. W. and Darling, D. A. (1954) A test of goodness of fit. *J. Amer. Statist. Assoc.* **49**, 765-769.

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

Baumgartner, W., Weiβ, P. and Schindler, H. (1998) A nonparametric test for the general two-sample problem. *Biometrics*, **54**, 1129-1135.

Birnbaum, Z. W. (1952) Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. *J. Amer. Statist. Assoc.* **47**, 425-441.

Birnbaum, Z. W. and Tingey, F. H. (1951) One-sided confidence contours for probability distribution functions. *Ann. Math. Statist.*, **22**, 595-596.

Burr, E. J. (1963) Distribution of the two-sample Cramér-von Mises criterion for small equal samples. *Ann. Math. Statist.*, **34**, 95- 101.

Burr, E. J. (1964) Small-sample distributions of the two-sample Cramér-von Mises W^2 and Watson's U^2. *Ann. Math. Statist.*, **35**, 1091-1098.

Cabaña, A. (1996) Transformations of the empirical measure and Kolmogorov-Smirnov tests. *Ann. Statist.*, **24**, 2020-2035.

Cabaña, A. and Cabaña, E. M. (1994) Goodness-of-fit and comparison tests of the Kolmogorov-Smirnov type for bivariate populations. *Ann. Statist.*, **22**, 1447-1459.

Cabaña, A. and Cabaña, E. M. (1997) Transformed empirical processes and modified Kolmogorov-Smirnov tests for multivariate distributions. *Ann. Statist.*, **25**, 2388-2409.

Conover, W. J. (1965) Several k-sample Kolmogorov-Smirnov tests. *Ann. Math. Statist.*, **36**, 1019-1026.

Conover, W. J. (1980) *Practical Nonparametric Statistics*, 2nd ed. New York: John Wiley.

Cramér, H. (1928) On the composition of elementary errors: II, Statistical applications. *Skand. Akt.* **11**, 141-180.

Cressie, N. and Reed, T. R. C. (1984) Multinomial goodness-of-fit tests. *J. R. Statist. Soc.* **B, 46**, 440-464.

Csörgő, S. and Faraway, J. J. (1996) The exact and asymptotic distributions of Cramér-von Mises statistics. *J. R. Statist. Soc.* **B, 58**, 221-234.

D'Agostino, R. B. (1971) An omnibus test of normality for moderate and large sample size. *Biometrika* **58**, 341-348.

D'Agostino, R. B. (1973) Monte Carlo power comparison of the W' and D tests of normality. *Comm. Statist.* **1**, 545-551.

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

Darling, D. A. (1957) The Kolmogorov-Smirnov, Cramér-von Mises tests. *Ann. Math. Statist.*, **28**, 823-838.

Dudewicz, E. J. and van der Meulen, E. C. (1981) Entropy-based tests of uniformity. *J. Amer. Statist. Assoc.* **76**, 967-974.

Dyer, A. R. (1974) Comparisons of tests for normality with a cautionary note. *Biometrika* **61**, 185-189.

Epps, T. W. and Singleton, K. J. (1986) An omnibus test for the two- sample problem using empirical characteristic function. *J. Statist. Comput. Simul.*, **26** 177-203.

Ferger, D. (2000) Optimal tests for the general two-sample problem. *J. Multivariate Anal.*, **74**, 1-35.

Friedrich, T. and Schellhaas, H. (1998) Computation of the percentage points and the power for the two-sided Kolmogorov-Smirnov one sample test. *Statist. Papers*, **39**, 361-375.

Gibbons, J. D. (1992) *Nonparametric Statistical Inference*, 3nd ed. New York: Dekker.

Gnedenko, B. V. (1954) Tests of homogeneity of probability distribution in two independent samples (in Russian). *Doklody Akademii Nauk SSSR*, **80**, 525-528.

Hájek, J. and Šidák, Z. (1967) *Theory of Rank Tests*, Academic Press, New York.

Hall, P. and Padmanabhan, A. R. (1997) Adaptive inference for the two- sample scale problem. *Technometrics*, **39**, 412-422.

Hodges, J. L. Jr. (1958) The significance probability of Smirnov two- sample test. *Arkivfoer Matematik, Astronomi och Fysik*, **3**, 469- 486.

Justel, A., Peña, D. and Zamar, R. (1997) A multivariate Kolmogorov- Smirnov test of goodness of fit. *Statist. Probab. Lett.*, **35**, 251- 259.

Kiefer, J. (1959) K-sample Analogues of the Kolmogorov-Smirnov, Cramér-von Mises tests. *Ann. Math. Statist.*, **30**, 420-447.

Kim, T. Y. (1999) On tail probabilities of Kolmogorov-Smirnov statistics based on uniform mixing processes. *Statist. Probab. Lett.*, **43**, 217-223.

Knott, M. (1974) The distribution of the Cramér-von Mises statistic for small sample sizes. *J. R. Statist. Soc.* **B, 36**, 430-438.

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

Kruskal, W. H. and Wallis, W. A. (1952) Use of ranks in one-criterion analysis of variance. *J. Amer. Statist. Assoc.*, **47**, 583-621; errata, ibid., **48**, 907-911.

Kulinskaya, E. (1995) Coefficients of the asymptotic distribution of the Kolmogorov-Smirnov statistic when parameters are estimated. *J. Nonparametr. Statist.*, **5**, 43-60.

Marshall, A. W. (1958) The small sample distribution of $n\omega_n^2$. *Ann. Math. Statist.*, **29**, 307-309.

Massey, F. J. (1951) The Kolmogorov-Smirnov test for goodness of fit. *J. Amer. Statist. Assoc.*, **46**, 68-77.

Massey, F. J. (1952) Distribution table for the deviation between two sample cumulatives. *Ann. Math. Statist.*, **23**, 435-441.

Miller, L. H. (1956) Table of percentage points of Kolmogorov statistics. *J. Amer. Statist. Assoc.* **51**, 111-121.

von Mises, R. (1931) Wahrscheinlichkeitsrechnung und Ihre Anwendung in de Statistik und Theoretischen Physik. Leipzig: Deuticke.

Nair, V. N. (1984) On the behavior of some estimators from probability plots. *J. Amer. Statist. Assoc.*, **79**, 823-830.

Pearson, E. S. and Stephens, M. A. (1962) The goodness-of-fit tests based on W_N^2 and U_N^2. *Biometrika*, **49**, 397-402.

Pettitt, A. N. (1976) A two-sample Anderson-Darling rank statistic. *Biometrika*, **63**, 161-168.

Paramasamy, S. (1992) On the multivariate Kolmogorov-Smirnov distribution. *Statist. Probab. Lett.*, **15**, 149-155.

Pratt, J. W. and Gibbons, J. D. (1981) *Concepts of Nonparametric Theory*, Springer-Verlag, New York.

Rama K. Y. S. (1993) On tail probabilities of Kolmogorov-Smirnov statistic based on strong mixing processes. *Statist. Probab. Lett.*, **16**, 369-377.

Reed, T. R. C. and Cressie, N. (1988) *Goodness-of-Fit Statistics for Discrete Multivariate Data.* Nork York: Springer-Verlad.

Scholz, F. W. and Stephens, M. A. (1987) K-sample Anderson-Darling tests. *J. Amer. Statist. Assoc.*, **82**, 918-924.

Shapiro, S. S. and Francia, R. S. (1972) Approximate analysis of variance test for normality. *J. Amer. Statist. Assoc.* **67**, 215- 216.

Shapiro, S. S. and Wilk, M. B. (1965) Analysis of variance test for normality (complete samples). *Biometrika* **52**, 591-611.

Shapiro, S. S. and Wilk, M. B. (1968) Approximation for the null distribution of the W statistic. *Technometrics* **10**, 861-866.

Shapiro, S. S., Wilk, M. B. and Chen H. J. (1968) A comparative study of various tests for normality. *J. Amer. Statist. Assoc.* **63**, 1343-1372.

Sinclair, C. D. and Spurr, B. D. (1988) Approximations to the distribution function of Anderson-Darling test statistic. *J. Amer. Statist. Assoc.* **83**, 1190-1191.

Smirnov, N. V. (1936) Sur la distribution de ω^2 (critérium de M. R. v. Mises). *C. R. Acad. Sci. Paris*, **202**, 449-452.

Smirnov, N. V. (1937) On the distribution of Mises' ω^2- criterion (in Russian). *Mat. Sb. (Nov. Ser.)*, **2**, 973-993.

Smirnov, N. V. (1939) Estimate of derivation between empirical distribution functions in two independent samples (in Russian). *Bulletin of Moskow University*, **2**, 3-16.

Spinelli, J. J. and Stephens, M. A. (1997) Cramér-von Mises tests of fit for the Poisson distribution. *Canad. J. Statist.* **25**, 257-268.

Stephens, M. A. (1963) The distribution of the goodness-of-fit statistic U_N^2. I. *Biometrika*, **50**, 303-313.

Stephens, M. A. (1964) The distribution of the goodness-of-fit statistic, U_N^2. II. *Biometrika*, **51**, 393-397.

Stephens, M. A. (1970) Use of the Kolmogorov-Smirnov, Cramér-von Mises and related statistics without extensive tables. *J. R. Statist. Soc.* **B, 32**, 115-122.

Stephens, M. A. (1974) EDF statistics for goodness-of-fit and some comparisons. *J. Amer. Statist. Assoc.* **69**, 730-737.

Stephens, M. A. and Maag, U. R. (1968) Further percentage points for W_N^2. *Biometrika*, **55**, 428-430.

Tiku, M. L. (1965) Chi-square approximations for the distributions of goodness-of-fit statistics U_N^2 and W_N^2. *Biometrika*, **52**, 630-633.

Watson G. S. (1961) Goodness-of-fit tests on a circle. *Biometrika*, **48**, 109-114.

Wolf, E. H. and Naus, J. I. (1973) Tables of critical values for a k- sample Kolmogorov-Smirnov test statistic. *J. Amer. Statist. Assoc.*, **68**, 994-997.

Zhang, J. (2001a) Powerful goodness-of-fit tests based on likelihood ratio. *J. R. Statist. Soc.* **B**, in revision.

Zhang, J. (2001b) Powerful general two-sample tests. *Technometrics*, in revision.

Zhang, J. and Wu, Y. (2001a) Beta approximation to the distribution of Kolmogorov-Smirnov statistic. *Ann. Inst. Statist. Math.*, in revision.

Zhang, J. and Wu, Y. (2001b) A family of simple distribution functions to approximate complicated distributions. *J. Statist. Com. Sim.*, to appear.