SOME CONTRIBUTIONS TO THE ANALYSIS OF CIRCULAR DATA

By

J. S. RAO

A Thesis submitted to the Indian Statistical Institute in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Calcutta

February, 1969

ACKNOWLEDGEMENTS

It is with great pleasure I express my deep sense of gratitude to Prof. C. R. Rao, F.R.S., Director of the Research and Training School, Indian Statistical Institute, who has introduced me to this field and under whose supervision and constant encouragement this work has been carried out. I would also like to express my indebtedness to Prof. J. Sethuraman for the numerous valuable discussions and suggestions and for his permission to include some of the results of our joint work on the Pitman efficiencies of spacings tests.

I would like to take this opportunity to express my sincere thanks to all the members of the faculty of the Research and Training School whose help I have received at various stages of the work. Finally I wish to thank Mr. Gour Mohon Das for his efficient typing.

J. S. RAO

C O N T E N T S

CHAPTER I

INTRODUCTION AND REVIEW

1.1 Circular data - examples and special statistical problems

In many diverse scientific fields, one comes across experiments where the basic variable under observation is a direction. We shall be concerned, throughout this thesis, only with directions in two-dimensions and we refer to such a collection of directions as 'circular data' or 'directional data'. In order to make a statistical analysis of circular data, the first prerequisite is to put them in a quantitative form. One way of doing this is to represent the directions as 'angles' measured with respect to some suitably chosen zero direction. Alternately, since a direction has no magnitude, it can be represented as a unit vector i.e., as a point on the circumference of a unit circle in two dimensions. Neither of these representations for a direction is unique or absolute since the angular value assigned to a direction would depend on the choice of the zero direction as much as the components of the unit vector depend on the coordinate system chosen for representing the direction. Therefore, it is important to see that the conclusions arrived at on

analysing the circular populations, are independent of these arbitrary values assigned to the directions.

Examples of situations giving rise to circular data can be found in various scientific fields. For example, geologists study the orientations of cross-beddings and particle long-axes to interpret the direction of depositing currents of winds or water. Studies of the directions of remanent magnetism are sometimes made to interpret palaeomagnetism and possible magnetic pole migrations during geological times. Similarly biologists working on bird-migrations perform experiments on 'homing-in-pigeons' which involve observing the direction of flight of the birds as they go out of sight after release. Besides such cases where directions are observed as a direct result of the experiment, observations on any random phenomenon occuring over time or space with a regular period of known length L, can be conceived as observations on a circle with circumference of length L. Thus the study of any periodic phenomenon, with a period of known length, gives rise to an essentially similar problem. For example, if we are studying a phenomenon supposed to have say diurnal variation, we can treat the 24 hours of a day as making up a cycle and thus obtain a distribution of the occurrences on the circumference of a circle, pooling up the

observations over several such periods if necessary. Studies of say the plane-crashes around the year or the biological rhythms of living organisms, fall in this category and can be treated as studies involving directional or circular data.

The analysis of circular data gives rise to a host of novel statistical problems and does not fit into the familiar patterns of statistical analysis on the line. Suppose the directions have been given in terms of angles, then their arithmetic mean would not, in general, give us a meaningful mean direction of the sample, nor would the usual 'standard deviation' give a good measure of dispersion, in general (c.f. Batschelet, 1965). The reason for this is simple. If the zero direction is shifted through a fixed angle, the values of the arithmetic mean and the standard deviation calculated on the basis of the new values of the observations, bear no reasonable algebraic relation to the original values since the directions are measured as angles modulo 2π. Thus it is not possible on the circle to define an arithmetic mean or standard deviation in such a way that it is invariant under rotation of the circle in itself, the equivalent of a shift of origin on the straight line. The higher moments and cumulants also suffer from the same draw-back and this deprives one, of the set of valuable analytical tools like the various generating functions and so on. Though the directions can be

represented as vectors, the usual multivariate statistical techniques can not be applied readily since the vectors are all of a restricted nature namely that they all have unit length. This restriction drastically alters the problems.

1.2 Probability distributions on the circle

Even from the distributional view point, the standard linear models like the 'normal' do not, in general, provide the appropriate statistical models for describing the circular data. We describe, here, some probability distributions of a special nature which give a better description of the angular data. We call a probability distribution in two-dimensions, which has its total probability concentrated on the circumference of the unit circle, a 'circular distribution' (CD). Clearly it is singular with respect to the two-dimensional Lebesgue measure. However, if it is absolutely continuous with respect to the Lebesgue measure on the circumference of the circle, it can be specified by its density function $f(\alpha)$, which is a periodic function (with period 2π) satisfying

$$f(\alpha) \geq 0, \quad \int_0^{2\pi} f(\alpha)\,d\alpha = 1. \tag{1.2.1}$$

In general, a CD will have one or more preferred directions (also called mean directions or polar directions) and certain dispersion around these directions. We give here a few probability models for the circular data, to which we will have occasion to refer later on.

A uniform distribution on the circle has the constant density

$$(1.2.2) \qquad f(\alpha) = 1/2\pi \quad , \quad 0 \leq \alpha < 2\pi.$$

All the directions are equally likely under this distribution and there is no preferred direction. Among the unimodal CD's, the one that occupies a place of prominence is called the circular normal distribution (CND) or von Mises distribution after von Mises (1918) who first introduced it. A random angle α, with reference to an arbitrary vector in two-dimensions is said to have a CND if it has the density

$$(1.2.3) \qquad f(\alpha) = [2\pi I_0(k)]^{-1} \exp[k \cos(\alpha - \gamma)],$$

$$0 \leq \alpha < 2\pi$$

where $0 \leq \gamma < 2\pi$ is the population mean direction and $0 \leq k < \infty$ is a parameter of concentration, large values of

k standing for more concentration around the mean direction γ. When $k = 0$, (1.2.3) gives the uniform density defined in (1.2.2). $I_0(k)$, here, is a Bessel function of purely imaginary argument and has the expansion

$$(1.2.4) \qquad I_0(k) = \sum_{r=0}^{\infty} (k/2)^{2r} (1/r!)^2 .$$

The CND is a symmetric unimodal distribution. More details regarding the distribution can be found in Gumbel et al (1953). We will discuss more about this distribution in Chapter II.

One can obtain a CD by 'wrapping a linear distribution around the unit circle'. The process of 'wrapping around the circle' amounts to reducing the linear random variable modulo 2π and adding up the probabilities that correspond to the same point on the circumference. For example by wrapping the linear normal distribution with density

$$f(x) = (\sqrt{2\pi} \cdot \sigma)^{-1} \exp[-x^2/2\sigma^2], \quad -\infty < x < \infty$$

we get the so called wrapped normal distribution with density

$$(1.2.5) \quad f(\alpha) = 1/2\pi + 1/\pi \sum_{n=1}^{\infty} \cos n\alpha \exp[-n^2\sigma^2/2].$$

And similarly a Cauchy density

$$f(x) = \frac{\sigma}{\pi\,[1 + (x/\sigma)^2]}\,, \quad -\infty < x < \infty$$

gives rise to a wrapped Cauchy distribution with density

$$(1.2.6) \qquad f(\alpha) = \frac{(1-\sigma^2)}{2\pi(1+\sigma^2 - 2\sigma\cos\alpha)}\,.$$

Both the last mentioned distributions are symmetric and unimodal and resemble the CND closely.

When the circular data has more than one preferred direction, we have to consider multimodal circular densities. For instance, the following is a bimodal density which gives rise to an axially symmetric CD

$$(1.2.7) \qquad f(\alpha) = [2\pi\, I_0(k)]^{-1} \exp\,[\,k\cos 2(\alpha - \gamma)\,].$$

Another interesting bimodal density, which can be obtained from a spherically symmetric bivariate normal distribution is given by

$$(1.2.8) \qquad f(\alpha) = \frac{\sqrt{(1-\rho^2)}}{2\pi\,[1 - \rho\,\sin 2\alpha\,]}\,.$$

If (X, Y) denote a pair of random variables having a bivari-

correlation ρ, and $X = R\cos A$, $Y = R\sin A$, then the random angle A has the density given in (1.2.8). (1.2.8) becomes a uniform distribution if and only if $\rho = 0$.

1.3 General review of the thesis

In Chapter II, we give a general method of estimating the parameters of a circular distribution and show that this method ensures us of consistent and asymptotically normal (CAN) estimators of the parameters. We then discuss the circular normal distribution (CND) from estimation and distributional view points. We verify that the maximum likelihood (ML) estimates of the CN parameters are asymptotically independent so that one can construct simple large-sample tests for any hypotheses on the parameters. We give a heuristic motivation for the circular measures of dispersion and find a simple relation between the circular and linear measure of dispersion, when all the sample points are restricted to an arc of sufficiently small length on the circumference. We then, utilise this relation to get the approximate distributional results in CN populations, which are basic to the approximate analysis of variance (see e.g. Watson (1966)) for angular populations. We discuss some sampling distributions for the CN populations and obtain a generalisation

of a conditional distribution involving the lengths of the sample resultants. Finally we derive some likelihood ratio (LR) tests for single-sample and two-sample situations in CN populations and show that some approximations to these LR tests for small and large values of k, yield reasonable and useful tests, in general. They also provide further likelihood support to some of the known approximate tests.

In Chapter III, we discuss the basic problem of finding whether or not a sample indicates a preferred direction i.e., the problem of testing for uniformity (or a uniform distribution for the observations) on the circle. Since the problem of goodness of fit on the circle is canonically equivalent to testing for uniformity, the discussion and results of this Chapter can also be related to the case of goodness of fit problems on the circle. When testing for uniformity on the basis of a grouped data, we consider an invariant version of the usual χ^2 test and find its asymptotic distribution. Further if a specific class of plausible alternatives are given, we show that a special χ^2 test (See Rao (1961)) gives rise to a test based on the length of the sample resultant. We then suggest tests based on the sample arc lengths i.e., the differences between the successive observations on the circumference. We study in particular, a statistic U_n and give a table of percentage points along with an illustrative example. Finally we find the

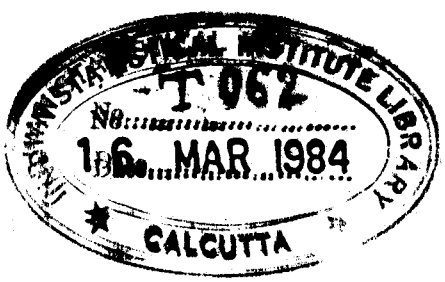
distribution of 'circular range' under uniformity, making use of the distributional results on arc lengths stated earlier in the Chapter.

In Chapter IV, we obtain the asymptotic efficiencies of tests based on arc lengths, which have been introduced in Chapter III. We derive Pitman's asymptotic relative efficiencies (ARE's) in this Chapter in a completely rigorous form. We prove many useful convergence theorems involving the empirical stochastic processes and then, appealing to these general theorems, obtain the asymptotic normality of the test statistics under the alternatives of interest. Since there is essentially no difference, in the asymptotics, between the circular and linear cases as far as the results of this Chapter are concerned, the results proved here also hold good for the spacings tests on the line. Thus the results of Chapter IV are quite general and illuminate the theory of spacings tests on the line.

In Chapter V, we compare the limiting efficiencies of the several tests that have been suggested for testing goodness of fit or uniformity on the circle. We compute the Bahadur efficiencies of the following tests: Rayleigh's test, Kuiper's test, Watson's test, Ajne's tests and a spacings test, U_n. The results of a small Montecarlo study, which sho that the spacings test U_n has quite satisfactory performance in small samples, are also reported in this chapter.

In Chapter VI, we go into the two-sample problem and discuss two nonparametric tests for testing the identity of two circular populations on the basis of independent samples from them. We also evaluate the asymptotic relative efficiency (ARE) of one test against the other.

In the final Chapter, we give some large sample homogeneity tests for comparing several angular populations with respect to their mean directions and dispersions. These tests do not assume any specific distribution for the observations and are quite robust. In testing the equality of polar directions, we show that the homogeneity test, besides being valid without any restrictions on the concentration parameters of the different populations, is asymptotically as efficient as the F-test due to Watson (1956, 1966). We then give the large sample standard errors of estimates of the parameters, if the underlying distribution is circular normal. Finally we illustrate these tests by means of a numerical example.



CHAPTER II

A GENERAL METHOD OF ESTIMATING THE PARAMETERS OF A CIRCULAR DISTRIBUTION AND SOME RESULTS CONNECTED WITH THE CIRCULAR NORMAL DISTRIBUTION

2.1 Estimating the parameters of a circular distribution

Let $\alpha_1, \ldots, \alpha_n$ be n independent observations from a parametric family of unimodal circular densities, $f(\alpha/\gamma, \rho)$ with parameters γ and ρ. Generally, γ gives the polar direction and ρ, the concentration. The maximum likelihood (ML) method of estimating the parameters, often turns out to be very much involved and the likelihood equations some times become intractable. In this section, we suggest an alternative method of estimating the parameters of a circular distribution (CD) which ensures us of consistent and asymptotically normal (CAN) estimators of the parameters.

First represent the observations as unit vectors

$$u_i = (\cos \alpha_i, \sin \alpha_i) = (x_i, y_i), \qquad i = 1, \ldots, n \tag{2.1.1}$$

and write

$$V = \sum_1^n \cos \alpha_i, \quad W = \sum_1^n \sin \alpha_i, \quad R = \sqrt{V^2 + W^2}\ . \tag{2.1.2}$$

For the parametric family under consideration, express the centre of gravity or the polar point,

$$(2.1.3) \qquad P = (E(\cos\alpha), E(\sin\alpha)) = (\xi(\gamma, \vartheta), \eta(\gamma, \vartheta)),$$

in terms of the original parameters γ and ϑ. We now estimate the parameters γ and ϑ by minimising the quantity

$$(2.1.4) \qquad D = \sum_{i=1}^{n} \left\{ (x_i - \xi(\gamma, \vartheta))^2 + (y_i - \eta(\gamma, \vartheta))^2 \right\}$$

i.e., by minimising the sum of squares of distances of the observed points u_i from the polar point P. On differentiating (2.1.4) with respect to γ and ϑ, the estimating equations become

$$(2.1.5) \qquad \sum_{1}^{n} (x_i - \xi) \frac{\partial \xi}{\partial \gamma} + \sum_{1}^{n} (y_i - \eta) \frac{\partial \eta}{\partial \gamma} = 0$$

$$\sum_{1}^{n} (x_i - \xi) \frac{\partial \xi}{\partial \vartheta} + \sum_{1}^{n} (y_i - \eta) \frac{\partial \eta}{\partial \vartheta} = 0 .$$

We might call this method of estimation, the 'minimum distance method', about which we have the following simple

Lemma 2.1.1: The estimates of γ and φ obtained by minimising (2.1.4) are consistent and asymptotically normal (CAN) provided the first order partial derivatives of ξ and η are continuou and the Jacobian

$$(2.1.6) \qquad J = \begin{vmatrix} \partial\xi/\partial\gamma & \partial\eta/\partial\gamma \\ \partial\xi/\partial\varphi & \partial\eta/\partial\varphi \end{vmatrix}$$

is non-vanishing.

Proof: If the Jacobian in (2.1.6) is non-zero, the equations (2.1.5) yield

$$(2.1.7) \qquad \bar{x} = \sum_{1}^{n} \frac{x_i}{n} = \frac{V}{n} \quad \text{and} \quad \bar{y} = \sum_{1}^{n} \frac{y_i}{n} = \frac{W}{n}$$

as estimates of ξ and η respectively. Now by the Kolmogorov's version of the strong law of large numbers (see e.g. Rao (1965)), as $n \to \infty$

$$(2.1.8) \qquad \begin{aligned} \bar{x} &\to E(x) = \xi \\ \bar{y} &\to E(y) = \eta \end{aligned}$$

with probability one. And further, the multivariate central limit theorems ensure the bivariate normality of the sample resultant $(\bar{x}, \bar{y})$. One sufficient condition for the distribution of the resultant to approach the normal law, is that all

the second order moments of x_i and y_i be finite which is trivially satisfied in our case since $x_i = \cos \alpha_i$, $y_i = \sin \alpha_i$. Thus $\bar{x}$ and $\bar{y}$ provide CAN estimators of ξ and η respectively. Now under the conditions of the Lemma, the new parameters ξ and η are one-one functions of the original parameters γ and ϑ. Further, they are invertible and the inverse transformations, say

$$\gamma = g(\xi, \eta)$$
$$\vartheta = h(\xi, \eta)$$

admit continuous first order partial derivatives. The CAN property of the estimators

$$\hat{\gamma} = g(\hat{\xi}, \hat{\eta}) \tag{2.1.9}$$
$$\hat{\vartheta} = h(\hat{\xi}, \hat{\eta})$$

is an immediate consequence of the fact that the functions g and h are continuous and totally differentiable.

As an example, consider the circular normal distribution (CND) defined in (1.2.3) for which

$$E(\cos \alpha) = \xi(\gamma, \vartheta) = \vartheta \cos \gamma \tag{2.1.10}$$
$$E(\sin \alpha) = \eta(\gamma, \vartheta) = \vartheta \sin \gamma$$

with $\varrho = \varrho(k) = I_1(k) / I_0(k)$. Hence from the estimating equations (2.1.5), we get

$$-\sum_1^n (x_i - \varrho \cos \gamma) \sin \gamma + \sum_1^n (y_i - \varrho \sin \gamma) \cos \gamma = 0$$

$$\sum_1^n (x_i - \varrho \cos \gamma) \cos \gamma + \sum_1^n (y_i - \varrho \sin \gamma) \sin \gamma = 0.$$

These equations lead to the estimates

(2.1.11) $$\hat{\gamma} = \mathrm{Tan}^{-1}(W/V)$$

(2.1.12) $$\hat{\varrho} = R/n$$

where the quadrant in which $\hat{\gamma}$ of (2.1.11) lies is determined by the signs of V and W. These estimates coincide with the ML estimates of γ and ϱ (see e.g. Gumbel et al (1953)) in the case of the CND.

When we apply this method to estimate the parameters of the wrapped Cauchy distribution with density

(2.1.13) $$f(\alpha) = \frac{(1-\sigma^2)}{2\pi (1+\sigma^2 - 2\sigma \cos(\alpha - \gamma))}$$

or the cardioid distribution with density

(2.1.14) $$f(\alpha) = \frac{(1 + 2\sigma \cos(\alpha - \gamma))}{2\pi}$$

we get similar results, $\hat{\sigma}$ being given by R/n and $\hat{\gamma}$ as in (2.1.11). This is because of the fact that the centre of gravity in both these situations is given by

$$P = (\sigma \cos \gamma, \sigma \sin \gamma).$$

The minimum distance method thus avoids the lengthy computations involved in the ML method, for instance, in the case of the densities (2.1.13) and (2.1.14). When the Jacobian is non-zero, this method is equivalent to estimating the pole P by the sample resultant, which is intuitively appealing for any unimodal CD.

In view of this, the fact that the minimum distance method and ML method coincide in the case of the CND is no surprise since the CND has been characterised (see e.g. Gumbel et al (1953)) as the distribution for which the direction of the sample resultant provides the ML estimate of the polar direction γ. The ML estimate of the concentration parameter k is obtained by solving the equation

$$\frac{I_1(\hat{k})}{I_0(\hat{k})} = R/n. \tag{2.1.15}$$

Tables for getting $\hat{k}$ from this are given in Gumbel et al (1953) and reproduced in Batschelet (1965). The ratio of the

Bessel functions $I_1(k)/I_0(k)$, we shall write as $\rho(k)$ and some times more briefly as ρ. $\rho(k)$ is a monotonic increasing function of k, increasing from zero to one as k goes from zero to infinity. When k is large

$$\rho(k) \sim 1 - \frac{1}{2k} \qquad (2.1.16)$$

where '$\sim$' means that the ratio of the two sides goes to one as $k \to \infty$. Because of this, when there is a high concentration, a fairly good estimate of k is provided by

$$\frac{n}{2(n - R)} \qquad (2.1.17)$$

By standard calculations, the asymptotic variance-covariance matrix of the ML estimates $\hat{\gamma}$ and $\hat{k}$ (inverse of the information matrix) can be shown to be

$$V = \left(\frac{1}{n}\right) \begin{bmatrix} 1/\rho k & 0 \\ 0 & 1/(1 - \rho/k - \rho^2) \end{bmatrix} \qquad (2.1.18)$$

Thus the ML estimates $\hat{\gamma}$ and $\hat{k}$ are asymptotically uncorrelated. Further, they have asymptotic normal distributions from the ML theory so that one can, when dealing with large sample construct tests of simple and composite hypotheses involving the parameters γ and k.

2.2 Circular measures of dispersion and some approximate sampling distributions in circular normal populations

In this section, we give a heuristic justification for some measures of dispersion on the circle and then using a simple relation between the circular and linear measures of dispersion, establish some approximate distributional results for circular normal (CN) populations. These results, due to Watson (1956), are basic to the approximate analysis of variance on the circle.

Let $\alpha_1, \ldots, \alpha_n$ be n independently and identically distributed angular observations. They can be written as unit vectors

$$u_i = (\cos \alpha_i, \sin \alpha_i) \qquad i = 1, \ldots, n \tag{2.2.1}$$

or as points on the circumference of the unit circle. If $A = (a, b)$ be any fixed arbitrary unit vector, a measure of distance of the observed point u_i from A is given by the 'circular' distance between u_i and the point A i.e., by the smaller of the two angles u_i makes with A, say θ_i. Clearly $0 \leq \theta_i \leq \pi$. Since for θ_i lying between 0 and π, $(1 - \cos \theta_i)$ is a monotonic function of θ_i, an algebraically more amenable measure of distance of u_i from A is given by

$d(u_i, A) = 1 - \cos \theta_i$. Therefore the quantity

$$D_A(u_1,\ldots,u_n) = \sum_1^n d(u_i, A) \tag{2.2.2}$$

may be taken as a reasonable measure of sample dispersion as measured from A. This quantity takes large values when the observations u_i are far removed from A and inversely. Now if we look upon (2.2.2) as a function of the arbitrary vector A, from which the dispersion has been measured, the natural question that arises is about the choice of A. It is simple to verify that the quantity (2.2.2) is minimum, for given $(u_1,\ldots, u_n)$, when A is

$$A^* = \left(\sum_1^n \cos \alpha_i/R, \quad \sum_1^n \sin \alpha_i/R \right) \tag{2.2.3}$$

which is the vector resultant, appropriately normalised. This minimum value corresponding to A^* is

$$D_{A^*}(u_1,\ldots,u_n) = n - R \tag{2.2.4}$$

which, thus, represents the sample dispersion about the estimated mean direction i.e., the direction of the resultant. This quantity $(n-R)$ lies between 0 and n taking large values when dispersion is high and small values when the dispersion is small and serves as a useful measure of dispersion on the circle.

Suppose, on the other hand, the polar vector $P = (\cos \gamma, \sin \gamma)$ is known. Then it is natural to measure the sample dispersion about this point P and then the quantity D, given in (2.2.2), turns out to be

$$(2.2.5) \qquad D_P(u_1, \ldots, u_n) = n - \sum_1^n \cos(\alpha_i - \gamma) = n - V^*$$

where V^* is the length of the projection of the resultant on the polar direction. This quantity $(n - V^*)$ may be called the 'sample dispersion about the population mean direction'. The quantities $(n - R)$ and $(n - V^*)$ have been used as measures of dispersion in the approximate analysis of dispersion due to Watson (1956, 1966). The relations (2.2.2), (2.2.4) and (2.2.5) serve to illustrate the analogy between circular measures of dispersion and the measures on the line where we consider

$$(2.2.6) \qquad D_a(x_1, \ldots, x_n) = \frac{1}{n} \sum_1^n (x_i - a)^2$$

for measuring the dispersion of the sample $(x_1, \ldots, x_n)$ from an arbitrary point 'a'.

We have remarked earlier that the quantity

$$s^2 = [\sum_1^n \alpha_i^2 - (\sum_1^n \alpha_i)^2/n]/n$$

computed directly from the observations $\alpha_1,\ldots,\alpha_n$ is not in general, a good measure of scatter in the case of the circle. However, it can serve as a useful measure of dispersion in situations, where the sample points are all restricted to an arc of sufficiently small length on the circumference of the circle. Several attempts have been made to link this quantity s^2 with the circular measures of dispersion (see e.g. Pincus (1956), Batschelet (1965)). s^2 is easier to compute than $(n - R)$ and a relation between the two may be desirable in situations where both the quantities are meaningful.

Suppose the observations are all restricted to an arc of sufficiently small length. Then, by a suitable choice of the origin, one can assume that the angular values themselves are small. Suppose $\theta_1,\ldots,\theta_n$ are the observations and the θ's, in radian measure, are so small that powers of θ of order greater than two can be ignored. Then

$$\begin{aligned} R^2/n^2 &= n^{-2}\,[(\sum_1^n \cos\theta_i)^2 + (\sum_1^n \sin\theta_i)^2] \\ &\doteq n^{-2}\,[(n - \sum_1^n \theta_i^2/2)^2 + (\sum_1^n \theta_i)^2] \\ &\doteq 1 - [\sum_1^n \theta_i^2/n - (\sum_1^n \theta_i/n)^2] \qquad (2.2.7) \\ &= 1 - s_\theta^2 \end{aligned}$$

where the symbol, '≐' here as elsewhere stands for 'approximately equal to'. Thus the sample variance s_θ^2 of the θ's is related to the length of the vector resultant, R by the relation (2.2.7). Pincus (1956), based on empirical studies, concludes that s^2 and R^2/n^2 have a high negative correlation and Reiche (1938) finds that $\bar{a} = R/n$ and s have a correlation of -0.87 for the samples he considered. These results are only natural in view of the relation (2.2.7).

Now suppose that the observations θ_i come from a circular normal (CN) population, with a high value of the concentration parameter k. Then using the relation (2.2.7), we show that

$$(2.2.8) \qquad \begin{aligned} 2k(n-R) &\sim \chi^2_{n-1} \\ 2k(n-V^*) &\sim \chi^2_n \end{aligned}$$

where '$\sim$' stands for 'approximately distributed as' and χ^2_p denotes a chi-square distribution with p degrees of freedom. If α has a CN distribution with parameters γ and k, then it is known that $\beta = \sqrt{k}\,(\alpha - \gamma)$ approaches a standard normal variate for k large (c.f. Gumbel et al (1953)). The CN parameter k, thus plays the role of $1/\sigma^2$ where σ^2 is the variance of α when it is treated as a linear normal variate. Some further similarities regarding k and

$1/\sigma^2$ are given in Gumbel et al (1953). Further, for k large, R is near n with high probability, so that the relation (2.2.7) gives

$$(2.2.9) \qquad s^2 = \frac{(n^2 - R^2)}{n^2} \doteq 2(n-R)/n.$$

Thus for k large, from (2.2.9) and the fact that $k = 1/\sigma^2$

$$2k(n-R) \doteq ns^2/\sigma^2$$

which from the normal distribution theory, is known to have a χ^2_{n-1} distribution. Similarly if the polar direction γ is known, writing

$$(2.2.10) \qquad s^{*2} = \sum_1^n (\alpha_i - \gamma)^2/n$$

and using similar arguments, it can be shown that

$$(2.2.11) \qquad 2k(n - V^*) \doteq ns^{*2}/\sigma^2$$

which, from the known linear normal theory, has a χ^2_n distribution. Thus the approximate distributional results (2.2.8), may be established following an essentially linear approach, which is not quite so apparent in the derivation of Stephens (1962). It may also be observed that the estimate of σ^2 viz. s^2 given in (2.2.9) is the reciprocal of the estimate of k given in (2.1.17) for large k . This provides a justification for making an approximate variance-ratio F-test for testing the

hypothesis $k_1 = k_2$ in two CN populations or for making use of the Hartley's maximum F-ratio for testing the equality of concentrations for several populations.

2.3 Some sampling distributions associated with the CND and a generalisation

The density of the length R of the resultant of n 'random' unit vectors (unit vectors with the direction having a uniform density) is well known (see e.g. Greenwood and Durand (1955)) and is given by

$$(2.3.1) \qquad f_o(R) = R \int_0^\infty J_o^n(t)\, J_o(Rt)\, t\, dt, \qquad 0 \leq R \leq n$$
$$= R.\ \psi_n(R), \quad \text{say}$$

where $J_o(t)$ is the Bessel function of first kind. We will use the suffixes 'k' and 'o' for the density functions (or sampling distributions) to indicate that the observations come from a CND with parameter $k > 0$ and from a uniform distribution corresponding to $k = 0$ respectively. When n is large, an approximation due to Rayleigh (1919) for $f_o(R)$ is given by

$$(2.3.2) \qquad f_o(R) \doteq (2R/n) \exp[-R^2/n]$$

which implies that $(2R^2/n)$ is distributed as a χ^2 with two degrees of freedom.

If β denotes the angle between the zero direction and resultant of n 'random' unit vectors, then β is uniformly distributed and is distributed independently of R. Hence the joint density of R and $c = \cos\beta$ is

$$(2.3.3) \qquad f_o(R, c) = R(\sqcup\!\!\sqcup_n)(R)/\pi \sqrt{1 - c^2}, \quad 0 \leq R \leq n, \quad -1 \leq c \leq 1.$$

Now the density of R and c for observations from a CND with $k > 0$, may be obtained by utilising the fact that Rc is sufficient for k. Therefore, we get

$$(2.3.4) \qquad f_k(R,c) = \frac{e^{kRc}}{I_o^n(k)} f_o(R, c) = \frac{e^{kRc}}{I_o^n(k)} \frac{R(\sqcup\!\!\sqcup_n)(R)}{\pi\sqrt{1 - c^2}}, \quad 0 \leq R \leq n, \ -1 \leq c \leq 1.$$

From this, one can obtain the distribution of R, when the observations come from a CND with concentration parameter $k \geq 0$, namely

$$(2.3.5) \qquad f_k(R) = \frac{I_o(kR)}{I_o^n(k)} R(\sqcup\!\!\sqcup_n)(R).$$

Suppose now that R_1 and R_2 denote the lengths of the resultants based on samples of sizes n_1 and n_2 respectively from two CN populations with the same concentration k. If R denotes the length of the overall resultant based on $n = n_1 + n_2$ observations, Watson and Williams (1956) give the conditional density of R_1 and R_2 given R and this is

$$(2.3.6) \quad f_k(R_1, R_2 | R) = \frac{2R_1R_2\left(\psi\right|_{n_1}(R_1)\left(\psi\right|_{n_2}(R_2)}{\pi\left(\psi\right|_{n}(R)\sqrt{(R_1+R_2)^2 - R^2)(R^2-(R_1-R_2)^2)}},$$

$$0 \leq R_i \leq n_i, \; i = 1,2, \quad |R_1 - R_2| \leq R \leq R_1 + R_2 .$$

Its importance lies in the fact that it is independent of the parameters and can be utilised for constructing significance tests for CN populations. Watson and Williams (1956) suggested a conditional test based on the sum $S = R_1 + R_2$ given R, for testing the hypothesis of equality of polar directions for two CN populations with common concentration parameter k. The critical point s_o is to be chosen so as to satisfy the equation

$$P(S \geq s_o \mid R) = \alpha.$$

Similarly if k_1 and k_2 denote the concentrations of the two CN populations, an exact conditional test of the hypothesis of equality of concentrations can also be made using

(2.3.6) as this gives the null distribution under the hypothesis $k_1 = k_2$. The appropriate test statistic is the difference $D = \left|\frac{R_1}{n_1} - \frac{R_2}{n_2}\right|$ which simplifies to $D = |R_1 - R_2|$ in the case of equal sample sizes. The critical difference $d_o = d_o(n, R, \alpha)$ is to be obtained by solving the equation

$$P(D \geq d_o \mid R, n) = \alpha.$$

Work on the construction of the nomograms for these situations is in progress. We now obtain here a generalisation of the conditional density (2.3.6) for the case of q ($q \geq 2$) samples.

Suppose we have q samples of sizes $n_1, \ldots, n_q$ from CN populations with the same concentration. Let $R_1, \ldots, R_q$ denote the lengths of individual sample resultants based on the $n_1, \ldots, n_q$ observations respectively of the q samples and let R denote the length of the overall resultant based on $N = \sum_{i=1}^{q} n_i$ observations. We derive, here, the conditional distribution of $R_1, \ldots, R_q$ given R. Following Watson and Williams (1956), we obtain as a first step, the joint density of $R_1, \ldots, R_q$, R and c i.e., $f_o(R_1, \ldots, R_q, R, c)$, where the subscript 'o' for the density denotes the density when $k = 0$, i.e., under uniformity. Once this is known, the density for $k > 0$ can again be obtained using the relation

$$(2.3.7) \qquad f_k(R_1, \ldots, R_q, R, c) = \frac{e^{kRc}}{I_o^N(k)} f_o(R_1, \ldots, R_q, R, c).$$

Now to obtain the joint density $f_o(R_1,\ldots,R_q, R,c)$, we shall make use of the conditional density $f_o(R|R_1,\ldots,R_q)$. It is difficult to obtain this straight forwardly as has been done for $q = 2$ by Watson and Williams (1956), though one can express R explicitly as a function of $R_1,\ldots,R_q$ and the angles in between them. Using a special case of Weber's discontinuous factor theorem (see e.g. Watson, G.N. (1944)), it can be shown that

$$(2.3.8) \qquad P_o(R \leq r \mid R_1,\ldots,R_q) = r \int_0^\infty J_1(r\,t)\Big[\prod_{i=1}^{q} J_o(R_i t)\Big]dt$$

where R is the resultant of q vectors of lengths $R_1,\ldots,R_q$, any two of them making a 'random' angle in between them. This is a generalisation of Kluyver's solution of the famous Pearson's random walk problem (see e.g. Watson, G.N. (1944)). Now $f_o(r \mid R_1,\ldots,R_q)$ is obtained from this by a single differentiation with respect to r and we get

$$(2.3.9) \qquad f_o(r \mid R_1,\ldots,R_q) = r \int_0^\infty J_o(r\,t)\Big[\prod_{1}^{q} J_o(R_i t)\Big]t\,dt\,.$$

Since β, the angle made by the resultant with the zero direction, is rectangularly distributed when $k = 0$, $c = \cos\beta$ has the density

$$(2.3.10)\,. \qquad f_o(c) = 1/\pi\sqrt{1-c^2}\,, \qquad -1 \leq c \leq 1.$$

This c is further, distributed independently of the lengths of the resultants $R_1,\ldots,R_q$ and R. The density of R_i, which is the length of the resultant of n_i random unit vectors is given by

$$(2.3.11)\qquad f_o(R_i) = R_i \int_0^\infty J_o(R_i t)[J_o(t)]^{n_i}\, t\,dt$$

$$= R_i \left(\bigcup\right)_{n_i}(R_i).$$

Now from (2.3.9), (2.3.10) and (2.3.11) and because of the independence of c with the lengths $R_1,\ldots,R_q$ and R when $k = 0$, we have

$$(2.3.12)\qquad f_o(R_1,\ldots,R_q,R,c) = \frac{\left\{R\int_0^\infty J_o(Rt)\left[\prod_{i=1}^{q} J_o(R_i t)\right]t\,dt\right\}\left\{\prod_{i=1}^{q} R_i \left(\bigcup\right)_{n_i}(R_i)\right\}}{\pi\sqrt{1-c^2}}.$$

Using the relation (2.3.7) and integrating out c in the density $f_k(R_1,\ldots,R_q, R, c)$, we obtain

$$(2.3.13)\qquad f_k(R_1,\ldots,R_q,R) = \frac{I_o(kR)}{I_o^N(k)}\left\{R\int_0^\infty J_o(Rt)\left[\prod_{i=1}^{q} J_o(R_i t)\right]t\,dt\right\}\cdot \prod_{i=1}^{q} R_i \left(\bigcup\right)_{n_i}(R_i).$$

And since the density of R for $k > 0$, is known, from (2.3.5), to be

$$(2.3.14) \qquad f_k(R) = \frac{I_0(kR)}{I_0^N(k)} \cdot R \left(\frac{1}{N}\right)(R)$$

we find the required conditional density by dividing (2.3.13) by (2.3.14) to get

$$(2.3.15) \quad f_k(R_1,\ldots,R_q|R) = \frac{\left\{\int_0^\infty J_0(Rt)\left[\prod_{i=1}^{q} J_0(R_i t)\right]t\,dt\right\}\prod_{i=1}^{q} R_i\left(\frac{1}{n_j}\right)(R_i)}{\left(\frac{1}{N}\right)(R)}$$

This expression, which generalises the result of Watson and Williams (1956), can be utilised for constructing exact conditional tests for q (≥ 2) populations, since it is independent of the nuisance parameters. For instance, in testing the equality of polar directions of q CN populations, given that they all have the same concentration, we can use the test statistic $S = R_1 + \ldots + R_q$, large values of S, for given R, being critical. Similarly the distribution of any test function for testing the equality of concentrations will be free of k, if it is obtained from (2.3.15). The expression (2.3.15) looks apparently quite involved. An interesting

approximation, in this connection is given in Uspensky (1937, p. 329) which says

$$(2.3.16)\quad \int_0^\infty J_0(Rt)\Big[\prod_{i=1}^{q} J_0(R_i t)\Big]t\,dt \doteq \frac{2}{\left(\sum_1^q R_i^2\right)} \exp\Big[-R^2/\sum_1^q R_i^2\Big]$$

if $\sum_1^q R_i^3/\left(\sum_1^q R_i^2\right)^{3/2} \to 0$ as $q \to \infty$. If the vectors $R_1,\ldots, R_q$ are each of unit length, this is the same as the Rayleigh's approximation given in (2.3.2).

In particular, when $q = 2$, the expression (2.3.15) gives

$$(2.3.17)\quad f_k(R_1,R_2|R) = \frac{R_1 \bigcirc_{n_1}(R_1) R_2 \bigcirc_{n_2}(R_2)}{\bigcirc_N(R)} \int_0^\infty J_0(Rt)J_0(R_1t)J_0(R_2t)\,t\,dt.$$

The integral involved here, can be evaluated using equation (3) on p. 411 of Watson G.N. (1944) and we have

$$(2.3.18)\quad \int_0^\infty J_0(Rt)J_0(R_1t)J_0(R_2t)\,t\,dt = 1/2\pi A$$

where A is the area of the triangle whose three sides are of

lengths R, R_1 and R_2. This, from a standard formula, is

$$(2.3.19) \quad A = \sqrt{s(s-R)(s-R_1)(s-R_2)} \quad \text{where} \quad s = (R+R_1+R_2)/2$$

$$= (\tfrac{1}{4}) \sqrt{(\overline{R_1+R_2}^2 - R^2)(R^2 - \overline{R_1-R_2}^2)} .$$

From (2.3.17), (2.3.18) and (2.3.19),

$$f_k(R_1,R_2|R) = \frac{2R_1 \left(\downarrow\right)_{n_1}(R_1) R_2 \left(\downarrow\right)_{n_2}(R_2)}{\pi \left(\downarrow\right)_N(R) \sqrt{(\overline{R_1+R_2}^2 - R^2)(R^2 - \overline{R_1-R_2}^2)}} .$$

There seems to be an omission of the factor 2 in the numerator, in the derivation of Watson and Williams (1956). In fact the conditional density $f_o(R|R_1,R_2)$ which they denote by $G(R_1,R_2,R)$, over the appropriate range, integrates only to 1/2. Similarly when $q = 3$, the integral involved in (2.3.15) can be expressed in terms of complete elliptic integrals of the first kind, using equation (9) on p. 414 of Watson (1944).

2.4. Likelihood ratio tests for circular normal populations and some approximations

In this section, we derive some likelihood ratio (LR) tests for circular normal (CN) populations in one-sample and two-sample situations and show that their approximations for large or small values of k, turn out to be reasonable tests of the corresponding hypotheses, in general. In some cases, they coincide with the known approximate tests derived by other means, thus giving further likelihood support to the latter. And in all cases, the approximations used give simple and meaningful tests. Some comments, regarding the range of situations where the tests given here may be used, are given at the end of the section. Consider, first, the one-sample case, where we have n independent observations from a CND with parameters γ and k.

(a) Test for $H_1: k = 0$:

This is a test of the hypothesis of no concentration or uniformity. The LR for the hypothesis H_1 turns out to be (ref. e.g. Greenwood and Durand (1955))

$$\lambda_1 = I_0^n(\hat{k}) \exp[-\hat{k} R] \tag{2.4.1}$$

where R is the length of the sample resultant and $\hat{k}$ is the

ML estimate of k. This, for fixed n, is a decreasing function of R alone and hence is equivalent to a test in which large values of R are critical. The null distribution of R is given by (2.3.1) and Greenwood and Durand (1955) give a short table of percentage points for this test. When k is small, we use the approximations

$$I_0(k) \doteq 1 + k^2/4 \tag{2.4.2}$$

$$g(k) = I_1(k)/I_0(k) \doteq k/2 \tag{2.4.3}$$

so that for small values of k, the ML estimate of k (from (2.1.15) and (2.4.3)) is given by

$$\hat{k} \doteq 2R/n \tag{2.4.4}$$

and hence

$$\log I_0(\hat{k}) \doteq R^2/n^2. \tag{2.4.5}$$

Using these approximations (2.4.4) and (2.4.5) in (2.4.1),

$$\begin{aligned} -2 \log \lambda_1 &= -2n \log I_0(\hat{k}) + 2 \hat{k} R \\ &\doteq 2R^2/n \end{aligned} \tag{2.4.6}$$

which is the statistic of the well-known Rayleigh's test (see

also Section 3.1). Under H_1, $-2 \log \lambda_1 = 2R^2/n$ has a χ^2_2 distribution, which agrees with the Rayleigh's approximation given in (2.3.2).

(b) Test for H_2: $\gamma = \gamma_o$.

This is the hypothesis of prescribed polar vector for which the LR is

$$\lambda_2 = [I_o(\hat{k})/I_o(\hat{k}')]^n \exp [\hat{k}' V_o^* - \hat{k} R]$$

where $V_o^* = \sum_1^n \cos(\alpha_i - \gamma_o)$ and $\hat{k}$ and $\hat{k}'$ are the solutions of $g(k) = R/n$ and V_o^*/n respectively. For k sufficiently large using the approximation

$$I_o(k) \doteq e^k / \sqrt{2\pi k} \qquad (2.4.7)$$

along with the approximation (2.1.17) for $\hat{k}$,

$$\begin{aligned} -2 \log \lambda_2 &\doteq 2\hat{k}(R-n) + 2\hat{k}'(n-V_o^*) + n \log(\hat{k}/\hat{k}') \\ &\doteq n \log [(n - V_o^*)/(n-R)] \\ &= n \log [1 + (R - V_o^*)/(n-R)] \\ &= n [(R - V_o^*)/(n-R) - (R-V_o^*)^2/2(n-R)^2 + \ldots]. \end{aligned}$$

This has a χ^2_1 distribution under H_2. Since for large k,

V_0^* is near to R with high probability, one can ignore terms other than the first in (2.4.8). This gives the approximate test statistic suggested by Watson (1956, 1966) for testing H_2. Our arguments show that this test has the flavour of a LR test, for large values of k . On the other hand, if k is small, the approximations (2.4.4) and (2.4.5) give

$$(2.4.9) \qquad -2 \log \lambda_2 \doteq 2(R^2 - V_0^{*2})/n$$

which again has a χ_1^2 distribution in large samples. This statistic (2.4.9), with χ_1^2 distribution has been utilised earlier by Stephens (1962, test 3.3). Our derivation again provides an alternate argument and support for the same on the basis of the LR principle.

Coming to the two-sample situation, suppose we have two CN populations say with parameters (γ_1, k_1) and (γ_2, k_2) respectively. Suppose we have n_1 independent observations from the first and n_2 independent observations from the second and let $n_1 + n_2 = n$. Let R_1 and R_2 denote the lengths of the individual sample resultants and R the length of the overall resultant.

(c) Test for H_3: $k_1 = k_2$.

This is a test of the hypothesis of equal concentrations. If k_1 and k_2 are small enough, the approximations (2.4.4)

and (2.4.5) yield

$$(2.4.10) \qquad -2 \log \lambda_3 \doteq 2R_1^2/n_1 + 2R_2^2/n_2 - 2(R_1 + R_2)^2/n$$

$$= \frac{2n_1 n_2}{n} (R_1/n_1 - R_2/n_2)^2$$

where λ_3 is the LR corresponding to the hypothesis H_3. (2.4.10), under H_3, has χ_1^2 as the large sample distribution. Since the statistic (2.4.10) depends on the difference between the ratios (R_1/n_1) and (R_2/n_2) of the two samples, it may be used for testing the equality of concentrations for unimodal circular populations, in general.

(d) Test for H_4: $\gamma_1 = \gamma_2$ (k same).

This is a test of the hypothesis of the equality of polar directions given that the two CN populations have the same concentration. When the common value of k is large, the LR test for the situation yields a test based on the statistic

$$(2.4.11) \qquad (n-2)(R_1 + R_2 - R)/(n - R_1 - R_2)$$

suggested by Watson.(1956). This statistic has a F distribution with 1 and (n-2) degrees of freedom. However, when k is small, the approximations (2.4.4) and (2.4.5) give the log likelihood ratio

(2.4.12) $\quad -2 \log \lambda_4 \doteq (2/n)\cdot[(R_1 + R_2)^2 - R^2]$

which has the null distribution χ^2_1.

(e) Test for H_5: $k_1 = k_2$, $\gamma_1 = \gamma_2$.

This is a test of the identity of the two CN populations. The LR for this hypothesis turns out to be

$$\lambda_5 = I_o^{n_1}(\hat{k}_1) I_o^{n_2}(\hat{k}_2) I_o^{-n}(\hat{k}) \cdot \exp\left[\hat{k} R - \hat{k}_1 R_1 - \hat{k}_2 R_2\right]$$

where $\hat{k}_1$, $\hat{k}_2$ and $\hat{k}$ are respectively the solutions of $\vartheta(k) = R_1/n_1$, R_2/n_2 and R/n. Using the approximations for small k

(2.4.13) $\quad -2 \log \lambda_5 \doteq 2(R_1^2/n_1 + R_2^2/n_2 - R^2/n)$

$$= \frac{2n_1 n_2}{n}\left[(\bar{x}_1 - \bar{x}_2)^2 + (\bar{y}_1 - \bar{y}_2)^2\right]$$

where $(n_1 \bar{x}_1, n_1 \bar{y}_1)$ and $(n_2 \bar{x}_2, n_2 \bar{y}_2)$ denote the resultants for the two samples. The statistic in (2.4.13) has the large sample distribution χ^2_2, under the hypothesis. The statistic (2.4.13) tests the difference in the means of 'cos' and 'sin' values between the two samples and hence is reasonable, in general, for testing the discrepancy between the two populations. The χ^2_2 distribution for (2.4.13) may also be alternately justified from the fact that for k small and n large, $(\bar{x}_1 - \bar{x}_2)$

and $(\bar{y}_1 - \bar{y}_2)$ have independent normal distributions with means zero and variance $(1/2n_1 + 1/2n_2) = n/2n_1n_2$ under the hypothesis.

The hypotheses H_2 and H_4 involve the polar directions. At the same time, we assumed k to be small enough for the approximations (2.4.2) and (2.4.3) to hold. Clearly, a test involving polar directions becomes meaningful only when a preferred direction has been established i.e., when k is so large that the hypothesis of uniformity is rejected. We remark here that there is, in fact, a large range of values of k under which a preferred direction is indicated and the required approximations also hold good. For instance the approximations (2.4.2) and (2.4.3) hold with good accuracy (involve less than about 1% error) if k is less than about 0.9, which corresponds to a value of 0.4 or less for the ratio R/n. The Rayleigh test, for example, favours the hypothesis of preferred direction for values of R much less than $(0.4)n$ (In fact, for any $R \geq R_0 = \sqrt{n \log (1/\alpha)}$ at level α, using the approximate distribution (2.3.2)), provided n is large. Thus there is a wide range of situations (n large, $R_0 \leq R \leq (0.4)n$) where the approximate LR tests given in this section could profitably be utilised in fairly large samples.

CHAPTER III

TESTING FOR GOODNESS OF FIT OR UNIFORMITY ON THE CIRCLE

3.1 Introduction and Summary

One of the basic problems in the analysis of circularly distributed data is to find whether a given set of observations on the circumference of the unit circle indicate any preferred direction or whether the data can be considered to have come from a uniform distribution on the circumference. We shall assume, throughout this discussion, that the observations are given in terms of angles measured with respect to some suitably chosen origin (or zero direction), taking say, the anticlockwise direction as positive. Then a goodness of fit' problem, on the circle, is to test whether a random sample $(\alpha_1,\ldots,\alpha_n)$ comes from a population with a completely specified distribution function $G_o(\alpha)$, $0 \leq \alpha < 2\pi$. If the specified distribution function is continuous, then the points $x_i = G_o(\alpha_i)$ may be considered as observations on the circle of unit circumference, where now, the problem is to test whether the observations $(x_1,\ldots, x_n)$ come from a uniform distribution. Thus a goodness of fit problem on the circle can also be reduced to testing for uniformity on the

circle and the two problems are canonically equivalent just as they are on the line.

Broadly speaking, the test procedures available for this purpose on the line may be grouped into three categories viz., (i) the methods based on χ^2 (ii) the methods utilising the empirical distribution functions and (iii) those based on sample spacings, i.e., differences between successive order statistics. However, these methods are not, in general, directly applicable for observations on the circle because of special problems posed by the arbitrary choice of the zero direction. A test statistic should, clearly be independent of this arbitrary origin, in order that it can be meaningfully used with the circular data.

In some cases, modifications of the usual test statistics on the line, so as to make them independent of the choice of origin, have been introduced for use with the circular data. For instance, when employing the methods based on empirical distribution functions, Kuiper (1960) and Watson (1961) suggested such modifications for the standard Kolmogorov-Smirnov and Cramer-von Mises tests respectively. On the other hand, if the χ^2 methods were to be exploited for testing uniformity, one can make the usual χ^2 test for uniformity invariant under choice of origin (or equivalently, the choice

of different class intervals), by considering the maximum possible value of χ^2 (for a given number of class intervals) or by taking the average such χ^2. We obtain the asymptotic distribution of the latter, in Section 3.2, by using methods similar to those in Watson (1967). Looking at the problem from another angle, if one has a suitable class of parametric alternatives for the observations, one can improve on the usual χ^2 test by concentrating on those alternatives. We show, in Section 3.2 that a special type of χ^2, due to Rao (1961), gives a test based on the length of the sample resultant, when testing for uniformity amongst the class of 'close CN alternatives'. In Sections 3.3 to 3.6, we suggest and study the third group of tests, based on sample arc lengths, which correspond to the spacings tests on the line (c.f. Greenwood (1946), Kimball (1950), Sherman (1950), Darling (1953), Pyke (1965) etc.). We consider, in particular, a test suggested by Kendall (c.f. Discussion on Greenwood (1946)) and studied by Sherman (1950) which has a particularly nice interpretation in the case of the circle. A table of percentage points and a numerical example illustrating the simplicity in using the test, are given in Section 3.5. Finally in Section 3.6, we find the null distribution of the 'circular range', the length of the smallest arc containing all the sample observations, using the distributional results of Section 3.3

A remark regarding the wider validity of the tests based on arc lengths than, for example, the classical Rayleigh's test, is in order. The classical method for testing uniformity of circular data is based on the length of the sample resultant R, large values of R being critical. This test is known as the Rayleigh's test (c.f. eg. Batschelet (1965)) and is valid, in general, against any unimodal circular density but should not be used when multimodal alternatives are suspected since R can be small not only under uniformity but also under a symmetric multimodal alternative. But the arc length tests, suggested here, are clearly valid over a much wider class of alternatives as they detect any sort of clustering of the observations on the circumference. Pyke (1965, Discussion) remarks that the tests based on spacings would be sensitive to differences in density functions whereas the methods based on empirical distribution functions detect significant differences in distribution functions. Apart from such general comments, no effort has yet been made to compare the efficiencies of the spacings-tests with others. In the next chapter, we study in detail, the asymptotic relative efficiencies of the tests based on arc lengths.

3.2 Some variants of the χ^2 for testing uniformity

In this section, we consider the problem of testing for uniformity on the basis of a grouped data, by using the χ^2 methods which involve comparing the observed frequencies with those expected. The value of the usual χ^2 statistic depends, in general, on the particular grouping adopted and we therefore suggest an average type of χ^2 and find its asymptotic distribution (a.d.). We then consider this problem from a different view point. Suppose a specific class of parametric distributions can be considered as plausible alternatives to uniformity. One may then compare, as we do in this section, the estimated frequencies under the alternative with those under the hypothesis by using a χ^2 statistic suggested by Rao (1961).

Suppose the circular data consisting of n independent observations $\alpha_1, \ldots, \alpha_n$ is grouped into m class intervals of equal width, the first class starting with a suitably chosen direction α. The number of classes, m, is held fixed throughout our discussion. Let the i^{th} class interval be $I_i(\alpha) = [\alpha + (i-1)2\pi/m, \alpha + i2\pi/m) = [\alpha^i, \alpha^{i+1})$, say, for $i = 1, \ldots, m$. Suppose $n_i = n_i(\alpha)$ is the number of observations that fall in $I_i(\alpha)$ and let $n = \sum_1^m n_i$ be the total number of observations. For testing uniformity on the basis of this grouped data, the usual χ^2 statistic with equal expected

frequencies under the hypothesis is

$$\chi_n^2(\alpha) = \sum_1^m [n_i(\alpha) - n/m]^2/(n/m)$$

(3.2.1)

$$= n \sum_1^m [p_i(\alpha) - \pi_i^0(\alpha)]^2/\pi_i^0(\alpha)$$

where $p_i(\alpha) = n_i(\alpha)/n$ and $\pi_i^0(\alpha) = 1/m$ denote the observed and hypothetical relative frequencies in the i^{th} class. The statistic given in (3.2.1) has, for n large, a $\chi^2_{(m-1)}$ distribution under the hypothesis of uniformity. But this statistic clearly depends on the particular grouping adopted or in other words on the starting point α. However, this dependence, it may be remarked, is not peculiar to observations on the circle alone.

The χ^2 statistic (3.2.1) can be made independent of α (or the particular choice of grouping) by considering, for instance, $\sup_\alpha \chi_n^2(\alpha)$ or $\int_0^{2\pi} \chi_n^2(\alpha)d\alpha$. We shall now find the asymptotic distribution (a.d.) of the statistic

$$\chi_n^2 = \frac{1}{2\pi} \int_0^{2\pi} \chi_n^2(\alpha)\, d\alpha$$

(3.2.2)

$$= \frac{m}{2\pi n} \int_0^{2\pi} \sum_{i=1}^m [n_i(\alpha) - \frac{n}{m}]^2 d\alpha$$

under the hypothesis of uniformity. (Here the subscript n

for χ^2 is used to denote the number of observations on which it is based). The a.d. of χ^2_n can be evaluated by adopting the following standard method. First, it can be established that the empirical process $\{\frac{1}{\sqrt{n}}(n_i(\alpha) - \frac{n}{m}),$ $i = 1,\ldots, m,\ 0 \leq \alpha < 2\pi\}$ converges to a m-variate Gaussian process on $[0, 2\pi)$. Then appealing to the invariance principle, the a.d. of χ^2_n is the same as that of the limiting statistic χ^2, expressed in terms of the Gaussian process. The Fourier representation of this Gaussian process reduces this χ^2 to an infinite summand of m-variate complex Laplacian variables, whose theory is by now well known (See eg. Goodman (1963)). A more elementary but essentially equivalent approach is given in Watson (1967) and we adopt this approach to find the a.d. in our case. Define the indicator random variables.

$$(3.2.3)\qquad \chi^i_j(\alpha) = \begin{cases} 1 & \text{if } \alpha^i \leq \alpha_j < \alpha^{i+1} \\ & \text{i.e., if the } j^{th} \text{ observation,} \\ & \alpha_j \in I_i(\alpha) \\ 0 & \text{otherwise} \end{cases}$$

for $j = 1,\ldots, n$ and $i = 1,\ldots, m$. Then

$$(3.2.4)\qquad (n_i(\alpha) - \frac{n}{m}) = \sum_{j=1}^{n} (\chi^i_j(\alpha) - \frac{1}{m})$$

which, being a periodic function, may be expressed in Fourier

$$(3.2.5)\qquad (n_i(\alpha) - \frac{n}{m}) = a_{io} + \sum_{\ell=1}^{\infty} (a_{i\ell}\cos \ell\alpha + b_{i\ell} \sin \ell\alpha)$$

where

$$a_{io} = \frac{1}{2\pi} \int_0^{2\pi} \sum_j (\chi_j^i(\alpha) - \frac{1}{m}) d\alpha = 0$$

and for $\ell \neq 0$

$$a_{i\ell} = \frac{1}{\pi} \sum_j \int_0^{2\pi} (\chi_j^i(\alpha) - \frac{1}{m}) \cos \ell\alpha \, d\alpha$$

$$(3.2.6)\qquad = \frac{1}{\pi} \sum_j \int_{\alpha_j - \frac{2\pi i}{m}}^{\alpha_j - \frac{2\pi(i-1)}{m}} \cos \ell\alpha \, d\alpha$$

since

$$(3.2.7)\qquad \chi_j^i(\alpha) = \begin{cases} 1 \text{ if } \alpha_j - \frac{2\pi}{m} \leq \alpha^i < \alpha_j \text{ or} \\ \qquad \alpha_j - \frac{2\pi i}{m} \leq \alpha < \alpha_j - \frac{2\pi(i-1)}{m} \\ 0. \text{ otherwise.} \end{cases}$$

Thus, from (3.2.6),

$$(3.2.8)\qquad a_{i\ell} = \frac{2 \sin \frac{\ell\pi}{m}}{\ell\pi} \sum_{j=1}^{n} \cos \ell \, (\alpha_j - \frac{2\pi(i-1)}{m}).$$

Similarly $b_{i\ell}$, the coefficient of $\sin \ell\alpha$, can be shown to be

$$(3.2.9)\qquad b_{i\ell} = \frac{2 \sin \frac{\ell\pi}{m}}{\ell\pi} \sum_{j=1}^{n} \sin \ell \, (\alpha_j - \frac{2\pi(i-1)}{m}).$$

Now for a fixed ℓ, consider the set of $2m$ coefficients

$\{ (a_{i\ell}, b_{i\ell}), i = 1, \ldots, m \}$ given in (3.2.8) and (3.2.9).

Since the α_j are independently and uniformly distributed on $[0, 2\pi)$, it is easy to check that

$$E(a_{i\ell}) = E(b_{i\ell}) = 0$$

(3.2.10)

$$E(a_{i\ell}\, a_{j\ell}) = E(b_{i\ell}\, b_{j\ell}) = \frac{2n \sin^2 \frac{\ell\pi}{m}}{\ell^2 \pi^2} \cdot \cos (j-i) \frac{2\pi\ell}{m}$$

$$E(a_{i\ell} b_{j\ell}) = \frac{2n \sin^2 \frac{\ell\pi}{m}}{\ell^2 \pi^2} \sin (i-j) \frac{2\pi\ell}{m}$$

for $i,j = 1, \ldots, m$. Further, for every fixed ℓ, when n is large, by the multivariate central limit theorem, the random vector of $2m$ variables

(3.2.11)

$$\eta'_{n\ell} = \frac{1}{\sqrt{n}} \{ a_{1\ell}\ b_{1\ell}\ a_{2\ell}\ b_{2\ell} \cdots a_{m\ell}\ b_{m\ell} \}$$

converges in distribution to a random vector η'_ℓ which has a $2m$ variate normal distribution with means zero and variance covariance matrix

$$\underset{(2m\times 2m)}{\Sigma_{\ell}} = E\eta_{\ell}\eta_{\ell}' = \left(\frac{2\sin^2\frac{\ell\pi}{m}}{\pi^2\ell^2}\right)\times$$

$$\begin{bmatrix} 1 & 0 & \cdots\cdots & \cos(m-1)\frac{2\pi\ell}{m} & -\sin(m-1)\frac{2\pi\ell}{m} \\ 0 & 1 & \cdots\cdots & \sin(m-1)\frac{2\pi\ell}{m} & \cos(m-1)\frac{2\pi\ell}{m} \\ \vdots & \vdots & \cdots\cdots & \vdots & \vdots \\ \cos(m-1)\frac{2\pi\ell}{m} & \sin(m-1)\frac{2\pi\ell}{m} & \cdots\cdots & 1 & 0 \\ -\sin(m-1)\frac{2\pi\ell}{m} & \cos(m-1)\frac{2\pi\ell}{m} & \cdots\cdots & 0 & 1 \end{bmatrix}$$

(3.2.12)

with the elements corresponding to those defined in (3.2.10). Moreover in view of the orthogonality of the Fourier coefficients, for $\ell \neq \ell'$, the random vectors $\eta_{n\ell}$ and $\eta_{n\ell'}$ converge to independent normal vectors η_{ℓ} and $\eta_{\ell'}$. Now from (3.2.2) and (3.2.5),

$$\begin{aligned} \chi_n^2 &= \frac{m}{2\pi n}\sum_{i=1}^{m}\int_0^{2\pi}\left[n_i(\alpha)-\frac{n}{m}\right]^2 d\alpha \\ &= \frac{m}{2\pi n}\sum_{i=1}^{m}\sum_{\ell=1}^{\infty}(a_{i\ell}^2 + b_{i\ell}^2)\cdot\pi \\ &= \frac{m}{2}\sum_{\ell=1}^{\infty}(\eta_{n\ell}'\,\eta_{n\ell}) \\ &= \frac{m}{2}\sum_{\ell=1}^{\infty} Q_{n\ell} \end{aligned} \qquad (3.2.13)$$

where $Q_{n\ell} = \eta'_{n\ell}\,\eta_{n\ell}$ is a quadratic form in $\eta_{n\ell}$. However, since $\eta_{n\ell}$ converges in law to η_ℓ as $n \to \infty$, the a.d. of $Q_{n\ell}$ is that of $Q_\ell = \eta'_\ell\,\eta_\ell$ where η_ℓ is the random vector with means zero and covariance matrix Σ_ℓ given in (3.2.12). Because of this and the independence of the quadratic forms $Q_{n\ell}$ and $Q_{n\ell'}$ for $\ell \neq \ell'$, the a.d. of

$$S_{nN} = \sum_{\ell=1}^{N} Q_{n\ell} \tag{3.2.14}$$

is the same as that of

$$S_N = \sum_{\ell=1}^{N} Q_\ell \tag{3.2.15}$$

for any finite N. i.e.,

$$S_{nN} \xrightarrow{L} S_N \tag{3.2.16}$$

where $\xrightarrow{L}$ denotes convergence in law. If $S_{n\infty}$ and S_∞ stand for the corresponding infinite summands of the quadratic forms, we show below that

$$S_{n\infty} \xrightarrow{L} S_\infty \tag{3.2.17}$$

by arguments similar to those in Beran (1968).

In $F_X(\cdot)$ denotes the distribution function of the subscripted random variable X, then for any arbitrary continuity point x of $F_{S_\infty}(x)$, we have

$$(3.2.18) \qquad |F_{S_{n\infty}}(x) - F_{S_\infty}(x)| \leq |F_{S_{n\infty}}(x) - F_{S_{nN}}(x)| + |F_{S_{nN}}(x) - F_{S_N}(x)| + |F_{S_N}(x) - F_{S_\infty}(x)|$$

But since

$$E(Q_{n\ell}) = 4m \frac{\sin^2(\frac{\ell\pi}{m})}{\ell^2 \pi^2} = E(Q_\ell),$$

$$E|S_{n\infty} - S_{nN}| = \sum_{\ell=N+1}^{\infty} E(Q_{n\ell})$$

and

$$E|S_\infty - S_N| = \sum_{\ell=N+1}^{\infty} E(Q_\ell)$$

are the tails of convergent series. Therefore by Markov's inequality

$$(S_{n\infty} - S_{nN}) \xrightarrow{p} 0$$

$$(S_\infty - S_N) \xrightarrow{p} 0$$

uniformly in n as $N \to \infty$. Hence for any $\epsilon > 0$, there exists an N independent of n such that

(3.2.19) $$|F_{S_{n\infty}}(x) - F_{S_{nN}}(x)| < \epsilon/3$$

$$|F_{S_{\infty}}(x) - F_{S_N}(x)| < \epsilon/3 \quad .$$

Now for this choice of N, we can get a n_o such that for all $n > n_o$,

(3.2.20) $$|F_{S_{nN}}(x) - F_{S_N}(x)| < \epsilon/3$$

in view of (3.2.16). For such an n, (3.2.18), (3.2.19) and (3.2.20) imply

(3.2.21) $$|F_{S_{n\infty}}(x) - F_{S_{\infty}}(x)| < \epsilon.$$

Further since the distribution function of S_∞ is continuous, by Polya's theorem, this convergence in (3.2.21) is uniform in x. Thus the distribution of $\chi_n^2 = \frac{m}{2} \sum_{\ell=1}^{\infty} Q_{n\ell}$ converges not only weakly but uniformly to that of

$$\frac{m}{2} \sum_{\ell=1}^{\infty} Q_\ell \quad \text{as} \quad n \to \infty.$$

Now the distribution of $Q_\ell = \eta_\ell' \eta_\ell$ is not difficult to obtain but notice that $\Sigma_\ell = ((0))$ if ℓ is a multiple of m

and for $\ell \not\equiv 0 \pmod{m}$, it is only of rank 2 as can be seen from the fact that the third order principal minors of Σ_ℓ vanish. Therefore η_ℓ can be reduced to a two-dimensional random variable Y_ℓ by means of a transformation

$$(3.2.22) \qquad \underset{(2m \times 1)}{\eta_\ell} = \underset{(2m \times 2)}{B_\ell} \; \underset{(2 \times 1)}{Y_\ell}$$

where B_ℓ is such that $B_\ell B_\ell' = \Sigma_\ell$, B_ℓ is of rank 2 and Y_ℓ is distributed as $N(0, I_2)$ (See eg. Rao (1965), p 440). Because of the fact that Y_ℓ is distributed as $N(0, I_2)$, the characteristic function (c.f.) of

$$Q_\ell = \eta_\ell' \eta_\ell = Y_\ell' (B_\ell' B_\ell) Y_\ell$$

is given by

$$(3.2.23) \qquad \phi_\ell(t) = \det.^{-1/2} |I_2 - 2it\, B_\ell' B_\ell| .$$

In our case, Σ_ℓ given in (3.2.12), can be written as $B_\ell B_\ell'$ where

$$(3.2.24) \qquad B_\ell' = \frac{\sqrt{2} \sin \frac{\ell\pi}{m}}{\ell \pi} \times$$

$$\begin{bmatrix} 1 & 0 & \cos\frac{2\pi\ell}{m} & -\sin\frac{2\pi\ell}{m} & \cdots\cdots & \cos(m-1)\frac{2\pi\ell}{m} & -\sin(m-1)\frac{2\pi\ell}{m} \\ 0 & 1 & \sin\frac{2\pi\ell}{m} & \cos\frac{2\pi\ell}{m} & \cdots\cdots & \sin(m-1)\frac{2\pi\ell}{m} & \cos(m-1)\frac{2\pi\ell}{m} \end{bmatrix} .$$

For this B_ℓ, since

$$(3.2.25) \qquad B_\ell' B_\ell = \frac{2m \sin^2 \frac{\ell\pi}{m}}{\ell^2 \pi^2} I_2,$$

the c.f. of Q_ℓ, from (3.2.23), is

$$(3.2.26) \qquad \phi_\ell(t) = \text{det.}^{-1/2} \left| I_2 - 4it\, m \frac{\sin^2 \frac{\ell\pi}{m}}{\ell^2 \pi^2} I_2 \right|.$$

$$= \left(1 - \frac{4m \sin^2 \frac{\ell\pi}{m}}{\ell^2 \pi^2} it\right)^{-1}.$$

Now from the independence of η_ℓ and $\eta_{\ell'}$ for $\ell \neq \ell'$, the asymptotic c.f. of χ_n^2 can be written down using (3.2.13) and (3.2.26) as

$$(3.2.27) \qquad \phi(t) = \prod_{\ell=1}^{\infty} \left(1 - \frac{2m^2 \sin^2 \frac{\ell\pi}{m}}{\ell^2 \pi^2} it\right)^{-1}$$

$$= \prod_{\ell=1}^{\infty} (1 - it\, \lambda_\ell)^{-1}$$

with

$$(3.2.28) \qquad \lambda_\ell = 2 \sin^2 \frac{\ell\pi}{m} / \left(\frac{\ell\pi}{m}\right)^2.$$

If ℓ is a multiple of m, the corresponding λ_ℓ is zero so that it does not contribute anything to the c.f. of χ^2. We mentioned this earlier my saying that for such an ℓ, $\Sigma_\ell = ((0))$

so that the contribution of the corresponding quadratic form Q_ℓ is zero. From (3.2.27), the asymptotic distribution of the χ^2 statistic can be formally written down. If $f(x)$ denotes the density function we have by the inversion formula

$$(3.2.29) \quad f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \left[\prod_\ell (1-it\,\lambda_\ell)^{-1}\right] dt.$$

Since the λ_ℓ are positive and distinct there are only simple poles for the integrand and they occur in the lower half of the complex plane at $t_\ell = \frac{1}{i\lambda_\ell}$. The integral (3.2.29) can then be evaluated by closing a contour in the lower half plane and $f(x)$ is given by i times the sum of the residues at the poles. Thus

$$(3.2.30) \quad f(x) = \sum_\ell c_\ell \exp\left[-x/\lambda_\ell\right] \quad \text{for} \quad x \geq 0$$

where

$$c_\ell = \left[\lambda_\ell \prod_{\ell' \neq \ell} (1 - \lambda_{\ell'}/\lambda_\ell)\right]^{-1}$$

and zero otherwise. From (3.2.27) one can write down the μ^{th} cumulant, K_μ of the distribution and we have

$$K_\mu = (\mu-1)! \sum_{j=1}^{\infty} \lambda_j^\mu = 2^\mu(\mu-1)! \frac{m^{2\mu}}{\pi^{2\mu}} \sum_{j=1}^{\infty} \sin^{2\mu}\left(\frac{j\pi}{m}\right)/j^{2\mu}.$$

For getting the percentage points for this distribution, some numerical approximations may have to be found or use may be made of the first four cumulants to find a suitable Pearson curve approximation.

The statistic χ^2_n given in (3.2.2) in the integral form can now be expressed as a finite sum and given the following computational form

$$\chi^2_n = \frac{m}{2\pi n} \int_0^{2\pi} \sum_{i=1}^{m} [\eta_i(\alpha) - \frac{n}{m}]^2 \, d\alpha$$

$$= \frac{m}{2\pi n} \sum_i \int_0^{2\pi} \{[\sum_j \chi^i_j(\alpha)]^2 - 2\frac{n}{m}[\sum_j \chi^i_j(\alpha)] + \frac{n^2}{m^2}\} \, d\alpha$$

(3.2.31)

$$= \frac{m}{2\pi n} \sum_i \int_0^{2\pi} \{[\sum_j \chi^i_j(\alpha)][1 - \frac{2n}{m}] +$$

$$+ [\sum_{j \neq j'} \chi^i_j(\alpha) \chi^i_{j'}(\alpha)] + \frac{n^2}{m^2}\} \, d\alpha$$

since $[\chi^i_j(\alpha)]^2 \equiv \chi^i_j(\alpha)$. Now

(3.2.32)
$$\chi^i_j(\alpha) = \begin{cases} 1 & \text{if } \alpha_j - \frac{2\pi i}{m} \leq \alpha < \alpha_j - \frac{2\pi(i-1)}{m} \\ 0 & \text{otherwise} \end{cases}$$

and similarly

(3.2.33)
$$\chi^i_j(\alpha)\, \chi^i_{j'}(\alpha) = \begin{cases} 1 & \text{if either } \alpha_{j'} - \frac{2\pi i}{m} \leq \alpha < \alpha_j - \frac{2\pi(i-1)}{m} \\ & \text{or} \qquad \alpha_j - \frac{2\pi i}{m} \leq \alpha < \alpha_{j'} - \frac{2\pi(i-1)}{m} \\ 0 & \text{otherwise} \end{cases}$$

Using (3.2.32) and (3.2.33) in (3.2.31),

$$(3.2.34)\qquad \chi^2_n = \frac{m}{2\pi n}\sum_i \Big\{ \Big[1-\frac{2n}{m}\Big]\Big[\sum_j \frac{2\pi}{m}\Big] + \sum_j \sum_{j'} \Big(\frac{2\pi}{m} - D_{jj'}\Big) + \frac{2\pi n^2}{m^2} \Big\}$$
$$\Big\{ (j,j') : j \neq j',\ D_{jj'} \leq \frac{2\pi}{m} \Big\}$$

where $D_{jj'}$ denotes the 'circular' distance (i.e., smaller of the two distances) between the observations α_j and $\alpha_{j'}$ on the circumference. Simplifying (3.2.34) further

$$\chi^2_n = \frac{m^2}{2\pi n}\Big\{ (m-2n)\,\frac{2\pi n}{m^2} + \frac{2\pi n^2}{m^2} + \sum_{j\neq j'}\sum \Big(\frac{2\pi}{m} - D_{jj'}\Big) \Big\}$$
$$\Big\{(j,j') : D_{jj'} \leq \frac{2\pi}{m}\Big\}$$

$$= m - n + \frac{m^2}{2\pi n} \sum_{j\neq j'}\sum \Big(\frac{2\pi}{m} - D_{jj'}\Big)$$

$$(3.2.35)\qquad \Big\{ (j,j') :\ D_{jj'} \leq \frac{2\pi}{m} \Big\}$$

Thus the invariant version (3.2.2) of the usual χ^2 statistic, depends on the sample arc lengths between the observations. These arc lengths play a crucial role in the case of the circle as any invariant test statistic has to be a function of these values. We will discuss about these arc lengths in much greater detail in the other sections of this Chapter as well as in Chapter IV. The statistic χ^2_n, in this form

(3.2.35), looks like a 'U-statistic' (See eg. Fraser (1957)) which it is not in the strictest sense. The usual central limit theorem for the U-statistics thus fails in our case. Finally it may be remarked that the Ajne's statistic A_n discussed in Ajne (1966) and Watson (1967) is a special case of our χ^2 statistic when the number of class intervals into which the circumference is divided is only 2.

Now we look at the problem of testing for uniformity on the basis of a grouped data from another direction. If the alternatives of interest can be specified, at least approximately as can be done in most cases, the χ^2 given in (3.2.1) can be improved by concentrating on the specified class of alternative distributions. Suppose it is desired to test for uniformity, as against the class of unimodal symmetric densities given by

$$(3.2.36) \qquad g(\alpha \mid \vartheta, \gamma) = \frac{1}{2\pi} + \vartheta \cos(\alpha - \gamma), \quad 0 \leq \alpha < 2\pi$$

where $0 \leq \gamma < 2\pi$ and $\vartheta \geq 0$ denote the location and concentration parameters respectively. It may be observed that circular normal densities close to the hypothesis of uniformity (i.e., with a low value for the concentration parameter) can be put in this form. Under these alternatives (3.2.36), the i^{th} cell has probability

$$\pi_i(\alpha) = \int_{\alpha+(i-1)2\pi/m}^{\alpha+i2\pi/m} [1/2\pi + \vartheta \cos(\beta - \gamma)]\, d\beta$$

$$= \frac{1}{m} + \vartheta \cos\gamma \left\{ 2 \sin\frac{\pi}{m} \cos\left(\alpha + \frac{(2i-1)\pi}{m}\right)\right\}$$

(3.2.37)

$$+ \vartheta \sin\gamma \left\{ 2 \sin\frac{\pi}{m} \sin\left(\alpha + \frac{(2i-1)\pi}{m}\right)\right\}$$

$$= \frac{1}{m} \left\{ 1 + \xi x_i + \eta y_i \right\}$$

where

$$\xi = \vartheta \cos\gamma \quad , \qquad \eta = \vartheta \sin\gamma$$

$$x_i = 2m \sin\frac{\pi}{m} \cos\left(\alpha + \frac{(2i-1)\pi}{m}\right),$$

(3.2.38)

$$y_i = 2m \sin\frac{\pi}{m} \sin\left(\alpha + \frac{(2i-1)\pi}{m}\right).$$

Now in order that the required asymptotic distribution holds good for the statistic given in (3.2.44), these π_i's should be estimated by any method of estimation that is 'efficient' in the sense of Rao i.e., satisfying the assumption (3) of Rao (1961). The validity of this assumption can be verified easily for the estimates $\hat{\xi}$ and $\hat{\eta}$ obtained by minimising the quantity

$$L = \sum_1^m (p_i - \pi_i)^2/\pi_i^o \qquad (3.2.39)$$

$$= \frac{1}{m} \sum_1^m (mp_i - 1 - \xi x_i - \eta y_i)^2.$$

On using the trigonometric relations

$$\sum_{1}^{m} \cos\frac{(2i-1)\pi}{m} = \sum_{1}^{m} \sin\frac{(2i-1)\pi}{m} = \sum_{1}^{m} \cos\frac{(2i-1)\pi}{m} \sin\frac{(2i-1)\pi}{m}$$

$$= 0$$

(3.2.40)

$$\sum_{1}^{m} \cos^2 \frac{(2i-1)\pi}{m} = \sum_{1}^{m} \sin^2 \frac{(2i-1)\pi}{m} = m/2$$

the estimating equations simplify and we get

$$\hat{\xi} = m \sum_{1}^{m} p_i x_i \Big/ \sum_{1}^{m} x_i^2$$

$$= \bar{c} \Big/ m \sin\frac{\pi}{m}$$

(3.2.41)

$$\hat{\eta} = m \sum_{1}^{m} p_i y_i \Big/ \sum_{1}^{m} y_i^2$$

$$= \bar{s}/m \sin\frac{\pi}{m}$$

where

(3.2.42) $$n\bar{c} = \sum_{1}^{m} n_i \cos\left(\alpha + \frac{(2i-1)\pi}{m}\right) \text{ and } n\bar{s} = \sum_{1}^{m} n_i \sin\left(\alpha + \frac{(2i-1)\pi}{m}\right)$$

are nothing but the components of the vector resultant based on the grouped data. Thus from (3.2.37) and (3.2.41),

(3.2.43) $$\hat{\pi}_i = 1/m \left(1 + \hat{\xi} x_i + \hat{\eta} y_i\right)$$

$$= 1/m + (2/m)\left\{\bar{c} \cos\left(\alpha + \frac{(2i-1)\pi}{m}\right) + \bar{s} \sin\left(\alpha + \frac{(2i-1)\pi}{m}\right)\right\}$$

Let $E_{oi} = n \pi_i^o$ and $E_{1i} = n \hat{\pi}_i$, denote respectively the estimated frequencies in the i^{th} cell, under the restrictions imposed by the hypothesis, and when there are no restrictions on the parameters. Then Rao (1961) suggests the statistic

$$(3.2.44) \qquad T = \sum_{1}^{m} (E_{1i} - E_{oi})^2/E_{oi}$$

$$= n \sum_{1}^{m} (\hat{\pi}_i - \pi_i^o)^2/\pi_i^o$$

to test the hypothesis that the data follows a uniform distribution given that the admissible set of probabilities $\pi_i = \pi_i(\theta, \gamma)$ have the representation given in (3.2.37). Using the relations (3.2.40), the statistic (3.2.44) reduces to

$$T = nm \sum_{1}^{m} \left\{ (2/m)\left[\bar{c} \cos\left(\alpha + \frac{(2i-1)\pi}{m}\right) + \bar{s} \sin\left(\alpha + \frac{(2i-1)\pi}{m}\right)\right] \right\}^2$$

$$= 2n(\bar{c}^2 + \bar{s}^2) \qquad (3.2.45)$$

$$= 2R^2/n$$

where R^2 is the squared length of the resultant based on the given grouped data. The value of this R^2, computed on the basis of the grouped data obviously depends on the particular grouping adopted or in other words on α. But as is well known, the length of the resultant based on the ungrouped (or raw) data is independent of the choice of the zero direction so that the statistic (3.2.45), remains invariant if the effect of grouping is ignored. However, since the grouping correction needed for R^2 turns out to be quite negligible even when we have only about 10 to 15 class intervals (see eg. Batschelet (1965)), the statistic is almost as good as an invariant one if m is not too small.

Now, under the density (3.2.36), the cell probabilities π_i have a parametric representation in terms of two independent parameters ρ and γ or equivalently in terms of $\xi = \rho \cos \gamma$ and $\eta = \rho \sin \gamma$. On the other hand, the hypothesis of uniformity is equivalent to the simple hypothesis that the parameter point is the origin i.e., $\xi = 0$, $\eta = 0$. Hence the statistic T, given in (3.2.45), has asymptotically a χ^2 distribution with 2 degrees of freedom (c.f. Rao (1961)). It is interesting that the special type of χ^2 test (3.2.44) reduces to an analogue of the Rayleigh's test for grouped data on the circle.

3.3 Sample arc lengths and their distribution under the hypothesis of uniformity

In the rest of this chapter, we suggest tests based on sample arc lengths for testing uniformity and discuss some of them. As we mentioned in the beginning, we take a fixed sense of rotation (say, anticlockwise) as positive throughout. If, however, one takes a fixed point as the origin (or zero direction), then it can be shown (see eg. Feller (1966) p. 22) that the length of the sample arc containing this fixed origin would be equivalent distributionally to the sum of the lengths of two other simple arcs, whereas on symmetry considerations, we should expect all the arc lengths to be identically distributed. To

steer clear of such apparent paradoxes, which are peculiar to the spacings on the circle, we set forth the main distributional results for the arc lengths on the circle.

Let $\alpha_1,\ldots,\alpha_n$ be the n angular observations from a density $f(\alpha)$, $0 \leq \alpha < 2\pi$, measured with respect to some arbitrary origin. We will arrive at the distribution of the arc lengths via the distribution of maximal invariant in our case. We consider the problem of testing the hypothesis H_0: $f(\alpha)$ is uniform density, against the alternative H_1: $f(\alpha) \in \mathcal{A}$, where $\mathcal{A}$ is a family of circular densities which is invariant under the group of all rotations i.e. a density corresponding to a random variable continues to belong to $\mathcal{A}$ even if the origin of measurement of the random variable is changed. The choice of such a family of alternatives is natural for the circle since it makes the testing problem invariant under rotations of the circle within itself and has been suggested by Ajne (1966). From the theory of invariant tests, an invariant test must depend on the observations only through the maximal invariant, which, in this case, is the set of differences

$$(3.3.1) \qquad \theta_2 = \alpha_2 - \alpha_1, \quad \theta_3 = \alpha_3 - \alpha_1, \ldots, \theta_n = \alpha_n - \alpha_1$$

where the sums or differences are to be interpreted properly i.e. modulo 2π. This amounts to taking α_1 as the new origin

and renaming the other observations as $\theta_2, \ldots, \theta_n$. Clearly the joint density of $\theta_2, \ldots, \theta_n$ is

$$(3.3.2) \qquad \int_0^{2\pi} f(\alpha_1) \prod_{k=2}^{n} f(\theta_k + \alpha_1)\, d\alpha_1 .$$

Putting $\theta_1 = 0$, we see that $\theta_1, \theta_2, \ldots, \theta_n$ are identically distributed with a density $h(\theta) = f(\theta + \alpha_1)$, rotationally equivalent to the density $f(\alpha)$, and hence still a member of the family $\mathcal{A}$. Thus the problem of testing $(\alpha_1, \ldots, \alpha_n)$ have density $f(\alpha)$ is equivalent to testing $(\theta_1, \ldots, \theta_n)$ have a density $h(\theta)$. The property of rotational invariance of a distributional or testing problem is preserved if one proceeds through θ_i's instead of the original observations α_i's.

The sample arc lengths, $\{ T_i \}$ are now defined as follows. If

$$(3.3.3) \qquad 0 = \theta_1' \leq \theta_2' \leq \ldots \leq \theta_n' \leq 2\pi$$

denote the ordered values of $(\theta_1, \ldots, \theta_n)$, then the sample arc lengths are

$$(3.3.4) \qquad T_1 = \theta_2', \; T_2 = \theta_3' - \theta_2', \ldots, \; T_{n-1} = \theta_n' - \theta_{n-1}'$$

and $\qquad T_n = 2\pi - \theta_n'$.

These are, in fact, the lengths of the

segments into which the circumference is broken up by the sample observations, $(\alpha_1,\ldots,\alpha_n)$. These arc lengths, defined in (3.3.4), depend on the observations through the maximal invariant $(\theta_2,\ldots,\theta_n)$ and hence tests based on the arc lengths are rotation invariant, a fact, which is apparent even otherwise. If α_1 is taken as the new origin i.e., θ_1 is taken to be identically zero, the likelihood of obtaining a sample in an infinitesimal volume element $d\theta_2' \ldots d\theta_n'$ is given by

$$(3.3.5)\quad (n-1)!\prod_{j=2}^{n} h(\theta_j')d\theta_j', \text{ if } 0 \le \theta_2' \le \ldots \le \theta_n' \le 2\pi.$$

When in particular, the observations are from the uniform density, the joint density of $(\theta_2',\ldots,\theta_n')$ is

$$(3.3.6)\qquad (n-1)!/(2\pi)^{n-1} \cdot d\theta_2' \ldots d\theta_n'$$

which is also the unconditional density given by (3.3.2), in this case. The density of $(T_1,\ldots,T_{n-1})$, under the hypothesis of uniformity, is therefore obtained from (3.3.6) as

$$(3.3.7)\qquad g(T_1,\ldots,T_{n-1}) = \begin{cases} (n-1)!/(2\pi)^{n-1} \cdot dT_1 \ldots dT_{n-1}, \\ \qquad \text{if } T_i \ge 0,\ \sum_1^{n-1} T_i \le 2\pi \\ 0 \quad \text{otherwise} \end{cases}$$

since the Jacobian of transformation is unity. The frequency element of all the arc lengths i.e., $(T_1,\ldots,T_n)$ is still given by $[(n-1)!/(2\pi)^{n-1}]$ but now the distribution is degenerate because of the restriction $\sum_1^n T_i = 2\pi$.

Clearly, the marginal distribution of any subset of k arcs say $(T_{i_1},\ldots, T_{i_k})$ would be the same whatever the indices $(i_1,\ldots,i_k)$ are. In other words, these T_i's form a set of interchangeable random variables under the hypothesis of uniformity. From (3.3.7) the density of any single arc, say T_1, is

$$(3.3.8) \qquad g_1(t_1) = (n-1)/(2\pi)^{n-1}\cdot(2\pi-t_1)^{n-2}$$

$$\text{for } 0 \le t_1 \le 2\pi .$$

Similarly the density of any subset $(T_{i_1},\ldots,T_{i_k})$, $(k \le n-1)$ of gaps is

$$(3.3.9) \qquad g_k(t_{i_1},\ldots,t_{i_k}) =$$

$$= [(n-1)!/(n-k-1)!(2\pi)^{n-1}]\ (2\pi - \sum_{j=1}^{k} t_{i_j})^{n-k-1}$$

$$\text{for } 0 \le t_{i_1} \le t_{i_1} + t_{i_2} \le \cdots \le \sum_1^k t_{i_j} \le 2\pi.$$

3.4 Tests based symmetrically on the arc lengths - U_n statistic and its exact and asymptotic distributions

On the line segment, tests based on sample spacings are used in a number of situations as for example in goodness of fit problems and in testing departures from exponentiality (See e.g. the literature cited earlier). The main idea of this section is to introduce similar classes of tests, based on the circular spacings or the arc lengths defined in (3.3.4), for testing for uniformity or goodness of fit on the circle. The essential difference, with regard to spacings, between the line segment and the circle is that whereas n observations on the former give (n+1) spacings, we get only n arcs on the latter. However, as can be seen from the results of Section 3.3, the distribution under uniformity of the n arcs made by n sample points on the circle, is the same as that of n spacings made on the line segment $[0, 2\pi]$ by (n-1) sample points from a uniform distribution on $[0, 2\pi]$. This analogy tells us that we could apply any of the spacings tests on the line, to the circular case with only minor modifications. A test statistic based symmetrically on the arc lengths T_i can be written in the form

$$A_n = 1/n \sum_{i=1}^{n} m(nT_i) \qquad (3.4.1)$$

where $m(\cdot)$ is any reasonable function of the arguments nT_i. Such spacings tests have been made use of on the line (ref. e.g. Pyke (1965) and the references contained therein). For instance, $m(x) = x^r$, $r > 0$ gives a class of statistics due to Greenwood (1946) and Kimball (1950). The cases $m(x) = |x-1|$ and $m(x) = \log x$ are studied by Sherman (1950) and Darling (1953) respectively. We consider these different classes of spacings tests in more detail and compare them from the point of view of their efficiencies in the next chapter.

We now discuss a particular test statistic based on the arc lengths, which has an attractive interpretation on the circle. First we observe that the expected length of T_i, under uniformity, is $2\pi/n$. Given n sample points on the circumference, corresponding to the i^{th} sample point, we cut an arc of fixed length, $2\pi/n$, starting with the sample point and in the positive (anticlockwise) direction. When n such arcs, each of fixed length $2\pi/n$, are placed corresponding to the n sample points, a complete covering of the circumference occurs only if the n sample observations happen to be equi-spaced on the circumference. Stevens (1939) and Rao (1942) consider the probability that the circumference is completely covered when n arcs of arbitrary lengths $a_1, \ldots, a_n$ are randomly placed on the circumference of the circle with given radius. In our case

the probability of such a complete coverage is zero since, as we have noted, this can happen only when the sample points are equally spaced on the circumference. On the other hand when the observation are not equi-spaced, some of the arcs of fixed length would overlap, and to that extent a portion of the circumference would remain uncovered by these arcs. We shall denote this uncovered portion of the circumference by U_n. U_n takes large values when there is a clustering of the sample observations in one or more places, the maximum value, viz. $(1 - \frac{1}{n})2\pi$, occurring when all the sample points happen to coincide. Thus

$$(3.4.2) \qquad 0 \leq U_n \leq (1 - \frac{1}{n})\ 2\pi,$$

large values of U_n indicating clustering of the observations or departures from uniformity. The quantity

$$(3.4.3) \qquad S_i = \max.[T_i - 2\pi/n,\ 0]$$

gives the uncovered portion as contributed by the i^{th} sample arc T_i and then the statistic

$$(3.4.4) \qquad U_n = \sum_{i=1}^{n} S_i$$

$$= \sum_{i=1}^{n} \max.[T_i - 2\pi/n,\ 0].$$

S_i's are again interchangeable random variables and have the same range as U_n. Using the results in Section 3.3 and the interchangeability of S_i's, one can show that

$$E(U_n) = 2\pi\left(1 - \frac{1}{n}\right)^n \tag{3.4.5}$$

$$\text{Var}\,(U_n) = 4\pi^2\left[2/(n+1)\cdot\left(1 - \frac{1}{n}\right)^{n+1} + \frac{(n-1)}{(n+1)}\cdot\left(1 - \frac{2}{n}\right)^{n+1} - \left(1 - \frac{1}{n}\right)^{2n}\right].$$

The exact distribution of U_n can also be derived rather easily, following an elegant approach based on Laplace transforms due to Darling (1953). But a moment's reflection would show that the U_n defined in (3.4.4) is the same as the quantity

$$\frac{1}{2}\sum_{i=1}^{n}\left|T_i - \frac{2\pi}{n}\right| \tag{3.4.6}$$

studied for the line by Sherman (1950). In view of the remarks made earlier, regarding the sample spacings on the line and the sample arc lengths on the circle, the distribution of U_n can immediately be written down from that of Sherman (1950) or Darling (1953). In fact, the density function of U_n is given by

$$(3.4.7) \qquad \phi_n(u) = (n-1)! \sum_{j=1}^{n-1} \binom{n}{j} (u/2\pi)^{n-j-1} \frac{f_j(nu)}{(n-j-1)!\,n^{j-1}}$$

$$\text{for } 0 \leq u \leq 2\pi\left(1 - \frac{1}{n}\right)$$

where $f_j(x)$ is the density function of the sum of j independent uniform random variables on $[0, 2\pi]$ and has the expression

$$(3.4.8) \qquad f_j(x) = \frac{1}{2\pi \cdot (j-1)!} \sum_{k=0}^{\infty} (-1)^k \binom{j}{k} \langle \frac{x}{2\pi} - k \rangle^{j-1}.$$

with the notation $\langle x \rangle = x$ if $x > 0$ and $= 0$ if $x \leq 0$.

The asymptotic normality of the statistic U_n has been shown by the cumbersome method of moments by Sherman (1950) and by the method of steepest descent by Darling (1953). A more general and elegant method, which also establishes the asymptotic normality of the statistic under suitably chosen alternatives, is given by us in the chapter that follows. For the present, we simply state the asymptotic normality of U_n in the following theorem, whose proof follows from the more general results of Chapter IV and hence is omitted here.

Theorem 3.4.1: The statistic $\sqrt{n}\,(U_n - 2\pi/e)$, where U_n is as defined in (3.4.4), has an asymptotic normal distribution with mean zero and variance $4\pi^2(2e^{-1} - 5e^{-2})$.

The property of consistency (power approaching unity as the sample size increases indefinitely) of the test procedure based on U_n, is also established in Serman (1950). But we give here a simpler alternate proof of the same. For doing this, it is sufficient to show that U_n converges in probability to a value different from $(2\pi/e)$ under the alternatives since the variances both under the hypothesis and under the alternative tend to zero. Thus the consistency of the test sequence U_n, for a large class of alternatives with continuous densities $g(\alpha)$ differing from the uniform density on $[0,2\pi]$, is established by the following

Theorem 3.4.2: Under the alternative distribution function $G(\alpha)$ on $[0,2\pi]$ with continuous density $g(\alpha)$, the spacings statistic U_n defined in (3.4.4), converges in probability to

$$\int_0^{2\pi} \exp\,[-2\pi\, g(\alpha)]\,d\alpha.$$

Proof: Let $\alpha_1,\ldots,\alpha_n$ be independently distributed with distribution function (d.f.) $G(\alpha)$ on $[0, 2\pi]$. If $0 \le \alpha_1' \le \ldots \le \alpha_n' \le 2\pi$ denote the ordered values, then

$$D_i = \alpha_i' - \alpha_{i-1}', \quad i = 1,\ldots, n$$

from the d.f. $G(\alpha)$. Using the fact that if A is a random variable (r.v.) with d.f. $G(\alpha)$, then the r.v. $U = G(A)$ has the uniform distribution on $[0, 1]$, we now relate these spacings D_i to spacings from a uniform distribution. Let $U_1, \ldots, U_n$ be independent observations from the uniform distribution on a circle of unit circumference and let $0 \leq U_1' \leq \ldots \leq U_n' \leq 1$ be the values arranged in increasing order. Then

$$T_i = U_i' - U_{i-1}'$$

with $U_0' = (U_n' - 1)$ give the arc lengths from the uniform distribution on the circle of unit circumference. For any two random variables X and Y, writing $X \sim Y$ to mean that X and Y are distributionally equivalent (that is the distributions of X and Y are identical), we have

$$\begin{aligned} \{ D_i = (\alpha_i' - \alpha_{i-1}'),\ i = 1, \ldots, n \} \\ \sim \{ G^{-1}(U_i') - G^{-1}(U_{i-1}'),\ i = 1, \ldots, n \} \\ = \{ (U_i' - U_{i-1}')/k(\tilde{U}_i),\quad i = 1, \ldots, n \} \\ = \{ T_i/k(\tilde{U}_i),\quad i = 1, \ldots, n \}. \end{aligned} \tag{3.4.9}$$

where $U_{i-1}' \leq \tilde{U}_i \leq U_i'$ and $k(p) = g(G^{-1}(p))$

This relation (3.4.9) connects the circular spacings D_i from

the density $g(\alpha)$ and the spacings T_i from the uniform density. If $W_1, \ldots, W_n$ are n independently and identically distributed exponential random variables with density e^{-w}, $w \geq 0$, and $W_n^* = W_1 + \ldots + W_n$, then it is well known that

$$(3.4.10) \qquad \{ T_i, \ i = 1, \ldots, n \} \sim \{ W_i / W_n^*, \ i = 1, \ldots, n \}.$$

Thus from (3.4.9) and (3.4.10)

$$(3.4.11) \qquad \{ D_i, \ i = 1, \ldots, n \} \sim \{ W_i / W_n^* \ k(\tilde{U}_i), \ i = 1, \ldots, n \}.$$

Therefore the empirical distribution function of $\{ nD_i, \ i = 1, \ldots, n \}$, say

$$(3.4.12) \qquad H_n(a) = n^{-1} [\text{ number of } nD_i \leq a]$$

$$\sim n^{-1} [\text{ number of } W_i / \bar{W}_n \ k(\tilde{U}_i) \leq a]$$

where $\bar{W}_n = W_n^* / n$. Now in view of the facts that for any $\delta > 0$,

$$n^{\frac{1}{2} - \delta} \sup_i |\tilde{U}_i - i/n| \xrightarrow{p} 0$$

$$n^{\frac{1}{2} - \delta} |\bar{W}_n - 1| \xrightarrow{p} 0$$

large enough n, stochastically equivalent to

$$(3.4.13) \qquad H_n^*(a) = n^{-1}[\text{number of } W_i \leq a\, k(i/n)].$$

But

$$EH_n^*(a) = 1/n \sum_{i=1}^{n} (1 - e^{-ak(i/n)})$$

which can be written, as $n \to \infty$, in the integral form

$$(3.4.14) \qquad \int_0^1 (1 - e^{-ak(p)})\,dp$$

$$= 1 - \int_0^{2\pi} e^{-ag(x)}\, dG(x).$$

And

$$\text{Var}\,(H_n^*(a)) = \frac{1}{n^2} \sum_{i=1}^{n} e^{-ak(i/n)} (1 - e^{-ak(i/n)})$$

which tends to zero as $n \to \infty$. Thus for each fixed a, $H_n^*(a)$ and hence $H_n(a)$, converges stochastically to

$$\left\{1 - \int_0^{2\pi} \exp\,[-ag(x)]dG(x)\right\}$$

and hence from Polya's theorem, the convergence in supremum norm also takes place. Now

$$U_n = \sum_{\{i: D_i \le 2\pi/n\}} (2\pi/n - D_i)$$

$$= \int_0^{2\pi} (2\pi - a)\, dH_n(a)$$

$$= \int_0^{2\pi} H_n(a)\, da$$

(3.4.15) $$\xrightarrow{p} \int_0^{2\pi} [1 - \int_0^{2\pi} e^{-ag(\alpha)} dG(\alpha)]\, da$$

$$= \int_0^{2\pi} \exp[-2\pi\, g(\alpha)]\, d\alpha .$$

This probability limit of U_n differs from $(2\pi/e)$ if $g(\alpha)$ is continuous and differs from the uniform density on a set of positive Lebesgue measure, thus establishing the consistency of the U_n-test against such alternatives.

3.5 Table of critical values for using U_n and an illustrative example

In this Section, we give a table of percentage points for the test statistic U_n, for sample sizes $n = 2(1)20$ and for three levels of significance, $\alpha = .01$, 0.05 and 0.10. This table is obtained by a modification of Sherman's table (1957). We illustrate, by means of a numerical example, the simplicity in using the statistic U_n for testing uniformity of a circular distribution.

Table 3.1

Table of critical points $U_0(\alpha,n)$ (in degrees) for the statistic U_n.

n \ α	0.01	0.05	0.10
2	178.20	171.00	162.00
3	219.24	193.68	174.24
4	221.04	186.48	171.72
5	212.04	183.60	168.84
6	206.04	180.72	166.32
7	202.68	177.84	164.88
8	198.36	175.68	163.44
9	195.12	173.52	162.36
10	192.24	172.08	161.28
11	189.72	170.28	160.20
12	187.56	169.20	159.48
13	185.76	167.76	158.40
14	183.96	166.68	157.68
15	182.16	165.60	156.96
16	180.72	164.88	156.60
17	179.64	164.16	155.88
18	178.20	163.08	155.16
19	177.12	162.36	154.80
20	176.04	161.64	154.44

If for a given sample size n and level α, the calculated value of the uncovered length U_n exceeds the tabulated value $U_o(\alpha, n)$, we reject the hypothesis of uniformity. The critical points have been given in terms of degrees for ready applicability.

Example: We consider here example 20.1 given in Batschelet (1965).

'In an experiment on homing orientation in pigeons, 10 birds were released singly at 25 Km west of their loft. Field glass observations yielded the vanishing point of each departing bird measured to the nearest 5° interval in true bearings. These ten vanishing points are 20°, 35°, 350°, 120°, 85°, 345°, 80°, 320°, 280° and 85°'.

It is required to know whether the birds have a preferred orientation of flight. The arc lengths $\{T_i\}$ made by these observations on the circle are easily seen to be 15°, 45°, 5°, 0°, 35°, 160°, 40°, 25°, 5° and 30° and the fixed arcs are of length $360/10 = 36^\circ$ in this case. Therefore

$$U_{10} = \sum_{i=1}^{10} \max [T_i - 36, 0]$$

$$= (1/2) \sum_{i=1}^{10} |T_i - 36|$$

$$= 137^\circ.$$

This value of 137°, for $n = 10$, is not significant even at the 10 percent level of significance as the critical point in this case is only 161.28 degrees. Therefore we conclude that the observations could have come from a uniform distribution. The same conclusion is reached using the Rayleigh's test (See e.g. Batschelet (1965)). But the test based on U_n is not only very much simpler computationally and conceptually but is also valid against a wider class of alternatives as we have remarked in the beginning.

3.6 Other tests based on arc lengths - circular range

As we have mentioned in Section 3.4, any of the spacings tests for randomness on the line can be adopted to the circular case with ease, in view of the distributional results of Section 3.3. Another simple statistic of interest in this connection, is the 'circular range', R, which is the length of the smallest arc that encompasses all the sample observations. This is the correct analogue of the linear range to the circular situation and may be used in testing uniformity. Small values of R should be considered critical as that indicates clustering of the observations. The distribution of R, under uniformity, can be obtained directly from its definition and the results of Section 3.3.

Since, for an r such that $0 \leq r < 2\pi$, R can be less than or equal to r if and only if at least one of the n arcs $\{T_i, i = 1,\ldots, n\}$ exceeds $(2\pi - r)$ in length, the distribution function of R, say $F(r)$, can be written as

$$F(r) = P(R \leq r)$$

$$= P\left(\bigcup_{i=1}^{n} E_i\right)$$

where E_i stands for the event that the i^{th} spacing T_i exceeds $(2\pi - r)$. Because of the identical distributions for T_i, this gives

$$F(r) = P\left(\bigcup_{1}^{n} E_i\right)$$

$$= nP(E_{i_1}) - \binom{n}{2}P(E_{i_1} \cap E_{i_2}) + \cdots$$

(3.6.1)

$$+ (-1)^{k-1}\binom{n}{k}P(E_{i_1} \cap \cdots \cap E_{i_k})$$

the series terminating with a k, where k is the integral part of $(2\pi/2\pi-r)$. This is clear since, if the integral part of $(2\pi/2\pi-r)$ is k, then at most k arcs can exceed $(2\pi-r)$. Now the probabilities $P(E_{i_1})$, $P(E_{i_1} \cap E_{i_2})$, etc., in (3.6.1) can be computed easily from the results of Section 3.3. It can be checked, for instance, that

3.6.2) $$P(E_{i_1} \cap \cdots \cap E_{i_m}) = [mr - 2\pi(m-1)]^{n-1}/(2\pi)^{n-1}$$

if $m \leq k$.

From (3.6.1) and (3.6.2), the distribution function of R, comes out to be

$$F(r) = \sum_{m=1}^{k} (-1)^{m-1} \binom{n}{m} \left[\frac{mr}{2\pi} - (m-1) \right]^{n-1}$$

$$= \sum_{m=1}^{\infty} (-1)^{m-1} \binom{n}{m} \left\langle \frac{mr}{2\pi} - (m-1) \right\rangle^{n-1} \qquad (3.6.3)$$

with the notation $\langle x \rangle = x$ if $x > 0$ and zero otherwise.

The circular range R, as defined here, is very closely related to the maximum arc length

$$T = \max_{1 \leq i \leq n} T_i \qquad (3.6.4)$$

and in fact

$$R = (2\pi - T). \qquad (3.6.5)$$

From (3.6.3) and the relation (3.6.5), one can obtain the distribution function of the maximum arc length, say $G(t)$ as

$$G(t) = \sum_{j=0}^{\infty} (-1)^{j} \binom{n}{j} \left\langle 1 - \frac{jt}{2\pi} \right\rangle^{n-1}. \qquad (3.6.6)$$

It may be noted that a distribution of similar form was

obtained by Fisher (1929), who used it to construct a test of significance of the largest amplitude in harmonic analysis. He also gave a table of percentage points for sample sizes $n = 5(1)50$ at two levels namely, 5 per cent and 1 per cent, which can as well be used for testing uniformity on the circle. But the test based on the circular range, R (or equivalently on the maximum arc length, T) is however not likely to be as good as that based on U_n. Some further comments regarding this will be made in the next Chapter.

CHAPTER IV

PITMAN EFFICIENCIES OF TESTS BASED ON ARC LENGTHS

4.1 Introduction and Summary

In this chapter we study the Pitman's asymptotic relative efficiencies (ARE's) of tests based on arc lengths, that have been introduced in Chapter III for testing uniformity on the circle. The investigations contained in this chapter are quite general and apply with equal force to the goodness of fit problems on the line, as will be clear from the remarks that follow. Tests based on spacings have been used for goodness of fit on the line by several authors. See e.g. Greenwood (1946), Kimball (1950), Sherman (1950) and Darling (1953). Given that $X_1, \ldots, X_n$ are n independently and identically distributed random variables with common distribution function (d.f.) $G(x)$, the goodness of fit problem on the line is to test whether $G(x)$ is a specified d.f. When the latter d.f. is continuous, a simple probability integral transformation on the random variables (r.v.'s) would permit us to equate the preassigned d.f. to the uniform d.f. on $[0,1]$. From now on, we assume that such a reduction has been effected and that under the hypothesis, $G(x)$ is the uniform d.f. on $[0, 1]$. If $X_1' \leq \ldots \leq X_n'$ are the order statistics, the sample spacings, $(n+1)$ in number unlike on the circle, are defined by

(4.1.1) $$D_i = X'_i - X'_{i-1}, \qquad i = 1,\ldots, n+1$$

where we put $X'_0 = 0$ and $X'_{n+1} = 1$. In order that this definition of sample spacings be meaningful under any alternative, the d.f. $G(x)$ must have the carrier $[0,1]$. (The carrier of a d.f. is the smallest closed set with probability one).

Now let $\alpha_1,\ldots,\alpha_n$ be n random variables distributed independently and identically on a circle of unit circumference. The null hypothesis is one which states that the distribution is uniform on the circle. Ordering the observations as $\alpha'_1 \leq \ldots \leq \alpha'_n$, the n sample spacings which are the arc lengths between the successive sample observations are given by

(4.1.2) $$D_i = \alpha'_i - \alpha'_{i-1}, \quad i = 1,\ldots, n$$

where we put $\alpha'_0 = \alpha'_n - 1$. The letter T_i, as used in (3.3.4), will be reserved for spacings from a uniform distribution while the spacings in general will be denoted by D_i. As we have noted in Section 3.3, under the null hypothesis, the distribution of these n circular spacings is the same as those from a sample of size $(n-1)$ from a uniform distribution on the unit interval $[0, 1]$. Under the alternative, we

can choose and fix an arbitrary point on the circumference as the origin and cut open the circle at that point to get the line segment [0, 1]. Now (n-1) of the circular spacings which do not contain the cut-off point, will have the same distribution as (n-1) spacings on [0, 1], not containing the end-points 0 and 1, while the n^{th} circular spacing containing the cut-off point will have the same distribution as the sum of the remaining two linear spacings. It is easy to see therefore that the limiting distribution of the empirical d.f. of the spacings, which forms the basic tool in our study, remains the same in the circular and linear cases. Hence all the statements regarding the ARE's of spacings tests made in this chapter hold for the circular as well as linear situations. We shall, therefore, deal with the spacings on [0, 1] from the next section, keeping in mind that all the results hold equally well for the circular situation, in which our main interest lies. It is fortunate that the study of the circular situation fits nicely into the linear case in the asymptotics and helps throw light on the spacings tests on the line.

To compute Pitman efficiencies, one usually obtains the limit distributions of the test statistics under the alternative which, in general, poses a difficult problem. However, for calculating the Pitman's ARE's, it is enough to obtain the

limiting distributions under a sequence of alternatives which converge to the hypothesis. This problem turns out to be somewhat simpler. We, therefore, choose the following sequence of alternatives

$$G(x) = x + L(x)/n^{\partial} , \qquad 0 \leq x < 1$$

where ∂ is a number $\geq 1/4$ and $L(x)$ is twice continuously differentiable on $[0, 1]$. These conditions already imply our earlier requirement that the carrier of $G(x)$ be $[0, 1]$. We shall say that this alternative is at a distance of order $n^{-\partial}$ from the hypothesis. For example, we may choose $L(x) = (\sin 2\pi x)/2\pi$, $0 \leq x < 1$, in which case the d.f. $G(x)$ refers to close CN alternatives or one may take for instance $L(x) = (x^2 - x)$.

For the D_i defined in (4.1.2), $E(D_i) = 1/n$ for all i under the null hypothesis. We will therefore call $\{nD_i, \ i = 1, \ldots, n\}$ as 'normalised' spacings. Further if $h_{n1}, \ldots, h_{nn}$ be some positive numbers, then we shall call $\{nD_i/h_{ni}, \ i = 1, \ldots, n\}$ as 'modified' or 'adjusted' spacings. The rationale behind dealing with these modified spacings is that, in some cases, one may choose to adjust the spacings by their expectations under some alternate distribution, or otherwise to enlarge the class of statistics based on spacings.

This may be thought of as being analogous to the use of 'normal scores' and other scores in rank tests.

In Section 4.5, we obtain the limiting distributions of the empirical d.f.'s of the normalised and modified spacings in the sense of weak convergence of measures in an appropriate complete separable metric space. These results are the most crucial ones in this chapter. Their proofs are quite long and require Sections 4.2, 4.3 and 4.4 in which, after describing some distributional equivalences, many interesting results of independent interest, concerning the empirical d.f.'s of 'perturbed random variables' and 'randomly scaled random variables' are proved. Appealing to the invariance principle, we immediately have the limiting distributions of a large class of statistics which are symmetric in the normalised and modified spacings.

From this, we deduce in Section 4.6 that tests which are symmetric in the normalised spacings can have limiting power greater than the test size only if $\delta = 1/4$ i.e., they can not discriminate alternatives which are at a distance of order $n^{-\delta}$ from the hypothesis, for any $\delta > 1/4$. This character of the symmetric spacings tests has been pointed out in a paper of Gibisov (1961), which has not attracted the attention it deserves. The importance of our results lies in

the fact that they allow for the computation of the ARE's of symmetric sample spacings tests. We find that among the many such standard tests due to Greenwood (1946), Kimball (1950), Sherman (1950) and Darling (1953), the one due to Darling based on $n^{-1} \sum_{1}^{n} \log (nD_i)$ has the maximum ARE. We note another interesting feature of the tests symmetric in normalised spacings, namely their ARE's do not depend on the particular choice of the alternative sequence chosen i.e., on $L(x)$. However, this feature is not shared by tests symmetric in the modified spacings, which are discussed below.

The theorems of Section 4.5 also enable us to study the ARE's of tests which are symmetrically based on the modified spacings. A new feature arises here. Consider for instance the modified spacings $\{(n-i+1)D_i, \ i = 1,\ldots, n\}$ and consider the test based on their sum. This is essentially a test based on the sample mean and hence can distinguish alternatives at a distance of order $n^{-1/2}$ from the hypothesis if the mean under the alternative is not equal to $1/2$ i.e., if

$$\int_0^1 x \, dL(x) \neq 0.$$

When $\int_0^1 x \, dL(x) = 0$, this test may still have limiting power greater than the test size when the alternative is at a distance of order $n^{-1/4}$ from the hypothesis and so be on a par.

with tests which are symmetric in the normalised spacings. In Section 4.6, we demonstrate this behaviour of tests symmetric in modified spacings.

It is however true that there exist tests which always distinguish alternatives at a distance of order $n^{-1/2}$ from the hypothesis. The Kolmogorov-Smirnov test is an example. That tests based symmetrically on normalised or modified spacings can not do as well is quite disturbing. One point to be learnt from the debacle of the symmetric spacings tests is this. The sample spacings form a sufficient statistic. They come quite close to the case of independently and identically distributed r.v.'s in the sense that they even form a set of interchangeable r.v.'s under the hypothesis. But the order of their occurrence is quite important in the case of these spacings, in complete constrast to the general principle of basing tests symmetrically on the available data, whenever it consists of independently and identically distributed observations.

4.2 Preliminaries

Let $X_1, X_2, \ldots, X_{n-1}$ be $(n-1)$ independently and identically distributed random variables with a continuous d.f. $G_n(x)$, whose carrier is $[0, 1]$, $n = 2,3,\ldots$. $X_1, \ldots, X_{n-1}$ may also depend on n, but we shall suppress this in our

notations throughout. $G_n(x)$ is a sequence of alternative distributions, which converges to the uniform distribution on [0, 1], the distribution specified by the hypothesis. In particular, we assume $G_n(x)$ to be of the form

$$G_n(x) = x + L(x)/n^{\partial} \tag{4.2.1}$$

for $x \in [0,1]$ where ∂ is a fixed constant $\geq 1/4$. We impose the following regularity condition on $L(x)$. $L(x)$ is twice continuously differentiable on $[0,1]$. (4.2.1) together with this condition will be referred to as assumption (A). If $\lambda(x)$ and $\lambda'(x)$ denote the first and second derivatives respectively of $L(x)$, then we note that there is a constant L_0 such that

$$|L(x)| \leq L_0, \quad |\lambda(x)| \leq L_0, \quad |\lambda'(x)| \leq L_0 \tag{4.2.2}$$

$$\text{for all} \quad x \in [0,1].$$

The inverse function of $G_n(x)$ is denoted by $G_n^{-1}(p)$, $0 \leq p \leq 1$. We define

$$k_n(p) = g_n[G_n^{-1}(p)] = [\, dG_n^{-1}(p)/dp]^{-1} \,. \tag{4.2.3}$$

It may be verified that in our case

$$G_n^{-1}(p) = p - L(p)/n^{\partial} + o(1/n^{\partial}) \tag{4.2.4}$$

$$k_n(p) = 1 + \alpha(p)/n^{\partial} - L(p)\,\alpha'(p)/n^{2\partial} + o(1/n^{2\partial}) \tag{4.2.5}$$

where $o(\cdot)$ is uniform in p.

We will obtain several limit distributions under the sequence of alternatives $G_n(x)$ satisfying assumption (A). It is clear, however, that the limit distributions under the hypothesis are obtained by putting $L(x) \equiv 0$. We will make some further remarks about these alternatives in Section 4.6. Let the random variables (r.v.'s) $X_1,\ldots,X_{n-1}$ be arranged in increasing order of magnitude thus

$$0 \le X_1' \le \cdots \le X_{n-1}' \le 1. \tag{4.2.6}$$

The sample spacings have been defined in Section 4.1 as

$$D_i = X_i' - X_{i-1}', \quad i = 1,\ldots, n \tag{4.2.7}$$

where we put $X_0' = 0$, $X_n' = 1$.

We first relate these sample spacings D_i to the spacings based on uniformly distributed r.v.'s on $[0,1]$ (to be called uniform sample spacings). Let $U_1,\ldots,U_{n-1}$ be

(n-1) independently and identically distributed r.v.'s with a uniform distribution on [0,1]. These are arranged in increasing order of magnitude thus

$$0 \leq U_1' \leq \cdots \leq U_{n-1}' \leq 1.$$

The uniform sample spacings are defined by

$$T_i = U_i' - U_{i-1}', \quad i = 1, \ldots, n \tag{4.2.8}$$

where again, we put $U_0' = 0$, $U_n' = 1$.

For two r.v.'s X and Y, we write $X \sim Y$ to mean that X and Y are distributionally equivalent, that is, the distributions of X and Y are identical. We know that

$$(X_i', \quad i = 0, \ldots, n) \sim (G_n^{-1}(U_i'), \quad i = 0, \ldots, n)$$

and thus

$$(D_i, \quad i = 1, \ldots, n) \sim (G_n^{-1}(U_i') - G_n^{-1}(U_{i-1}'), \ i = 1, \ldots, n)$$

$$= (T_i / k_n(\tilde{U}_i), \quad i = 1, \ldots, n)$$

where $U_{i-1}' \leq \tilde{U}_i \leq U_i'$

$$= (T_i / \alpha_{ni}^*, \quad i = 1, \ldots, n) \tag{4.2.9}$$

where

$$\alpha^*_{ni} = 1 + \beta^*_{ni}/n^{\partial} + \gamma^*_{ni}/n^{2\partial} + R^*_{ni} \tag{4.2.10}$$

with

$$\beta^*_{ni} = \alpha'(\hat{U}_i) \tag{4.2.11}$$

$$\gamma^*_{ni} = -L(\hat{U}_i)\,\alpha''(\hat{U}_i) \tag{4.2.12}$$

and

$$\sup_i \sqrt{n}\,|R^*_{ni}| \to 0 \quad \text{almost everywhere} \tag{4.2.13}$$

in view of (4.2.5). Also, from the existence of the limiting distribution of the Kolmogorov-Smirnov statistic,

$$\sup_i \sqrt{n}\,|U_i - i/n| = O_p(1). \tag{4.2.14}$$

Thus from the continuity of L, α and α',

$$\sup_i \sqrt{n}\,|\beta^*_{ni} - \beta(i/n)| = O_p(1) \tag{4.2.15}$$

$$\sup_i |\gamma^*_{ni} - \gamma(i/n)| = O_p(1) \tag{4.2.16}$$

where

$$\beta(p) = \alpha'(p) \tag{4.2.17}$$

$$\gamma(p) = -L(p)\,\alpha''(p), \quad 0 \le p \le 1. \tag{4.2.18}$$

Now let $W_1, \ldots, W_n$ be n independently and identically distributed exponential r.v.'s with density function $e^{-w}, w \ge 0$

Let $W_n^* = (W_1 + \ldots + W_n)$ and let $\bar{W}_n = W_n^*/n$. Then it is well known that

$$(T_i, \quad i = 1,\ldots, n) \sim (W_i/W_n^*, \quad i = 1,\ldots, n).$$

Thus (4.2.9) may be rewritten as

$$(4.2.19) \qquad (D_i, \quad i = 1,\ldots, n) \sim (W_i/\alpha_{ni}^{**} W_n^*, \quad i = 1,\ldots,n)$$

where

$$(4.2.20) \qquad (\alpha_{ni}^{**}, \; i = 1,\ldots, n) \sim (\alpha_{ni}^{*}, \; i = 1,\ldots, n)$$

In view of (4.2.20), we save on notation by writing α_{ni}^* for α_{ni}^{**} and retain its structure defined in (4.2.10) and will later on utilise the properties (4.2.13), (4.2.15) and (4.2.16).

The distributional equivalence in (4.2.19) is well known and has been used by others, e.g. Weiss (1965). However, this is the first time a systematic use of (4.2.19) has been made to obtain the asymptotic distributions of the empirical d.f.'s of the normalised spacings and modified spacings under the general alternatives described in (4.2.1).

The empirical d.f., $H_n(x)$ of the normalised spacings, which is of central interest in this chapter is defined as follows

$$H_n(x) = \sum_1^n I(nD_i;\ x)/n, \quad x \geq 0 \tag{4.2.21}$$

where

$$I(z;\ x) = \begin{cases} 1 & \text{if} \quad z \leq x \\ 0 & \text{if} \quad z > x\ . \end{cases} \tag{4.2.22}$$

The **rationale** for using nD_i's has been explained before. Using the equivalence (4.2.19), we note that

$$\{ H_n(x),\ x \geq 0 \} \sim \{ \sum_1^n I(W_i/\alpha^*_{ni}\ \bar{W}_n;\ x)/n, \quad x \geq 0 \}$$

$$= \{ F^*_n(x \cdot \bar{W}_n), \quad x \geq 0 \} \tag{4.2.23}$$

where

$$F^*_n(x) = \sum_1^n I(W_i/\alpha^*_{ni}\ ;\ x)/n. \tag{4.2.24}$$

(4.2.23) says that the distributions of the stochastic processes $\{ H_n(x), \quad x \geq 0 \}$ and $\{ \sum_1^n I(W_i/\alpha^*_{ni}\ \bar{W}_n;\ x)/n, \quad x \geq 0 \}$ in some suitable space (see discussion after (4.3.11)) coincide and this distributional equivalence is stronger than the distributional equivalence of the finite dimensional marginals. Anticipating our later definitions in Sections 4.3 and 4.4, $F^*_n(x)$ is the empirical d.f. of $W_1, \ldots, W_n$ with random perturbations and $F^*_n(x\ \bar{W}_n)$ is the empirical d.f. of $W_1, \ldots, W_n$ with random perturbations and a random scale factor $\bar{W}_n$.

If $(h_{n1}, h_{n2}, \ldots, h_{nn})$, $n = 1,2,\ldots$ be a triangular array of positive constants, define

(4.2.25) $\quad D_i^* = nD_i / h_{ni}, \quad i = 1,\ldots, n.$

We shall call $(D_1^*,\ldots, D_n^*)$ modified spacings and the empirical d.f. of these modified spacings is defined by $H_n^*(x)$ where

(4.2.26) $\quad H_n^*(x) = \sum_1^n I(D_i^* ; x)/n .$

From (4.2.19), it follows that

(4.2.27) $\quad \{ H_n^*(x), \quad x \geq 0 \} \sim \{F_n^*(x \bar{W}_n), \quad x \geq 0 \}$

where the α_{ni}^*'s used in the definition (4.2.24) of $F_n^*(x)$ here are distributionally equivalent as follows:

(4.2.28) $\quad \{ \alpha_{ni}^*, \quad i = 1,\ldots, n \} \sim$

$$\{ h_{ni}(1 + \beta_{ni}^*/n^{\partial} + \gamma_{ni}^* /n^{2\partial} + R_{ni}^*), \quad i = 1,\ldots, n\}$$

where β_{ni}^*, γ_{ni}^* and R_{ni}^* satisfy the conditions laid down in (4.2.15), (4.2.16) and (4.2.13).
As in the remark after (4.2.20), we replace the symbol in (4.2.28) by '=' in order to avoid introducing new notations.

Thus we have reduced the problem of finding the asymptotic distributions of $H_n(x)$ and $H_n^*(x)$ to finding that of $F_n^*(x \bar{W}_n)$. In Section 4.3, we derive the asymptotic

distribution of an empirical d.f. with non-random perturbations and then allow for random perturbations. Finally, in Section 4.4 we allow random scale factors. These terminologies are made precise in the following sections. The results presented in these sections are more general than are necessary for our purposes and are of independent interest by themselves.

4.3 Asymptotic distribution of the empirical d.f. of random variables subject to perturbations

Let $Z_1, Z_2, \ldots$ be independently and identically distributed r.v.'s with a common d.f. $F(x)$ with $F(0) = 0$. We assume that $F(x)$ is thrice differentiable. Let $f(x)$, $f'(x)$ and $f''(x)$ denote the first, second and third derivatives respectively of $F(x)$. We impose the following blanket condition (B) on $F(x)$

(B) $xf(x)$, $x^2f'(x)$ and $x^3f''(x)$ are bounded on $[0, \infty]$.

Let $\{\alpha_{nil}, \quad i = 1, \ldots, n, \; n = 1, 2, \ldots\}$

be a triangular array of constants. Then the random variables

$$\{Z_{ni} = Z_i/\alpha_{nil}, \quad i = 1, \ldots, n\}$$

are said to be perturbed random variables, $n = 1, 2, \ldots$.

Let

(4.3.1) $$F_{nl}(x) = \sum_{1}^{n} I(Z_{ni};\ x)/n$$
$$= \sum_{1}^{n} I(Z_i/\alpha_{nil};\ x)/n.$$

We refer to $F_{nl}(x)$ as the empirical d.f. of $(Z_1, \ldots, Z_n)$ under a perturbation by the non-random quantities $\{\alpha_{nil},\ i = 1, \ldots, n\}$. The following structure is assumed of $\{\alpha_{nil},\ i = 1, \ldots, n\}$. There exist continuous functions $\beta(p)$ and $\gamma(p)$ on $[0, 1]$ such that

(4.3.2) $$\alpha_{nil} = 1 + \beta(i/n)/n^{\partial} + \gamma(i/n)/n^{2\partial} + R_{ni}$$

where ∂ is a constant $\geq 1/4$ and

(4.3.3) $$\sup_i \sqrt{n}\ |R_{ni}| \to 0 \quad \text{as} \quad n \to \infty.$$

If $\partial > 1/2$, then the second and third terms on the right hand side (RHS) of (4.3.2) can be absorbed into R_{ni} and if $1/4 < \partial \leq 1/2$, then the third term of the RHS can be absorbed into R_{ni}. We note that $\alpha_{nil} \to 1$ uniformly in i, so that without loss of generality we may assume that

(4.3.4) $$1/2 \leq \alpha_{nil} \leq 2$$

for all n and i. Let

$$(4.3.5) \qquad F^{+}_{nl}(x) = E(F_{nl}(x)) = E\Big(\sum_{1}^{n} I(Z_i/\alpha_{nil};\ x)/n\Big) = \sum_{1}^{n} F(x\,\alpha_{nil})/n\ .$$

It is easy to see that $F^{+}_{nl}(x)$ tends to $F(x)$ uniformly in x.

The main theorem of this section will establish the limiting distribution, in an appropriate sense to be made precise very soon, of the stochastic process

$$(4.3.6) \qquad \{\ \eta_{nl}(x) = \sqrt{n}\,[F_{nl}(x) - F^{+}_{nl}(x)],\quad x \geq 0\ \}\ .$$

We now state two lemmas required in the proof.

Lemma 4.3.1: For $n = 1,2,\ldots$, let $\{\ Y_{ni},\quad i = 1,\ldots,n\ \}$ be independently distributed with

$$(4.3.7) \qquad P(Y_{ni} = 1) = p_{ni},\quad P(Y_{ni} = 0) = 1 - p_{ni},\quad i = 1,..,n.$$

Let

$$(4.3.8) \qquad Y_n = \sum_{1}^{n} (Y_{ni} - p_{ni}) \Big/ \sqrt{\sum_{1}^{n} p_{ni}(1 - p_{ni})}\ ..$$

Then as $n \to \infty$,

$$P(Y_n \leq x) \to \bar{\Phi}(x)$$

for each x, where

$$\bar{\Phi}(x) = \int_{-\infty}^{x} \exp(-t^2/2)\,dt\ /\ \sqrt{2\pi}$$

if and only if

$$(4.3.9) \qquad S_n^2 = \sum_1^n p_{ni}(1-p_{ni}) \to \infty.$$

See for example Fisz (1963) p. 207 for a proof. A sufficient condition for (4.3.9) to hold is

$$(4.3.10) \qquad \sum_1^n p_{ni}/n \text{ is bounded away from } 0 \text{ and } 1.$$

Lemma 4.3.2:

Let (Y_1, Y_2, Y_3) be a trinomial r.v. with $Y_1+Y_2+Y_3 = 1$, $Y_i = 0$ or 1, $P(Y_i = 1) = p_i$, $i = 1,2,3$ and $p_1+p_2+p_3 = 1$. If $Y_i^* = (Y_i - p_i)$, $i = 1,2,3$, then

$$(4.3.11) \quad E(Y_i^*) = 0 \qquad , E(Y_i^{*2}) = p_i(1-p_i)$$

$$E(Y_i^* Y_j^*) = -p_i p_j$$

$$E(Y_i^{*2} Y_j^{*2}) = p_i p_j (1-p_i)(1-p_j) - p_i p_j (1-2p_i)(1-2p_j)$$

$$i \neq j; \quad i, j = 1,2,3.$$

We now briefly describe the space in which our processes $\{\eta_{n1}(x), x \geq 0\}$, defined in (4.3.6), lie and define weak convergence of processes on this space. Consider the space $D[0,\infty]$ of functions, $p(x)$, on $[0,\infty]$ with the properties

i) $p(x+0)$, $p(x-0)$ *exist for each* x *in* $(0,\infty)$ *and*

$$p(x) = p(x+0)$$

ii) $p(0+0)$ exists and is equal to $p(0)$

iii) $\lim_{x \to \infty} p(x)$ exists and is equal to $p(\infty)$.

A sequence $\{p_n(x)\}$ in $D[0,\infty]$ converges to $p(x)$ in $D[0,\infty]$ if there exists a sequence $\{\lambda_n(x)\}$ of one-to-one monotonic continuous maps of $[0,\infty]$ onto $[0,\infty]$ such that

$$\sup_x |\lambda_n(x) - x| \to 0$$

and

$$\sup_x |p_n(\lambda_n(x)) - p(x)| \to 0.$$

This convergence corresponds to the J_1-topology of Skorohod on the space $D[0,\infty]$.

A sequence of stochastic processes $\{p_n(x),\ x \geq 0\}$ converges weakly to a process $\{p(x),\ x \geq 0\}$ in $D[0,\infty]$ if

$$(4.3.12) \qquad E(s(p_n(\cdot))) \to E(s(p(\cdot)))$$

for every function $s(\cdot)$ on $D[0,\infty]$ which is bounded and continuous in the topology just described. When this happens, we have by the invariance principle, the most useful conclusion that the distribution of the real valued r.v. $s(p_n(\cdot))$ converges weakly to the distribution of $s(p(\cdot))$ for every function $s(\cdot)$ on $D[0,\infty]$, which is continuous almost everywhere with respect to $p(\cdot)$.

Let $\mu(x)$ be a one-to-one monotone continuous map of $[0, \infty]$ onto $[0, 1]$. For any function $q(\mu)$ in $D[0,1]$, let $p(x) = q(\mu(x))$. Then this map from $D[0,1]$ to $D[0,\infty]$ is continuous and has a continuous inverse (when $D[0, 1]$ is endowed with the J_1-topology of Skorohod (1956)) and is therefore a homeomorphism. Thus the study of probability measures on $D[0,\infty]$ can be reduced to the study of probability measures on $D[0, 1]$, which is by now classical. See for instance Skorohod (1956), Sethuraman (1965). A sequence of stochastic processes $\{ q_n(\mu), 0 \leq \mu \leq 1 \}$ is said to converge weakly to a stochastic process $\{ q(\mu), 0 \leq \mu \leq 1 \}$ if for any bounded continuous function $r(\cdot)$ on $D[0,1]$,

$$E[r(q_n(\cdot))] \to E[r(q(\cdot))].$$

We shall make use of the following standard theorem of compactness and convergence of a sequence of stochastic processes in $D[0, 1]$.

Theorem 4.3.3: Let $\{q_n(\mu), 0 \leq \mu \leq 1, \}$ $n = 1,2,\ldots$ and $\{q(\mu), 0 \leq \mu \leq 1 \}$ be stochastic processes with values in $D[0, 1]$ such that

i) The marginal distributions of $\{ q_n(\mu_1),\ldots,q_n(\mu_k) \}$ converge to those of $\{q(\mu_1),\ldots,q(\mu_k)\}$ weakly, for every finite subset $\mu_1,\ldots,\mu_k$ of $[0,1]$.

ii) There exists a constant C such that

(4.3.13) $E \{ |q_n(\mu_1) - q_n(\mu_2)|^2 |q_n(\mu_2) - q_n(\mu_3)|^2 \} \leq C h^2$

whenever $0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq 1$ and $\mu_3 - \mu_2 \leq h$, $\mu_2 - \mu_1 \leq h$. Then the sequence of stochastic processes $\{ q_n(\mu),\ 0 \leq \mu \leq 1 \}$ converges weakly to $\{q(\mu),\ 0 \leq \mu \leq 1\}$.

This theorem is essentially due to Chentsov (1956) and can also be found in Sethuraman (1965).

Now, let $\mu(x)$ be a one-to-one monotone continuous transformations of $[0,\infty]$ onto $[0,1]$. Let $p_n(x) = q_n(\mu(x))$ and $p(x) = q(\mu(x))$, $0 \leq x \leq \infty$. Since $\mu(x)$ is a homeomorphism the sequence of stochastic processes $\{ q_n(\mu),\ 0 \leq \mu \leq 1 \}$ converges weakly to $\{ q(\mu),\ 0 \leq \mu \leq 1 \}$ if and only if the sequence of processes $\{p_n(x),\ 0 \leq x \leq \infty \}$ converges weakly to $\{ p(x),\ 0 \leq x \leq \infty \}$. This provides us the technique of investigating the convergence of our empirical d.f. processes.

After this digression on the definition of $D[0,\infty]$ and weak convergence of processes on it, we return to our empirical d.f. processes. We note that for each n, $\{ \eta_{n1}(x),\ x \geq 0 \}$ defined in (4.3.6) is in $D[0,\infty]$ and is measurable if we put $\eta_{n1}(\infty) = 0$.

Theorem 4.3.4: Let condition (B) hold. The sequence $\{ \eta_{n1}(x),\ x \geq 0 \}$ considered as a stochastic process in $D[0,\infty]$ converges weakly to a Gaussian process $\{\eta_1(x),\ x \geq 0\}$ with mean zero and covariance function

$$(4.3.14) \qquad K_1(x,y) = K_1(y,x) = F(x)(1 - F(y)) \qquad \text{for } x \leq y.$$

roof: Define the processes $\{ y_n(\mu),\ 0 \le \mu \le 1 \}$, $\{ z_n(\nu),\ \le \nu \le 1 \}$ and $\{ y(\mu),\ 0 \le \mu \le 1 \}$ in $D[0,1]$ by

4.3.15)
$$y_n(F_{nl}^{+}(x)) = \eta_{nl}(x)$$
$$z_n(F(x)) = \eta_{nl}(x)$$
$$y\ (F(x)) = \eta_1(x).$$

ur method of proof is to establish the following in order.

1. The sequence of processes $\{ y_n(\mu),\ 0 \le \mu \le 1 \}$ in $D[0,1]$ nverges weakly to the process $\{ y(\mu),\ 0 \le \mu \le 1 \}$.

2. The sequence of processes $\{ z_n(\nu),\ 0 \le \nu \le 1 \}$ in $D[0,1]$ nverges weakly to the process $\{ y(\nu),\ 0 \le \nu \le 1 \}$.

3. The sequence of processes $\{ \eta_{nl}(x),\ x \ge 0 \}$ in $D[0,\infty]$ nverges weakly to the process $\{ \eta_1(x),\ x \ge 0 \}$.

Fix x. Then,

$$\eta_{nl}(x) \Big/ \sqrt{\sum_1^n F(x\ \alpha_{nil})(1 - F(x\ \alpha_{nil}))/n}$$

$$= \sum_1^n (I(Z_i/\alpha_{nil};\ x) - F(x\ \alpha_{nil})) \Big/ \sqrt{\sum_1^n F(x\ \alpha_{nil})(1 - F(x\ \alpha_{nil}))}$$

nverges to the normal variable with mean zero and unit variance, $n \to \infty$. This can be seen easily from Lemma 4.3.1. Since

$$\sum_1^n F(x\ \alpha_{nil})(1 - F(x\ \alpha_{nil}))/n \to F(x)(1 - F(x)),$$

$\eta_{n1}(x)$ converges to a normal r.v. with mean zero and variance $F(x)(1-F(x))$. A similar application of the multivariate version of Lemma 4.3.1 will show that the finite dimensional marginal distributions of $\{\eta_{n1}(x),\ x \geq 0\}$ converges weakly to those of $\{\eta_1(x),\ x \geq 0\}$. Hence the finite dimensional marginal distributions of $\{y_n(\mu),\ 0 \leq \mu \leq 1\}$ converge to those of the Gaussian process $\{y(\mu),\ 0 \leq \mu \leq 1\}$. Next, let $0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq 1$. Then

$$y_n(\mu_j) = \eta_{n1}(x_j)$$
$$= \sum_{i=1}^{n} [I(Z_i/\alpha_{ni1};\ x_j) - F(x_j\ \alpha_{ni1})]/\sqrt{n}$$

where

$$F^{+}_{n1}(x_j) = \mu_j, \quad j = 1,2,3. \tag{4.3.16}$$

Thus

$$y_n(\mu_2) - y_n(\mu_1) = \sum_{i=1}^{n} V_{1i} / \sqrt{n} \tag{4.3.17}$$

$$y_n(\mu_3) - y_n(\mu_2) = \sum_{k=1}^{n} V_{2k}/\sqrt{n}$$

where

$$V_{1i} = I(Z_i/\alpha_{ni1};\ x_2) - I(Z_i/\alpha_{ni1};\ x_1) - F(x_2\alpha_{ni1}) + F(x_1\alpha_{ni1})$$
$$V_{2k} = I(Z_k/\alpha_{nk1};\ x_3) - I(Z_k/\alpha_{nk1};\ x_2) - F(x_3\alpha_{nk1}) + F(x_2\alpha_{nk1}) \tag{4.3.18}$$

$i,k = 1,\ldots, n$. Since $1/2 \leq \alpha_{ni1}$ from (4.3.4), these are binomial r.v.'s corrected for their means. Then

$$E\{(y_n(\mu_2) - y_n(\mu_1))^2 (y_n(\mu_3) - y_n(\mu_2))^2\}$$

$$= n^{-2} E\{(\sum_i V_{1i})^2 (\sum_k V_{2k})^2\}$$

$$= n^{-2} E\{(\sum_i V_{1i}^2 + \sum_{i\neq j} V_{1i}V_{1j})(\sum_k V_{2k}^2 + \sum_{k\neq \ell} V_{2k}V_{2\ell})\} \quad (4.3.19)$$

$$= n^{-2}\{\sum_{i,k} E(V_{1i}^2 V_{2k}^2) + \sum_{i\neq j,k} E(V_{1i}V_{1j}V_{2k}^2)$$

$$+ \sum_{i,k\neq \ell} E(V_{1i}^2 V_{2k}V_{2\ell}) + \sum_{i\neq j,k\neq \ell} E(V_{1i}V_{1j}V_{2k}V_{2\ell})\}.$$

The second and third terms of the last expression vanish whereas, in the last term non-zero contributions occur only when $i = k$, $j = \ell$ or $i = \ell$ and $j = k$. Let

$$p_{1i} = F(x_2\alpha_{ni1}) - F(x_1\alpha_{ni1})$$
$$(4.3.20)$$
$$p_{2k} = F(x_3\alpha_{nk1}) - F(x_2\alpha_{nk1}).$$

From Lemma 4.3.2

$$E(V_{1i}^2 V_{2k}^2) = p_{1i}(1 - p_{1i})p_{2k}(1 - p_{2k}) \quad \text{for} \quad i \neq k$$

$$= p_{1i}p_{2k}(1 - p_{1i})(1 - p_{2k}) - p_{1i}p_{2k}(1 - 2p_{1i})(1 - 2p_{2k})$$

$$\text{for} \quad i = k$$

which in any case is smaller than $p_{1i}p_{2k}$. Also

$$E(V_{1i}V_{2i}) = -p_{1i}p_{2i}.$$

Substituting these in (4.3.19)

$$E\{(y_n(\mu_2) - y_n(\mu_1))^2 (y_n(\mu_3) - y_n(\mu_2))^2\}$$

$$\leq 2\left(\sum_1^n p_{1i}/n\right)\left(\sum_1^n p_{2i}/n\right)$$

$$\leq 2C|\mu_2 - \mu_1| \cdot |\mu_3 - \mu_2|.$$

Thus from Theorem 4.3.3., the sequence $\{y_n(\mu),\ 0 \leq \mu \leq 1\}$ converges weakly to the Gaussian process $\{y(\mu),\ 0 \leq \mu \leq 1\}$. Consider the sequence of homeomorphisms $\lambda_n(\cdot)$ of $[0,1]$ onto $[0,1]$ given by

$$\lambda_n(\mu) = F(F_{n1}^{+-1}(\mu)) \tag{4.3.21}$$

where $F_{n1}^{+-1}(\cdot)$ is the inverse function of $F_{n1}^{+}(\cdot)$. It is easy to see that

$$\sup_{\mu} |\lambda_n(\mu) - \mu| \to 0,$$

and

$$z_n(\nu) = z_n(\lambda_n(\mu)) = y_n(\mu).$$

Hence the weak convergence of the sequence $\{y_n(\mu),\ 0 \leq \mu \leq 1\}$ implies the weak convergence of $\{z_n(\nu),\ 0 \leq \nu \leq 1\}$ to the Gaussian process $\{y(\nu), 0 \leq \nu \leq 1\}$. The transformation $x \to F(x)$ is a homeomorphism between $[0,\infty]$ and $[0,1]$ under our assumptions. From this and the relationships (4.3.15), it follows that the sequence of processes $\{\eta_{n1}(x),\ x \geq 0\}$ in $D[0,\infty]$ converges weakly to the Gaussian process $\{\eta_1(x),\ x \geq 0\}$

whose mean is zero and covariance is given by (4.3.14).

Remark: When condition (B) holds, we can replace $F^{+}_{n1}(x)$ which enters the definition of $\eta_{n1}(x)$ in (4.3.6) by

$$F^{+}_{n1}(x) = \begin{cases} F(x) & \text{if } \partial > 1/2 \\ F(x) + xf(x)\int_0^1 \beta(p)\,dp/n^{\partial} & \text{if } 1/4 < \partial \leq 1/2 \\ F(x) + xf(x)\int_0^1 \beta(p)\,dp/n^{1/4} & \\ \quad + [xf(x)\int_0^1 \gamma(p)\,dp + x^2 f'(x)\int_0^1 \beta^2(p)\,dp/2]/n^{1/2} & \text{if } \partial = 1/4 \end{cases}$$

after omitting terms which are of smaller order than $n^{-1/2}$ uniformly in x. The most general conditions under which $F^{+}_{n1}(x)$ can be replaced as above must depend on ∂. However since we are contemplating only the application with $F(x) = 1 - \exp(-x)$ in Section 4.6, we will content ourselves by imposing the blanket condition (B).

To allow for perturbations by constants which are more general than given in (4.3.2), we consider a triangular sequence $\{\alpha_{ni2},\ i = 1,\ldots,n\}$, $n = 1,2,\ldots$ with the following structure

$$\alpha_{ni2} = \theta(i/n)[1 + \beta(i/n)/n^{\partial} + \gamma(i/n)/n^{2\partial} + R_{ni}]$$

where $\beta(p)$ and $\gamma(p)$ are continuous functions on $[0,1]$ and R_{ni} satisfies (4.3.3). We put the following condition (C) on $\theta(p)$.

(C) $\theta(p)$ is continuous on $[0,1]$ except at a finite number of points and, for each x, the integrals of $F(x\theta(p))$, $\theta(p)f(x\theta(p))$ and $\theta^2(p)f'(x\theta(p))$ as functions of p on $[0,1]$ exist and are finite.

We shall see later that this generalisation generates theorems from which we will be able to obtain the limiting distributions of statistics based on modified spacings. Define the empirical d.f. $F_{n2}(x)$ of the Z's perturbed by the $\{\alpha_{ni2}\}$ given in (4.3.22), by a formula similar to (4.3.1). Let $F^+_{n2}(x)$ and $\eta_{n2}(x)$ be as defined in (4.3.5) and (4.3.6), with the perturbation constants $\{\alpha_{ni2}\}$ instead of $\{\alpha_{ni1}\}$. The following theorem is proved exactly as Theorem 4.3.4 and is stated without proof.

Theorem 4.3.5: Let conditions (B) and (C) hold. The sequence of stochastic processes $\{\eta_{n2}(x),\ x \geq 0\}$ in $D(0,\infty]$ converges to the Gaussian stochastic process $\{\eta_2(x),\ x \geq 0\}$ with mean zero and covariance function $K_2(x,y)$ defined by

(4.3.23) $$K_2(x,y) = K_2(y,x)$$
$$= \int_0^1 F'(x\theta(p))[1 - F(y\theta(p))]dp \quad \text{for } x \leq y.$$

Remark: Under conditions (B) and (C), $F^{+}_{n2}(x)$, which is defined by (4.3.5) through the constants $\{\alpha_{ni2}\}$ of (4.3.22), can be replaced by

$$F^{+}_{n2}(x) = \begin{cases} \int_0^1 F(x\theta(p))\,dp & \text{if } \partial > \frac{1}{2} \\ \int_0^1 F(x\theta(p))\,dp + \int_0^1 x\beta(p)\theta(p)f(x\theta(p))\,dp/n^{\partial} & \text{if } \frac{1}{4} < \partial \le \frac{1}{2} \\ \int_0^1 F(x\theta(p))dp + \int_0^1 x\beta(p)\theta(p)f(x\theta(p))\,dp/n^{1/4} & \\ \quad + \left[\int_0^1 x\gamma(p)\theta(p)f(x\theta(p))\,dp + \int_0^1 x^2\beta^2(p)\theta^2(p)f'(x\theta(p))dp/2\right]/n^{\frac{1}{2}} & \text{if } \partial = \frac{1}{4} \end{cases}$$

up to terms of smaller order than $n^{-1/2}$ uniformly in x.

We now proceed to establish a limit theorem for the empirical d.f. of randomly perturbed r.v.'s. A star will be, in general, used as a generic symbol for denoting random quantities or the associated functions and processes. Let $\{\alpha^{*}_{ni1}, i = 1,\ldots,n\}$ $n = 1,2,\ldots$ be a triangular scheme of random variables with the form

$$\alpha^{*}_{ni1} = 1 + \beta^{*}_{ni}/n^{\partial} + \gamma^{*}_{ni}/n^{2\partial} + R^{*}_{ni} \tag{4.3.24}$$

where

$$\sup_i \sqrt{n}\,|R^{*}_{ni}| = o_p(1) \tag{4.3.25}$$

and there are continuous functions $\beta(p)$ and $\gamma(p)$ on $[0,1]$

such that

$$\sup_i \; n^{\partial^*} |\beta^*_{ni} - \beta(i/n)| = o_p(1) \tag{4.3.26}$$

$$\sup_i \; | \gamma^*_{ni} - \gamma(i/n)| = o_p(1) \tag{4.3.27}$$

where

$$\partial^* = \begin{cases} \frac{1}{2} - \partial & \text{if } \partial < \frac{1}{2} \\ 0 & \text{if } \partial \geq \frac{1}{2} . \end{cases}$$

Let

$$F^*_{n1}(x) = \sum_1^n I(Z_i / \alpha^*_{ni1}; x)/n. \tag{4.3.28}$$

This is the empirical d.f. of $Z_1, \ldots, Z_n$ perturbed by the random quantities $\{ \alpha^*_{ni1}, i = 1, \ldots, n \}$. Let the non-random quantities $\{ \alpha_{ni1} \}$ be defined as

$$\alpha_{ni1} = 1 + \beta(i/n)/n^{\partial} + \gamma(i/n)/n^{2\partial} + R_{ni} \tag{4.3.29}$$

in terms of the new $\beta(p)$ and $\gamma(p)$, and $F^+_{n1}(x)$ be defined by the relation (4.3.5) with the new α_{ni1}'s. Consider the stochastic process

$$\{\eta^*_{n1}(x) = \sqrt{n}(F^*_{n1}(x) - F^+_{n1}(x)), \quad x \geq 0 \}. \tag{4.3.30}$$

Theorem 4.3.6: Let condition (B) hold. The sequence of processes $\{ \eta^*_{n1}(x), x \geq 0 \}$ in $D[0,\infty]$ converges weakly to the Gaussian process $\{\eta_1(x), x \geq 0\}$ with mean zero and

covariance function given by (4.3.14).

Proof: Given $\epsilon_1, \epsilon_2 > 0$, there is an n_1 such that for $n \geq n_1$,

$$(4.3.31) \qquad P\left\{ \sup_i n^{\partial^*} |\beta^*_{ni} - \beta(i/n)| > \epsilon_1 \right\} \leq \epsilon_2$$

$$(4.3.32) \qquad P\left\{ \sup_i | \gamma^*_{ni} - \gamma(i/n)| > \epsilon_1 \right\} \leq \epsilon_2$$

and

$$(4.3.33) \qquad P\left\{ \sup_i \sqrt{n} |R^*_{ni}| > \epsilon_1 \right\} \leq \epsilon_2$$

as can be readily seen from (4.3.26), (4.3.27) and (4.3.25). Let

$$\alpha_{nil,1} = 1 + \left\{\beta(i/n) + \epsilon_1/n^{\partial^*}\right\}/n^{\partial} + \left\{\gamma(i/n) + \epsilon_1\right\}/n^{2\partial} + \epsilon_1/\sqrt{n}$$

$$\alpha_{nil,2} = 1 + \left\{\beta(i/n) - \epsilon_1/n^{\partial^*}\right\}/n^{\partial} + \left\{\gamma(i/n) - \epsilon_1\right\}/n^{2\partial} - \epsilon_1/\sqrt{n}$$

and

$$(4.3.34) \qquad F_{nl,j}(x) = \sum_1^n I(z_i/\alpha_{nil,j};\, x)/n, \quad j = 1,2.$$

It is easy to see that $\{\alpha_{nil,1},\ i = 1,\ldots, n\}$ and $\{\alpha_{nil,2},\ i = 1,\ldots, n\}$ satisfy the structure defined in (4.3.2) and (4.3.3). We may also assume, without loss of generality that $\alpha_{nil,1} > 0$ and $\alpha_{nil,2} > 0$ for all i and $n \geq n_1$. From (4.3.31), (4.3.32) and (4.3.33),

$$(4.3.35) \qquad P\left\{\alpha_{nil,2} \leq \alpha^*_{nil} \leq \alpha_{nil,1},\ i = 1,\ldots, n\right\} \geq 1 - 3\epsilon_2$$

for $n \geq n_1$. This can be rewritten as

$$(4.3.36) \qquad P\left\{F_{n1,2}(x) \leq F^*_{n1}(x) \leq F_{n1,1}(x) \text{ for all } x\right\} \geq 1 - 3\epsilon_2$$

using the monotonicity of $I(z; x)$ in x. Appealing now to Theorem 4.3.4 we have

$$\left\{ \sqrt{n}\,[F_{n1,1}(x) - F^+_{n1,1}(x)], \quad x \geq 0 \right\}$$

and

$$\left\{ \sqrt{n}\,[F_{n1,2}(x) - F^+_{n1,2}(x)], \quad x \geq 0 \right\}$$

with $F^+_{n1,j}(x)$ as defined in (4.3.5) with perturbation constants $\{\alpha_{ni1,j}\}$, $j = 1,2$, converge weakly to Gaussian processes $\{\eta_1(x), x \geq 0\}$ in $D[0,\infty]$ with mean zero and covariance function given in (4.3.14). The functions $F^+_{n1,j}(x)$ differ from $F^+_{n1}(x)$ by constant times ϵ_1, uniformly in x. Since ϵ_1 and ϵ_2 are arbitrary, these assertions, along with (4.3.36) establish Theorem 4.3.6.

Remark: The remark after Theorem 4.3.4 is applicable to the above Theorem.

Now suppose that $\{\alpha^*_{ni2}, i = 1,\ldots,n\}$, $n = 1,2,\ldots$ is a triangular scheme of r.v.'s with the form

$$(4.3.37) \qquad \alpha^*_{ni2} = \theta^*_{ni}\left[1 + \beta^*_{ni}/n^{\partial} + \gamma^*_{ni}/n^{2\partial} + R^*_{ni}\right]$$

where R^*_{ni}, β^*_{ni} and γ^*_{ni} satisfy the assumptions (4.3.33),

(4.3.31) and (4.3.32) and further there is a function $\theta(p)$ on $[0,1]$ such that

$$(4.3.38) \qquad \sup_i \sqrt{n} \mid \theta^*_{ni} - \theta(i/n) \mid = o_p(1)$$

and $\theta(p)$ satisfies condition (C). Now define $F^*_{n2}(x)$, $F^+_{n2}(x)$ and $\eta^*_{n2}(x)$ similar to the expressions in (4.3.28), (4.3.29) and (4.3.30) respectively with $\{ \alpha^*_{ni2} \}$ instead of $\{\alpha^*_{ni1}\}$. The following theorem then follows from Theorem 4.3.5 in exactly the same way as Theorem 4.3.6 follows from Theorem 4.3.4.

Theorem 4.3.7: Let conditions (B) and (C) hold. The sequence of stochastic processes $\{ \eta^*_{n2}(x), x \geq 0 \}$ converges weakly to the Gaussian process $\{ \eta_2(x), x \geq 0 \}$ in $D[0,\infty]$ where $\{\eta_2(x), x \geq 0\}$ has mean zero and covariance function $K_2(x,y)$ given in (4.3.23).

Remark: The remark after Theorem 4.3.5 is applicable to the above Theorem.

4.4 Asymptotic distribution of the empirical d.f. when the random variables are subject to perturbations and a random scale factor

We retain the notations of the earlier sections. Let $\eta_{n1}(x)$ be defined as in (4.3.6) through $F_{n1}(x)$ and $F^{+}_{n1}(x)$ which are in turn defined as in (4.3.1) and (4.3.5) and the α_{nil}'s have the structure (4.3.2).

Let Z^{*}_{n} be a r.v. and let $\xi_n = \sqrt{n}\,(Z^{*}_{n} - 1)$. We now assume the following condition (D) on the stochastic process $\{\, \eta_{n1}(x),\ x \geq 0 \,\}$ and ξ_n:

(D) For any finite collection $(x_1,\ldots,x_k)$, the distribution of $\{\, \eta_{n1}(x_1), \eta_{n1}(x_2),\ldots, \eta_{n1}(x_k), \xi_n \,\}$ converges weakly to the distribution of $\{\, \eta_1(x_1),\ldots, \eta_1(x_k), \xi \,\}$, which is multivariate Normal with zero means and covariances given by

$$(4.4.1) \qquad \operatorname{cov}\,(\eta_1(x_i), \eta_1(x_j)) = K_1(x_i, x_j), \quad 1 \leq i, j \leq k$$

where $K_1(x,y)$ is as defined in (4.3.4) and

$$(4.4.2) \qquad \operatorname{cov}\,(\eta_1(x_i), \xi) = a_1(x_i), \quad i = 1,\ldots, k$$

and

$$(4.4.3) \qquad \operatorname{var}\,(\xi) = 1.$$

(B*) We add the following to the assumption (B) made on $F(x)$ in Section 4.3. (B*) There is an $\alpha > 0$ such that $x^{\alpha}(1 - F(x)) \to 0$ and $xf(x) \to 0$ as $x \to \infty$.

Theorem 4.4.1: Let the conditions (B), (B*) and (D) hold. Let

$$\zeta_{n1}(x) = \sqrt{n}[F_{n1}(xZ_n^*) - F_{n1}^+(x)]. \tag{4.4.4}$$

Then

$$\sup_{0 \le x \le \infty} | \zeta_{n1}(x) - \eta_{n1}(x) - \xi_n x f(x) | = o_p(1). \tag{4.4.5}$$

Thus $\{\zeta_{n1}(x), x \ge 0\}$ converges weakly to the Gaussian process $\{\zeta_1(x) = \eta_1(x) + xf(x)\xi, x \ge 0\}$ in $D[0,\infty]$. $\{\zeta_1(x), x \ge 0\}$ has mean zero and covariance function

$$K_3(x,y) = K_3(y,x) = K_1(x,y) + xyf(x)f(y) + xf(x)a_1(y) + yf(y)a_1(x) \quad \text{for} \quad x \le y. \tag{4.4.6}$$

Proof:

$$\begin{aligned}
\zeta_{n1}(x) &= \sqrt{n}[F_{n1}(xZ_n^*) - F_{n1}^+(x)] \\
&= \sqrt{n}\,[F_{n1}(x) - F_{n1}^+(x)] + \sqrt{n}\,[F_{n1}(xZ_n^*) - F_{n1}(x)] \\
&= \eta_{n1}(x) + \sqrt{n}(Z_n^* - 1) \sum_{i=1}^{n} x\alpha_{ni1}\, f(x\alpha_{ni1})/n \\
&\quad + \sqrt{n}\,[F_{n1}(xZ_n^*) - F_{n1}(x) - (Z_n^* - 1) \sum_i x\alpha_{ni1} f(x\alpha_{ni1})/n] \qquad (4.4.7) \\
&= \eta_{n1}(x) + \xi_n \sum_1^n x\alpha_{ni1} f(x\alpha_{ni1})/n + R_n(x) \quad \text{(say)}.
\end{aligned}$$

Since $\alpha_{ni1} \to 1$ uniformly in i,

$$\sum_{i=1}^{n} x\alpha_{ni1} f(x\alpha_{ni1})/n \to xf(x)$$

uniformly in x. Thus to prove (4.4.5), it is sufficient to prove that

$$(4.4.8) \qquad \sup_{0 \le x \le \infty} |R_n(x)| = o_p(1).$$

The proof of (4.4.8) is completed in (4.4.31). The main steps are (4.4.9), (4.4.24) and (4.4.30). Let

$$R_n(x,c) = \sum_1^n [I(Z_i/c\alpha_{nil};\ x) - I(Z_i/\alpha_{nil};\ x) - (c-1)x\alpha_{nil} f(x\alpha_{nil})] / \sqrt{n} .$$

Then

$$R_n(x,\ Z_n^*) = R_n(x).$$

Because of condition (D), for any given $\phi > 0$, we can find an L such that

$$(4.4.9) \qquad P\{ |Z_n^* - 1| \ge L / \sqrt{n} \} \le \phi .$$

We first obtain a bound for

$$(4.4.10) \qquad P\{ \sup_{0 \le x \le K} \sup_{|c-1| \le L/\sqrt{n}} |R_n(x,\ c)| > \omega \}$$

This involves the following standard but lengthy method.

Let $\epsilon, \theta > 0$. We will choose these constants suitably later on. The interval $[1 - L/\sqrt{n},\ 1 + L/\sqrt{n}\ ]$ is covered by

$L_n = [2L/\epsilon] + 1$ small intervals of length $\epsilon_n = \epsilon/\sqrt{n}$ each, the r^{th} interval being

$$(4.4.11) \qquad L_r^* = \{ c: c_r \le c < c_{r+1} \}$$

where $c_r = 1 - L/\sqrt{n} + r\epsilon_n$. Similarly, the interval $[0, K]$ is covered by $k_n = [\sqrt{n}\, K/\theta] + 1$ intervals of length $\theta_n = \theta/\sqrt{n}$ each, the s^{th} interval being

$$(4.4.12) \qquad K_s^* = \{ x: x_s \le x < x_{s+1} \}$$

where $x_s = s\theta_n$.

Let

$$m_{si} = \inf_{x \in K_s^*} x\alpha_{nil}\, f(x\alpha_{nil})$$

$$M_{si} = \sup_{x \in K_s^*} x\alpha_{nil}\, f(x\alpha_{nil}).$$

Fix an x. Recall that $1/2 \le \alpha_{nil} \le 2$ for all n and i. For $c \in L_r^*$,

$$R_n(x, c_r) - \epsilon \sum_1^n x\alpha_{nil}\, f(x\alpha_{nil})/n \le R_n(x, c)$$

$$\le R_n(x, c_{r+1}) + \epsilon \sum_1^n x\alpha_{nil}\, f(x\alpha_{nil})/n$$

due to the monotonicity of $I(z; x)$ in z. Therefore

$$R_n^*(x,L) = \sup_{|c-1| \le L/\sqrt{n}} |R_n(x,c)|$$

(4.4.13)

$$\le \max_{0 \le r \le L_n} |R_n(x,c_r)| + \epsilon \sum_1^n x\alpha_{nil}\, f(x\alpha_{nil})/n.$$

Next, for $x \in K_s^*$

$$\sum_1^n [I(Z_i/c_r\alpha_{nil};\, x_s) - I(Z_i/\alpha_{nil};\, x_{s+1}) - (c_r-1)\lambda_{sir}]/\sqrt{n}$$

(4.4.14)

$$\le R_n(x,\, c_r)$$

$$\le \sum_1^n [I(Z_i/c_r\alpha_{nil};\, x_{s+1}) - I(Z_i/\alpha_{nil};\, x_s) - (c_r-1)\lambda_{sir}]/\sqrt{n}$$

where λ_{sir} will be used generatively and will be m_{si} or M_{si} depending on whether $(c_r - 1)$ is negative or not and whether it enters in the expression for a lower bound or an upper bound. Thus

(4.4.15)

$$\sup_{0 \le x \le K} \sup_{|c-1| \le L/\sqrt{n}} |R_n(x,c)| = \sup_{0 \le x \le K} |R_n^*(x,L)|$$

$$\le \max_{0 \le s \le K_n} \max_{0 \le r \le L_n} R_n^{**}(s,r)$$

where

$$R_n^{**}(s,r) = \sum_1^n [I(Z_i/c_r\alpha_{nil};x_s) - I(Z_i/\alpha_{nil};x_{s\pm1})$$

(4.4.16)

$$-(c_r - 1)\,\lambda_{sir}]/\sqrt{n} + \epsilon \sum_1^n M_{si}/\sqrt{n}\,.$$

In (4.4.15), we omit any term in which the suffix of x becomes negative or larger than K_n. Let r and s be fixed. Then

$$R_n^{**}(s,r) \leq \Big| \sum_i \big[I(Z_i/c_r\alpha_{nil};\ x_s) - F(x_s c_r \alpha_{nil}) - I(Z_i/\alpha_{nil};\ x_{s\pm 1}) + F(x_{s\pm 1}\alpha_{nil}) \big] \Big| / \sqrt{n}$$

$$+ \Big| \sum_i \big[F(x_s c_r \alpha_{nil}) - F(x_{s\pm 1}\alpha_{nil}) - (c_r - 1)\lambda_{sir} \big] \Big| / \sqrt{n}$$

$$+ \epsilon \sum_i M_{si} / n$$

$$= \Big| \sum_i \big[I(Z_i/c_r\alpha_{nil};\ x_s) - F(x_s c_r \alpha_{nil}) - I(Z_i/\alpha_{nil};\ x_{s\pm 1}) + F(x_{s\pm 1}\alpha_{nil}) \big] \Big| / \sqrt{n}$$

$$+ r_n \quad \text{(say)} \tag{4.4.17}$$

where r_n is non-random. Given any $\omega > 0$, we can clearly choose ϵ and θ such that

$$r_n \leq \omega/2 \tag{4.4.18}$$

using condition (B). Thus from (4.4.17) and (4.4.18),

$$P\{ R_n^{**}(s,r) > \omega \} \leq P\Big\{ \Big| \sum_i \big[I(Z_i/c_r\alpha_{nil};x_s) - F(x_s c_r \alpha_{nil}) - I(Z_i/\alpha_{nil};\ x_{s\pm 1}) + F(x_{s\pm 1}\alpha_{nil}) \big] \Big| / \sqrt{n} > \omega/2 \Big\} . \tag{4.4.19}$$

We now obtain a bound for the probability on the RHS of (4.4.19). Similar bounds are known (Ref. e.g. Bahadur (1960a)), but none

suited to our situation is readily available in the literature. Let

$$(4.4.20)\quad W_{ni} = [I(Z_i/c_r\alpha_{nil};\ x_s) - F(x_s c_r \alpha_{nil}) - I(Z_i/\alpha_{nil};\ x_{s\pm 1}) + F(x_{s\pm 1}\alpha_{nil})].$$

If

$$(4.4.21)\quad p_{ni} = |F(x_s c_r \alpha_{nil} - F(x_{s\pm 1}\alpha_{nil})|$$

and B_{ni} is a binomial random variable such that

$$P(B_{ni} = 1) = p_{ni}$$

$$P(B_{ni} = 0) = (1 - p_{ni}) = q_{ni}$$

then W_{ni} of (4.4.20) may be written as

$$W_{ni} = \begin{cases} B_{ni} - p_{ni} & \text{if } x_s c_r > x_{s\pm 1} \\ -B_{ni} + p_{ni} & \text{if } x_s c_r < x_{s\pm 1}. \end{cases}$$

In either case, the required probability is

$$P\left\{\left|\sum_i W_{ni}\right|/\sqrt{n} > \omega/2\right\} = P\left\{\left|\sum_i (B_{ni} - p_{ni})\right| > \sqrt{n}\,\omega/2\right\}$$

$$= P\left\{\sum_i (B_{ni} - p_{ni}) > \sqrt{n}\,\omega/2\right\}$$

$$(4.4.22)\qquad + P\left\{\sum_i (B_{ni} - p_{ni}) < -\sqrt{n}\,\omega/2\right\}.$$

Now for any $t > 0$, in the region in which the moment generating function of W_{ni} exists,

$$P\left\{\sum_i (B_{ni} - p_{ni}) > \sqrt{n}\ \omega/2\right\} = P\left\{e^{t\sum_i (B_{ni}-p_{ni})} > e^{t\sqrt{n}\ \omega/2}\right\}$$

$$\leq e^{-t\sqrt{n}\ \omega/2} \prod_{i=1}^{n} E[e^{t(B_{ni} - p_{ni})}]$$

$$= a(n, t, \omega), \text{ say.}$$

The moment generating function of the binomial random variable B_{ni} is

$$E(e^{tB_{ni}}) = (q_{ni} + p_{ni} e^t) = (1 - p_{ni}(1 - e^t)).$$

Therefore

$$\frac{1}{n} \log a(n,t,\omega) = -t\omega/2\sqrt{n} - t \sum_i p_{ni}/n + \frac{1}{n} \sum_i \log [1 - p_{ni}(1 - e^t)]$$

$$= -t\omega/2\sqrt{n} - t \sum_i p_{ni}/n - (1 - e^t) \sum_i p_{ni}/n - (1 - e^t)^2 \sum_i p_{ni}^2/2n - \cdots .$$

It is easy to see from the definition (4.4.21) of p_{ni} that $\sqrt{n}\ p_{ni}$ is bounded uniformly in i. Thus if we write $\sum_i p_{ni}/n$ as $\bar{p}_n$ then $\sqrt{n}\ \bar{p}_n \to c_1$, a positive constant. Hence

$$\frac{1}{n} \log a(n,t,\omega) = -\frac{1}{\sqrt{n}} (t\omega/2 + t\sqrt{n}\ \bar{p}_n + (1 - e^t)\sqrt{n}\ \bar{p}_n) + o(1)$$

$$\sim -\frac{1}{\sqrt{n}} (t\omega/2 + c_1 t + (1 - e^t)c_1) + o(1)$$

$$= -\frac{c}{\sqrt{n}} + o(1)$$

for a suitable choice of t, which makes $c > 0$. For example, it is sufficient to choose $t = o(\omega/2c_1)$. One can proceed on exactly similar lines and obtain a similar bound for

$$P\left\{\sum_i (B_{ni} - p_{ni}) < -\sqrt{n}\,\omega/2\right\}.$$

Hence the required probability

$$P\left\{\left|\sum_i [\, I(Z_i/c_r\alpha_{nil};\, x_s) - F(x_s c_r \alpha_{nil}) - I(Z_i/\alpha_{nil};\, x_{s\pm 1})\right.\right.$$

$$\left.\left. + F(x_{s\pm 1}\,\alpha_{nil})\,]\right|/\sqrt{n} > \omega/2\right\}$$

$$= P\left\{\left|\sum_i^n W_{ni}\right|/\sqrt{n} > \omega/2\right\}$$

$$= \exp[-\text{constant}\ \sqrt{n}\,]. \qquad (4.4.23)$$

Thus, combining (4.4.15), (4.4.19) and (4.4.23),

$$P\left\{\sup_{0\le x\le K} |R_n(x)| > \omega\ ,\ |Z_n^* - 1| \le L/\sqrt{n}\right\}$$

$$\le P\left\{\sup_{0\le x\le K}\ \sup_{|c-1|\le L/\sqrt{n}} |R_n(x,c)| > \omega\right\}$$

$$\le ([\sqrt{n}\,K/\theta]+1)([2L/\theta]+1)\, e^{-\text{const.}\sqrt{n}} \qquad (4.4.24)$$

$$\le \phi$$

for n sufficiently large. Now, consider, on the other hand

$$P\left\{\sup_{x\geq K}|R_n(x)| > \omega,\ |Z_n^*-1| \leq L/\sqrt{n}\right\}$$

$$\leq P\left\{\sqrt{n}(1-F_{n1}(K)) + L\sup_{x\geq K/2}(xf(x)) > \omega\right\}$$

$$= P\left\{\sqrt{n}\,[F_{n1}^+(K) - F_{n1}(K)] > \omega - L\sup_{x\geq K/2}(xf(x)) - \sqrt{n}\,[1-F_{n1}^+(K)]\right\}$$

$$\leq P\left\{\sqrt{n}\,[F_{n1}^+(K) - F_{n1}(K)] > \omega - L\sup_{x\geq K/2}(xf(x)) - \sqrt{n}\,[1-F(K/2)]\right\} \tag{4.4.25}$$

since we can assume that $\alpha_{ni1} > 1/2$ for all n and i. Let us choose a K, which will depend on n, such that

$$\sup_{x\geq K/2}(xf(x)) \leq \omega/4 \tag{4.4.26}$$

and

$$\sqrt{n}\,[1-F(K/2)] \leq \omega/4. \tag{4.4.27}$$

Clearly (4.4.26) is satisfied if K is large enough because of the condition (B*). We now show that K can be chosen to satisfy condition (4.4.27) also. Since the rate of decrease condition (B*) on $F(x)$ gives

$$x^{\alpha}\,[1-F(x)] \to 0 \quad \text{as} \quad x \to \infty,$$

we can choose an x_0 such that for $x \geq x_0$

$$x^{\alpha}\,[1-F(x)] \leq \omega/4. \tag{4.4.28}$$

When $n \geq x_0^{2\alpha}$, putting $x^{\alpha} = \sqrt{n}$, from (4.4.28) we obtain

$$\sqrt{n}\,[1 - F(n^{1/2\alpha})] \leq \omega/4.$$

Thus if we choose

$$K \geq 2n^{1/2\alpha} \tag{4.4.29}$$

it satisfies (4.4.26) and (4.4.27) for sufficiently large n. Further since $\sqrt{n}\,[F_{n1}(K) - F^{+}_{n1}(K)]$ is asymptotically normal with mean zero and variance $F(K)(1-F(K))$, (see Theorem 4.3.4) using (4.4.25), (4.4.26) and (4.4.27)

$$P\left\{ \sup_{x \geq K} |R_n(x)| > \omega\,,\ |Z_n^* - 1| \leq L/\sqrt{n} \right\}$$

$$\leq P\left\{ \sqrt{n}\,[F^{+}_{n1}(K) - F_{n1}(K)] > \omega/2 \right\}$$

$$= 1 - \Phi\,(\omega/2\,F(K)(1-F(K)))$$

$$< \emptyset. \tag{4.4.30}$$

by choosing n sufficiently large. Our conditions assure us of a K which satisfies (4.4.26) and (4.4.27) and at the same time the probability in (4.4.30) can be made arbitrarily small.

Now combining (4.4.9), (4.4.24) and (4.4.30), we have

$$P\left\{ \sup_{0 \leq x \leq \infty} |R_n(x)| > \omega \right\}$$

$$\leq P\left\{ |Z_n^* - 1| > L/\sqrt{n} \right\}$$

$$+ P\left\{ \sup_{0 \leq x \leq K} |R_n(x)| > \omega\,,\ |Z_n^* - 1| \leq L/\sqrt{n} \right\}$$

$$+ P\left\{ \sup_{x \geq K} |R_n(x)| > \omega\,,\ |Z_n^* - 1| \leq L/\sqrt{n} \right\} \tag{4.4.31}$$

$$\leq \emptyset + \emptyset + \emptyset = 3\,\emptyset$$

for sufficiently large n. This completes the proof of Theorem 4.4.1.

We now extend this result to the more general non-random perturbation factors $\{ \alpha_{ni2} \}$ defined in (4.3.22). Let $F_{n2}(x)$, $F^{+}_{n2}(x)$ and $\eta_{n2}(x)$ be as defined and used in Theorem 4.3.5. If Z^{*}_{n} be a r.v., we assume that $\zeta_n = \sqrt{n}\,(Z^{*}_{n}-1)$ satisfies the following condition (D*) with the process $\{ \eta_{n2}(x),\ x \geq 0 \}$.

For any finite collection $(x_1,\ldots,x_k)$, the distribution of $\{ \eta_{n2}(x_1),\ldots,\eta_{n2}(x_k),\ \zeta_n \}$ converges weakly to that of $\{\eta_2(x_1),\ldots,\eta_2(x_k),\ \xi\}$ which is a multivariate Normal distribution with zero means and covariances given by

$$(4.4.32) \qquad \operatorname{cov}\,(\eta_2(x_i),\ \eta_2(x_j)) = K_2(x_i,x_j),\quad 1 \leq i,\ j \leq k$$

where $K_2(x,y)$ is as defined in (4.3.23) and

$$(4.4.33) \qquad \operatorname{cov}\,(\eta_2(x_i),\ \xi) = -a_2(x_i),\quad i = 1,\ldots,\ k$$

and

$$\operatorname{var}\,(\xi) = 1.$$

Then we have the following theorem whose proof follows on the lines of the proof of Theorem 4.4.1 and is omitted.

<u>Theorem 4.4.2.</u> Let the conditions (B), (B*), (C) and (D*) hold. Let

$$(4.4.34) \qquad \zeta_{n2}(x) = \sqrt{n}\,[F_{n2}(xZ^{*}_{n}) - F^{+}_{n2}(x)].$$

$$(4.4.35)\qquad \sup_{0\le x\le\infty} \left| \zeta_{n2}(x) - \eta_{n2}(x) - \xi_n x\left(\int_0^1 \theta(p) f(x\theta(p))dp\right)\right| = o_p(1).$$

Thus $\{\zeta_{n2}(x),\ x \ge 0\}$ converges weakly in $D[0,\infty]$ to the Gaussian process $\{\zeta_2(x) = \eta_2(x) + \xi x \left(\int_0^1 \theta(p) f(x\theta(p))dp\right),\ x \ge 0\}$ which has mean zero and covariance function

$$K_4(x,y) = K_4(y,x) = K_2(x,y) + xy\left(\int_0^1 \theta(p) f(x\theta(p))dp\right)\left(\int_0^1 \theta(p) f(y\theta(p))dp\right)$$
$$+ x a_2(y)\left(\int_0^1 \theta(p) f(x\theta(p))dp\right) + y a_2(x)\left(\int_0^1 \theta(p) f(y\theta(p))dp\right)$$

$$(4.4.36)\qquad \text{for } x \le y$$

with $K_2(x,y)$ as in (4.3.23).

Now coming to the case of random perturbation factors, let $\{\alpha^*_{nil}\}$ be as in (4.3.24) and $F^*_{n1}(x)$, $F^+_{n1}(x)$ and $\eta^*_{n1}(x)$ be as defined and used in Theorem 4.3.6. Let $\{\alpha_{nil}\}$, the non-random constants generated from $\{\alpha^*_{nil}\}$ be as in (4.3.29). Let Z^*_n, ξ_n be as used in Theorem 4.4.1 and satisfy the condition (D) with the process $\{\eta^*_{n1}(x),\ x \ge 0\}$. Then we have the following extension of Theorem 4.4.1 to the case of random perturbations.

Theorem 4.4.3: Let the conditions (B), (B*) and (D) hold. Let

$$(4.4.37)\qquad \zeta^*_{n1}(x) = \sqrt{n}\,[F^*_{n1}(xZ^*_n) - F^+_{n1}(x)]$$

Then

$$(4.4.38)\qquad \sup_{0\le x\le\infty} \left|\zeta^*_{n1}(x) - \eta^*_{n1}(x) - \xi_n x\, f(x)\right| = o_p(1).$$

Thus $\{ \zeta^*_{n1}(x),\ x \geq 0 \}$ converges weakly to the Gaussian process $\{ \zeta_1(x) = \eta_1(x) + \xi x f(x),\ x \geq 0 \}$ defined in Theorem 4.4.1.

We will omit the proof since it follows from Theorem 4.4.1 in almost the same manner as Theorem 4.3.6 was deduced from Theorem 4.3.4.

Finally let $\{ \alpha^*_{ni2} \}$ be the more general random perturbation factors defined in (4.3.37) and satisfy the conditions stipulated there. Let $\{ \alpha_{ni2} \}$, the non-random constants generated by $\{ \alpha^*_{ni2} \}$, be as defined in (4.3.22). Let $F^*_{n2}(x)$, $F^+_{n2}(x)$ and $\eta^*_{n2}(x)$ be as defined and used in Theorem 4.3.7. Let Z^*_n, ξ_n be as used in Theorem 4.4.2 and satisfy the condition (D*) with the process $\{ \eta^*_{n2}(x),\ x \geq 0 \}$. Then the following theorem can be deduced from Theorem 4.4.2 in the usual way.

Theorem 4.4.4: Let the conditions (B), (B*), (C) and (D*) hold. Let

$$\zeta^*_{n2}(x) = \sqrt{n}\,[F^*_{n2}(xZ^*_n) - F^+_{n2}(x)]. \tag{4.4.39}$$

Then

$$\sup_{0 \leq x \leq \infty} \left| \zeta^*_{n2}(x) - \eta^*_{n2}(x) - \xi_n x \left(\int_0^1 \theta(p) f(x\theta(p))\,dp \right) \right| = o_p(1). \tag{4.4.40}$$

Thus the process $\{ \zeta^*_{n2}(x),\ x \geq 0 \}$ converges weakly in $D[0,\infty]$ to the Gaussian process $\{ \zeta_2(x),\ x \geq 0 \}$ defined in Theorem 4.4.2 with mean zero and covariance function given by (4.4.36).

4.5 Asymptotic distributions of the empirical d.f.'s of normalised and modified spacings and tests based on them

In this section, we relate the results of the last two sections to the spacings statistics. First we give the asymptotic distributions of the empirical d.f.'s of the normalised spacings $H_n(x)$ and of the modified spacings $H_n^*(x)$, defined in (4.2.21) and (4.2.26) respectively using the distributional equivalences (4.2.23) and (4.2.27). We then establish the asymptotic normality of some classes of test statistics based on these spacings.

We shall first consider the empirical d.f. of the normalised spacings $H_n(x)$, which from (4.2.23) is distributionally equivalent to $F_n^*(x\,\bar{W}_n)$. The r.v.'s $W_1, W_2, \ldots$ have the exponential d.f.

$$F(x) = 1 - e^{-x}, \quad x \geq 0 \tag{4.5.1}$$

which satisfies all the regularity conditions of Theorem 4.4.1 and the assumptions (B) and (B*). Further the $\{ \alpha_{nil}^* \}$ used in the definition (4.3.28) of $F_{nl}^*(x)$ satisfy the conditions (4.3.25), (4.3.26) and (4.3.27) with $\beta(p)$ and $\gamma(p)$ given by (4.2.17) and (4.2.18). Hence we have from the definition of $B(p)$ and $\gamma(p)$,

$$(4.5.2) \qquad \int_0^1 \beta(p)dp = \int_0^1 \lambda(p)dp = 0$$

$$\int_0^1 \gamma(p)dp = -\int_0^1 L(p)\lambda'(p)dp = \int_0^1 \lambda^2(p)dp = \int_0^1 \beta^2(p)dp.$$

Let

$$(4.5.3) \qquad \xi^*_{n1}(x) = \sqrt{n}\,[H_n(x) - F^+_{n1}(x)]$$

$$\sim \sqrt{n}\,[F^*_{n1}(x\bar{W}_n) - F^+_{n1}(x)]$$

where

$$(4.5.4) \qquad F^+_{n1}(x) = \begin{cases} (1-e^{-x}) & \text{for} \quad \delta > 1/4 \\ (1-e^{-x}) + (\int_0^1 \lambda^2(p)dp)e^{-x}(x - x^2/2)/\sqrt{n} & \\ & \text{for} \quad \delta = 1/4, \end{cases}$$

ignoring terms, which are of smaller order than $n^{-1/2}$ uniformly in x. Further since the random scale factor here is $\bar{W}_n$, condition (D) is satisfied and $a_1(x)$, defined in (4.4.2) is easily seen to be $(-xe^{-x})$. In view of the these remarks we have the following theorem as a consequence of Theorem 4.4.3.

Theorem 4.5.1: The sequence of stochastic processes $\{\xi^*_{n1}(x),\ x \geq 0\}$ converges weakly to a Gaussian process $\{\xi_1(x),\ x \geq 0\}$ with mean zero and covariance function

$$(4.5.5) \qquad K_3(x,y) = e^{-y}(1 - e^{-x} - xy\,e^{-x}), \quad x \leq y.$$

This theorem together with Theorem 4.5.4 on the empirical d.f. of the modified spacings, form the basic results for deriving the asymptotic distributions and finding the efficiencies

of the spacings tests which we do in this and the next section. From the invariance principle, in view of Theorem 4.5.1, we have

Theorem 4.5.2: Let $g(\cdot)$ be a real valued function on $D[0,\infty]$ which is almost everywhere continuous with respect to the probability measure induced by the Gaussian process $\{ \zeta_1(x),\ x \geq 0 \}$ of the previous theorem. Then the distribution of the real valued r.v. $g(\zeta^*_{n1}(x))$ converges to the distribution of $g(\zeta_1(x))$ as $n \to \infty$.

As an application, we have the following

Theorem 4.5.3: Let m be an absolutely continuous function on $[0,\infty]$ with $m(0) < \infty$. Let $m'(x)$ be bounded on every finite interval and let

$$T_n = \sum_{i=1}^{n} m(nD_i)/n \tag{4.5.6}$$

be a statistic based on spacings from the d.f. $G_n(x)$. Then

$$\sqrt{n}\,[T_n - \int_0^\infty m'(x)(1 - F^+_{n1}(x))dx + m(0)] \tag{4.5.7}$$

has an asymptotic normal distribution with mean zero and variance

$$\int_0^\infty \int_0^\infty m'(x)m'(y)K_3(x,y)\,dx\,dy < \infty \tag{4.5.8}$$

if the function $g(\cdot)$ on $D[0,\infty]$ defined by

$$(4.5.9) \qquad g(y(x)) = \int_0^\infty m'(x)y(x)dx,\ y(\cdot)\in D[0,\infty]$$

is continuous almost everywhere with respect to the probability measure induced by the Gaussian process $\{\zeta_1(x),\ x \geq 0\}$. (This condition will be referred to as Assumption (E) later on.)

Proof:
$$T_n = \sum_1^n m(nD_i)/n$$
$$= \int_0^\infty m(x)dH_n(x)$$
$$= -\int_0^\infty m(x)d[1 - H_n(x)]$$
$$= \int_0^\infty m'(x)[1 - H_n(x)]dx - m(0)$$

on integration by parts. Now from (4.5.3), we get

$$(4.5.10) \qquad T_n = \int_0^\infty m'(x)[1 - F_{n1}^+(x)]dx - (1/\sqrt{n})\int_0^\infty m'(x)\,\zeta_{n1}^*(x)dx - m(0).$$

Thus the quantity defined in (4.5.7) is equal to

$$(4.5.11) \qquad \int_0^\infty m'(x)\,\zeta_{n1}^*(x)dx .$$

The result follows immediately in view of Theorems 4.5.1 and 4.5.2 and because of the assumption (E).

Remark 1: From the standard law of the iterated logarithm for the Weiner process, we can deduce that

$$P\left\{\overline{\lim_{x\to\infty}} \frac{\zeta_1(x)}{\sqrt{2(1-e^{-x})\log\log(1-e^{-x})}} = 1\right\} = 1$$

and

$$P\left\{ \overline{\lim_{x\to\infty}} \frac{\zeta_1(x)}{\sqrt{2e^{-x}\log x}} = 1 \right\} = 1.$$

Thus if $m(x)$ is an absolutely continuous function and $m'(x)$ is bounded in every interval of the form $[\epsilon, K]$ with $0 < \epsilon < K < \infty$ and satisfies

$$\int_0 m'(x) \sqrt{(1-e^{-x}) \log\log (1-e^{-x})^{-1}}\, dx < \infty$$

and

$$\int^{\infty} m'(x) \sqrt{e^{-x} \log x}\, dx < \infty$$

then the assumption (E) of Theorem 4.5.3 is satisfied. Examples of such $m(x)$ are

$$m(x) = x^r, \quad r > -1/2$$

or

$$m(x) = \tfrac{1}{2} |x-1|.$$

The spacings tests corresponding to these two classes have been studied by Greenwood (1946), Kimball (1950) and Sherman (1950).

Remark 2: The function $m(x) = \log x$ (c.f. Darling (1953)) does not satisfy condition (E). However, $\int_0^{\infty} \log x\, d\zeta^*_{n1}(x)$ and $\int_0^{\infty} \log x\, d\zeta_1(x)$ may be considered as stochastic integrals, existence of which can be proved using the tied-down Weiner process on $[0,1]$ after transforming $[0, \infty]$ into $[0,1]$ by the transformation $x \to (1-e^{-x})$. Thus it can be shown that Darling's statistic,

$$L(n) = \frac{1}{n} \sum_{i=1}^{n} \log (n D_i)$$

has an asymptotic normal distribution both under the hypothesis as well as under the alternatives (4.2.1).

Remark 3: We assume the condition (E) to hold. Under the alternatives given in (4.2.1), the statistic T_n has an asymptotic normal distribution with mean, say

$$\mu_{1n} = \int_0^\infty m'(x)(1 - F^+_{n1}(x))\,dx$$

$$= \begin{cases} \int_0^\infty m'(x)e^{-x}\,dx & \text{if } \partial > 1/4 \\ \int_0^\infty m'(x)e^{-x}\,dx + (1/\sqrt{n})\left(\int_0^1 \ell^2(p)\,dp\right) \cdot \int_0^\infty m'(x)e^{-x}\left(\frac{x^2}{2} - x\right)dx & \text{if } \partial = 1/4. \end{cases}$$

and with variance

$$(4.5.13) \qquad \sigma^2_{1n} = \left(\frac{1}{n}\right)^2 \int_0^\infty \int_x^\infty m'(x)m'(y)e^{-y}(1 - e^{-x} - xye^{-x})\,dx\,dy.$$

Under the hypothesis, the statistic T_n has an asymptotic normal distribution with mean, say

$$(4.5.14) \qquad \mu_{on} = \int_0^\infty m'(x)\, e^{-x}dx$$

and variance

$$(4.5.15) \qquad \sigma^2_{on} = \sigma^2_{1n}.$$

The distribution under the hypothesis follows from Theorem 4.5.3 by putting $\ell(x) \equiv 0$.

Comparing the distributions of T_n under the null hypothesis and under the alternatives $G_n(x)$ of (4.2.1), we observe that the null distribution coincides with the distribution under the alternative if the alternatives are such that $\delta > 1/4$. In other words, the test statistic T_n is incapable of discriminating the uniform density from the densities of the form $(1+\ell(x)/n^{\delta})$ if $\delta > 1/4$. But when we consider the sequence of alternatives

$$G_n(x) = x + L(x)/n^{1/4}, \quad 0 \leq x \leq 1 \tag{4.5.16}$$

the means μ_{0n} and μ_{1n} of T_n differ, though the variances agree up to order $1/n$. From this we conclude that it is only the alternatives converging to uniformity at the rate of $n^{-1/4}$ that give us an idea about the power behaviour of these symmetric spacings tests. Hence we should concentrate on the alternatives (4.5.16) for making efficiency comparisons of these tests.

We now consider each of these spacings statistics in turn and give their means and variances under the hypothesis as well as under the alternatives (4.5.16). The Pitman efficiencies that we compute in Section 4.6 depend on these

expressions. Consider, first, the statistic

$$(4.5.17) \qquad V_r(n) = \sum_{i=1}^{n} (nD_i)^r/n, \quad r > -1/2$$

due to Kimball (1950) of which Greenwood's statistic is a particular case for $r = 2$. We have the asymptotic normality of this statistic from our theorems both under the hypothesis as well as under the alternatives (4.5.16) with means

$$(4.5.18) \qquad \mu_{on}(V_r) = \overline{|r+1}$$

$$(4.5.19) \qquad \mu_{1n}(V_r) = \overline{|r+1} + r(r-1)(\overline{|r+1})(\int_0^1 \phi^2(p)\,dp)/2\sqrt{n}$$

and variance

$$(4.5.20) \qquad \sigma^2_{on}(V_r) = \sigma^2_{1n}(V_r) = \{\overline{|2r+1} - (r^2+1)(\overline{|r+1})^2\}/n.$$

The statistic

$$(4.5.21) \qquad U(n) = 1/2 \sum_{i=1}^{n} |D_i - 1/n|$$

was suggested by Kendall (c.f. Greenwood (1946) discussion) and its asymptotic normality under the null hypothesis was shown, by using the method of moments, by Sherman (1950). We have discussed this statistic in some detail in the earlier chapter. We have

$$U(n) = \Sigma\,(1/n - D_i)$$

$$\{i = D_i < 1/n\}$$

$$= \int_0^1 (1 - x)\,dH_n(x)$$

$$= \int_0^1 H_n(x)\,dx$$

$$(4.5.22) \qquad = \int_0^1 F_{n1}^{+}(x)\,dx + (1/\sqrt{n}) \int_0^1 \mathcal{L}_{n1}^{*}(x)\,dx.$$

Now appealing to our theorems U(n) has an asymptotic normal distribution both under the hypothesis and under the alternatives (4.5.16) with means

$$(4.5.23) \qquad \mu_{1n}(U) = \int_0^1 F_{n1}^{+}(x)\,dx$$

$$= e^{-1} + e^{-1} \left(\int_0^1 \mathcal{L}^2(p)\,dp\right)/2\sqrt{n}$$

and

$$(4.5.24) \qquad \mu_{on}(U) = e^{-1}$$

with variance

$$(4.5.25) \qquad \sigma_{on}^2(U) = \sigma_{1n}^2(U) = (2e^{-1} - 5e^{-2})/n.$$

We consider next, the statistic

$$(4.5.26) \qquad L(n) = \sum_{i=1}^{n} \log (nD_i)/n$$

suggested by Darling (1953). We have already remarked about the asymptotic normality of the statistic L(n) (see Remark 2 following Theorem 4.5.3) under the hypothesis of uniformity as well as under the alternatives (4.5.16). The means are given by

$$(4.5.27)\qquad \mu_{1n}(L) = \int_0^\infty \log x \, dF^{+}_{n1}(x)$$

$$= -\gamma - \left(\int_0^1 \chi^2(p)\,dp\right)/2\sqrt{n}$$

where γ is the Euler's constant ($\gamma = 0.5772\ldots$) and

$$(4.5.28)\qquad \mu_{0n}(L) = \int_0^\infty \log x . e^{-x} dx = -\gamma$$

with variance

$$(4.5.29)\qquad \sigma^2_{0n}(L) = \sigma^2_{1n}(L) = (\pi^2/6 - 1)/n.$$

Consider now the modified spacings

$$(4.5.30)\qquad D_i^* = n D_i / h_{ni}$$

where h_{ni} satisfy

$$(4.5.31)\qquad \sup_i \sqrt{n}\,\left| h_{ni} - h(i/n)\right| = o(1)$$

where $h(p)$ is a function on $[0,1]$ having atmost a finite number of discontinuities. Then the empirical d.f.

$\{H_n^*(x),\ x \geq 0\}$ of $\{D_1^*,\ldots,D_n^*\}$ defined in (4.2.26) is distributionally equivalent to $\{F_{n2}^*(x\bar{W}_n),\ x \geq 0\}$ where $F_{n2}^*(x)$ is the empirical d.f. of the exponentially distributed r.v.'s $W_1,\ldots,W_n$ perturbed by the random factors $\{\alpha_{ni2}^*,\ i = 1,\ldots,n\}$ which have the structure defined in (4.3.37) i.e.,

$$\alpha_{ni2}^* = \theta_{ni}^*\ (1 + \beta_{ni}^*/n^{\partial} + \gamma_{ni}^*/n^{2\partial} + R_{ni}^*)$$

where θ_{ni}^*, β_{ni}^*, γ_{ni}^* and R_{ni}^* satisfy (4.3.38), (4.3.28), (4.3.27) and (4.3.25) respectively with

$$\begin{aligned} \theta(p) &= h(p) \\ \beta(p) &= \lambda(p) \\ \gamma(p) &= -L(p)\lambda'(p), \qquad 0 \leq p \leq 1. \end{aligned} \tag{4.5.32}$$

Let

$$\zeta_{n2}^*(x) = \sqrt{n}\,[H_n^*(x) - F_{n2}^+(x)] \tag{4.5.33}$$

where

$$F_{n2}^+(x) = \begin{cases} \int_0^1 (1 - e^{-xh(p)})\,dp & \text{if } \partial > 1/2 \\ \int_0^1 (1 - e^{-xh(p)})\,dp + \left(\int_0^1 xe^{-xh(p)}\lambda(p)h(p)\,dp\right)/n^{\partial} & \text{if } 1/4 < \partial \leq 1/2 \\ \int_0^1 (1 - e^{-xh(p)})\,dp + \left(\int_0^1 xe^{-xh(p)}\lambda(p)h(p)\,dp\right)/n^{1/4} & \\ \quad + \int_0^1 [-xL(p)\lambda'(p)h(p)e^{-xh(p)} & \\ \quad - x^2h^2(p)\lambda^2(p)e^{-xh(p)}/2]\,dp/n^{1/2} & \text{if } \partial = 1/4, \end{cases} \tag{4.5.34}$$

to terms of order $n^{-1/2}$. As an immediate consequence of Theorem 4.4.4, since $a_2(x)$ defined in (4.4.33) is $(-x \int_0^1 h(p)e^{-xh(p)}dp)$, we have the following

Theorem 4.5.4: The sequence of stochastic processes $\{\zeta^*_{n2}(x),\ x \geq 0\}$ in $D[0,\infty]$ converges weakly to the Gaussian process $\{\zeta_2(x),\ x \geq 0\}$ with mean zero and covariance function

$$(4.5.35)\quad K_4(x,y) = K_4(y,x) = \int_0^1 e^{-yh(p)}(1 - e^{-xh(p)})dp - xy\left(\int_0^1 h(p)e^{-xh(p)}dp\right)\left(\int_0^1 h(p)e^{-yh(p)}dp\right)$$

for $x \leq y$.

Theorems corresponding to 4.5.2 and 4.5.3 and remarks analogous to those following Theorem 4.5.3 can be immediately written for the case of modified spacings.

Theorem 4.5.5: Let $m(x)$ be an absolutely continuous function on $[0,\infty]$ with $m(0) < \infty$. Let $m'(x)$ be bounded on every finite interval and let the function

$$\int_0^\infty m'(x)y(x)dx, \quad y(\cdot) \text{ in } D[0,\infty]$$

be almost everywhere continuous with respect to the Gaussian process $\{\zeta_2(x),\ x \geq 0\}$ defined in Theorem 4.5.4. Let

$$(4.5.36)\quad T_n = \sum_1^n m(nD_i^*)/n.$$

Then the distribution of

$$(4.5.37) \qquad \sqrt{n}\,[T_n - \int_0^\infty m'(x)(1 - F^+_{n2}(x))\,dx + m(0)]$$

where $F^+_{n2}(x)$ is defined in (4.5.34), converges weakly to the normal distribution with mean zero and variance

$$(4.5.38) \qquad \int_0^\infty \int_0^\infty K_4(x,y)\; m'(x)m'(y)\,dxdy$$

where $K_4(x,y)$ is as defined in (4.5.35).

This theorem covers a very wide range of statistics based on spacings.

4.6 Asymptotic relative efficiencies of tests based on arc lengths

The Pitman asymptotic relative efficiency (ARE) of a test relative to another test is defined as the limit of the inverse ratio of sample sizes required to obtain the same limiting power at a sequence of alternatives converging to the null hypothesis. This limiting power should be a value in between the limiting size, α and the maximum power 1, in order that it can give an insight into the power behaviour of the test. If the limiting power of a test at a sequence of alternatives is α, then its ARE with respect to any other test whose

limiting power (with same size) is greater than α, is zero. On the other hand, if the limiting power of a test at a sequence of alternatives converges to a number in the interval $(\alpha, 1)$, then a measure of the rate of this convergence, called 'efficacy' can be computed. Under certain standard regularity assumptions (see eg. Fraser (1957)) which include a condition about the nature of alternatives, asymptotic normal distribution of the test statistic under these alternatives, etc., this 'efficacy' is given by

$$\text{efficacy} = \mu_\partial^2 / \sigma^2. \tag{4.6.1}$$

Here μ_∂ and σ^2 are the mean and variance of the limiting normal distribution under the sequence of alternatives when the test-statistic has been normalised to have a limiting normal distribution with mean zero and finite variance under the hypothesis. In such a situation, the ARE of one test with respect to another is simply the ratio of their efficacies.

We illustrate this concept by first computing the efficacies of some tests which are symmetric in the normalised spacings. As we have seen in Theorem 4.5.3, if

$$T_n = \sum_1^n m(nD_i)/n \tag{4.6.2}$$

then

$$\sqrt{n}\ [T_n - \int_0^\infty m'(x)e^{-x}dx\ ]$$

has a limiting normal distribution with mean zero and variance

$$\sigma^2 = \int_0^\infty \int_0^\infty m'(x)m'(y)\, K_3(x,y)\,dxdy \tag{4.6.3}$$

under the hypothesis. Again under the sequence of alternatives (4.2.1) satisfying assumption (A), the normalised statistic in (4.6.2) has a limiting normal distribution with mean μ_∂ and variance σ^2 where

$$\mu_\partial = \begin{cases} 0 & \text{if } \partial > 1/4 \\ \left(\int_0^1 \ell^2(p)\,dp\right)\left(\int_0^\infty m'(x)e^{-x}(x^2/2 - x)\,dx\right) & \text{if } \partial = 1/4. \end{cases} \tag{4.6.4}$$

Thus if $\partial > 1/4$, the efficacy of such a test is zero. The correct rate, therefore, at which the alternatives must converge to the hypothesis to give a non-zero efficacy is $n^{-1/4}$. When $\partial = 1/4$, the efficacy of the test based on T_n is

$$\left[\left(\int_0^1 \ell^2(p)\,dp\right)\left(\int_0^\infty m'(x)e^{-x}(x^2/2 - x)\,dx\right)\right]^2 \Big/ \int_0^\infty \int_0^\infty m'(x)\, m'(y)\, K_3(x,y)\,dx\,dy \tag{4.6.5}$$

The following table gives the efficacies of some standard tests at the sequence of alternatives in (4.2.1) with $\partial = 1/4$.

Table 4.1: Efficacies of some test statistics based on arc lengths.

Test Statistic		Efficacy/ $(\int_0^1 \ell^2(p)dp)^2$
$V_r(n)$:	$r = 0.0$	0.0000
	$r = 0.5$	0.6760
	$r = 1.0$	0.0000
	$r = 1.5$	0.9700
	$r = 2.0$	1.0000
	$r = 2.5$	0.9728
	$r = 3.0$	0.9000
	$r = 3.5$	0.7976
	$r = 4.0$	0.6792
U(n):		0.5726
L(n):		1.5505

From this table, we conclude that Darling's test statistic L(n), defined in (4.5.26), has maximum efficacy among the tests considered above. This is not altogether surprising in view of the fact that L(n) corresponds to Bartletts' M-test for homogeneity of variances, since each (nD_i) has, under the hypothesis, asymptotically the same distribution as that of a variance with 2 degrees of freedom from a normal sample.

It is interesting to note that $(\int_0^1 \ell^2(p)dp)$ enters in all the above efficacies in a multiplicative way. Thus the relative

efficiency of two tests which are symmetric in the spacings is independent of the particular sequence of alternatives (the choice of $L(x)$) in (4.2.1). The test statistic $V_2(n)$ has an efficiency of 64.5 per cent as against the statistic $L(n)$ whereas $U(n)$ has an efficiency of only 36.9 per cent when compared with $L(n)$.

The above discussion presents for the first time a complete and rigourous account of the computation of the efficacies and efficiencies of tests based on spacings. A number of previous attempts at the computation of the ARE's (see eg. Pyke (1965), Proschan and Pyke (1965) and Jackson (1967)) have not fully justified the method of computing the efficiencies. They obtain the limiting distribution of the test statistics under fixed alternatives and the efficiencies were then obtained in terms of the derivative of the asymptotic mean and variance under this alternative. Weiss suggested the use of alternatives (4.2.1) in the discussion of Pyke (1965), following the work of Cibisov (1961). Weiss (1965) himself obtained the distribution of $\sum_1^n D_i^2$ under these alternatives, following a specific method. The idea of finding the power function for near alternatives brings to mind the Lecam-Hajek's concept of 'contiguous' alternatives, which however, does not turn out to be very fruitful in our case.

Similarly using Theorem 4.5.5, we can compute the ARE's of tests which are symmetric in modified spacings. We defined $\{ D_i^* = nD_i / h_{ni},\ i = 1,\ldots, n \}$ as the modified spacings where the factors $\{ h_{ni} \}$ satisfy the condition

$$\sup_i \sqrt{n}\ |h_{ni} - h(i/n)| = o(1). \tag{4.6.6}$$

If m be any function on $[0,\infty]$ satisfying the conditions of Theorem 4.5.5 we define a symmetric statistic based on the modified spacings

$$T_n^* = \sum_{i=1}^{n} m(nD_i^*)/n. \tag{4.6.7}$$

The mean under the hypothesis of this T_n^* is

$$\mu_{on}^* = \int_0^\infty \int_0^1 m'(x) e^{-xh(p)}\, dx\, dp \tag{4.6.8}$$

and under the alternatives (4.2.1)

$$\begin{aligned} \mu_{1n}^* &= \mu_{on}^* \quad \text{if} \quad \delta > 1/2 \\ &= \mu_{on}^* + A(m,L,h)/n^{\delta}, \quad \text{say if} \quad 1/4 < \delta \leq 1/2 \\ &= \mu_{on}^* + A(m,L,h)/n^{1/4} + B(m,L,h)/n^{1/2}, \\ &\qquad \text{say if} \quad \delta = 1/4 \end{aligned} \tag{4.6.9}$$

where

$$A(m,L,h) = -\int_0^\infty \int_0^1 m'(x)x\lambda(p)h(p)e^{-xh(p)}\,dxdp$$

(4.6.10)

$$B(m,L,h) = \int_0^\infty \int_0^1 m'(x)e^{-xh(p)}[xL(p)\lambda'(p)h(p) + (x^2/2)\lambda^2(p)h^2(p)]\,dxdp.$$

If $A(m,L,h) \neq 0$, then the sequence of tests based on $T_n^{*'}$ can distinguish alternatives of the form (4.2.1) at a distance of order $n^{-1/2}$ from the hypothesis. This shows that such tests have a better performance than tests considered earlier which are symmetric in the normalised spacings. However there is no surety that $A(m,L,h) \neq 0$ for all L. Consider the following example. Let

(4.6.11)

$$h_{ni} = n/(n-i+1)$$

$$h(p) = 1/(1-p), \qquad 0 \leq p < 1$$

$$D_i^* = (n-i+1)D_i$$

$$m(x) = x .$$

Then

$$T_n^* = \sum_1^n m(nD_i^*)/n$$

(4.6.12)

$$= \sum_{i=1}^n [(n-i+1)D_i]/n$$

$$= \sum_{i=1}^n X_i/n + 1/n.$$

A simple computation shows

$$A(m,L,h) = \int_0^1 p\,\ell\,(p)dp \qquad (4.6.13)$$

which is n^{∂} times the excess of the mean under the alternative over that under the hypothesis and is zero for alternatives under which T_n^* has a mean $1/2$. But if this excess is non-zero, the test based on $\sum_1^n m(nD_i^*)/n$ has a better performance than symmetric normalised spacings statistics considered in Theorem 4.5.3. However if $A(m,L,h) = 0$, this test statistic T_n^* discriminates such alternatives, if at all, only when they are at a distance of $n^{-1/4}$, which puts this on par with the symmetric spacings tests. This phenomenon was mentioned and demonstrated in an easier fashion in Section 4.1. It therefore seems that if we take tests which are symmetric in modified spacings, we may gain over the tests which are symmetric in the normalised spacings on the swing but lose on the round abouts.

Another final remark before we conclude. It is known that the Kolmogorov-Smirnov statistic discriminates alternatives of the form (4.2.1) which are at a distance of order $n^{-1/2}$ from the hypothesis and hence compares favourably with any symmetric test in the spacings (see eg. Cibisov (1961)). However, Weiss (1962) has given an example of a sequence of alternatives for which the Kolmogorov-Smirnov test of level α has asymptotic

efficiency zero as compared to the symmetric spacings test based on $D_{(n)} = \max_i D_i$. The sequence of alternatives considered there by Weiss, do not, however, have a constant carrier $[0, 1]$ but only a subset of the unit interval for each n. As we remarked earlier, the definition (4.2.7) of spacings is not very meaningful for alternatives with a non-constant carrier. The anamolous situation described by Weiss arises possibly because of the lack of this feature for the alternative sequence considered by him.

CHAPTER V

BAHADUR EFFICIENCIES OF SOME TESTS FOR UNIFORMITY ON THE CIRCLE

5.1 Introduction and Summary

In this chapter we compare the asymptotic efficiencies of several tests that are available for testing uniformity on the circle. The alternatives to uniformity considered here, are the symmetric unimodal distributions, the circular normal distributions (CND's) introduced in Chapter I. We take the CND with density

$$g(\alpha) = [2\pi I_0(k)]^{-1} \exp[k \cos \alpha], \qquad -\pi \leq \alpha < \pi \tag{5.1.1}$$

for convenience. When $k = 0$, this gives the uniform density so that the null hypothesis is one which formulates $H : k = 0$. The tests compared here are (i) Rayleigh's test, R (ii) Kuiper's test, V (iii) Watson's test, W (iv) Ajne's test, N (v) Ajne's test, A and (vi) Spacings test, U_n. Each of the tests is briefly described before its efficiency is computed.

We compute the asymptotic efficiency due to Bahadur (1960b), of each test, by evaluating in most cases the 'exact slopes' of the test statistics, using large deviation results. In some cases we take the 'approximate slopes' as given by the asymptotic distributions. On the basis of these comparisons, we find that the limiting efficiencies of the three tests viz., Ajne's test A, Watson's W and Rayleigh's test based on R are identical while the other tests have asymptotic efficiencies, which are lower. Further conclusions, based on the comparisons of Bahadur efficiencies, are given in Section 5.9. Finally in Section 5.10, a simple inequality between the Ajne's N and Kuiper's V, whose asymptotic performances are identical, has been noted.

5.2 Some preliminaries

In this section, we first briefly set forth the concept of Bahadur efficiency and give some results that would be used later on. We shall say that a sequence of test statistics $\{ T_n \}$ is a 'standard' sequence for testing the hypothesis $H_o : \theta = \theta_o$ if

(a) Under H_o, there exists a non-degenerate d.f. F_n such that

$$P_{\theta_o} [T_n < t] = F_n(t) \tag{5.2.1}$$

for all real t.

(b) There exists a non-negative function $s(\cdot)$ on $[0, \infty)$ such that for any sequence of numbers $\{\lambda_n\}$ such that $\lambda_n^2/n \to \lambda$, we have

$$-(2/n) \log P_{\theta_0}[T_n \geq \lambda_n] \to s(\lambda) \tag{5.2.2}$$

and

(c) There is a non-negative real-valued function $b(\theta)$ such that

$$T_n/\sqrt{n} \to b(\theta) \tag{5.2.3}$$

almost surely or in probability, for θ in the alternative.

If $F_n(x)$ denotes the exact null d.f. of T_n, then the level of significance attained by the test T_n is given by

$$\alpha_n = 1 - F_n(T_n). \tag{5.2.4}$$

If $F_n(x)$ is continuous, then α_n is uniformly distributed in $[0, 1]$ under the null hypothesis and, under the alternative, $\alpha_n \to 0$, usually at an exponential rate. Thus for a sufficiently small level of significance, if $\theta \neq \theta_0$ is a parameter point in the alternative, then

(5.2.5) $$-2(\log L_n)/n \rightarrow s(b^2(\theta))$$

in probability if $b(\theta)$ is the probability limit of $(T_n/\sqrt{n})$ as in (5.2.3). This function, $s(b^2(\theta))$, is well-defined and we shall call this the 'exact slope' of the test sequence $\{T_n\}$. It is a measure of the performance of the test since, the larger this quantity, the faster the test based on T_n rejects H_0, for an arbitrarily fixed level of significance. Therefore the ratio of the slopes of two standard sequences of test statistics $\{T_n^{(1)}\}$ and $\{T_n^{(2)}\}$ for the hypothesis H_0, gives the 'efficiency' at θ

(5.2.6) $$E_{1,2}(\theta) = s_1(b_1^2(\theta))/s_2(b_2^2(\theta))$$

of the sequence $\{T_n^{(1)}\}$ relative to $\{T_n^{(2)}\}$. The limiting efficiency of $\{T_n^{(1)}\}$ relative to $\{T_n^{(2)}\}$ is defined as the limit of (5.2.6), for a sequence of parameter points $\{\theta\}$ in the alternative, converging to θ_0 of the hypothesis i.e., the limiting efficiency is given by

(5.2.7) $$L_{1,2}(\theta_0) = \lim_{\theta \rightarrow \theta_0} E_{1,2}(\theta).$$

When the exact slope is difficult to compute, one finds the 'approximate slope' (see eg. Gleser (1964)), which is generally

much easier to obtain. Let $F(x)$ be the asymptotic null d.f. of the test sequence $\{T_n\}$. That is

$$\lim_{n \to \infty} P_{\theta_o}(T_n < x) = F(x).$$

Suppose as $x \to \infty$, $F(x)$ satisfies the condition on the tail, namely

$$-2 \log (1 - F(x)) = ax^2(1 + o(1)) \tag{5.2.8}$$

where 'a' is some positive constant. Then

$$s^*(b^2(\theta)) = a\,[b(\theta)]^2 \tag{5.2.9}$$

would be referred to as the 'approximate slope' of the test sequence $\{T_n\}$ where $b(\theta)$ is the function given by the condition (c) i.e., (5.2.3). One can, then, compute the efficiencies as in (5.2.6) and (5.2.7) on the basis of this approximate slope $s^*(b^2(\theta))$. We now give here a few lemmas that will be useful in the computation of the slopes of the test statistics.

Lemma 5.2.1 If $X_1, X_2 \ldots$ are independently and identically distributed binomial random variables with mean 'p', then

$$(5.2.10)\qquad (1/n)\log P\left\{\left|\sum_{i=1}^{n} X_{i/n} - p\right| \geq \lambda\right\} \to \log \varphi^*(p,\lambda)$$

where

$$(5.2.11)\qquad \varphi^*(p,\lambda) = \text{Max}\,[\varphi_1(p,\lambda),\ \varphi_2(p,\lambda)]$$

with

$$(5.2.12)\qquad \varphi_1(p,\lambda) = \begin{cases} \left[p/(p+\lambda)\right]^{p+\lambda}\left[(1-p)/(1-p-\lambda)\right]^{1-p-\lambda}, \\ \qquad\qquad \text{for } 0 \leq p \leq 1-\lambda \\ 0 \qquad\qquad \text{otherwise} \end{cases}$$

and

$$(5.2.13)\qquad \varphi_2(p,\lambda) = \begin{cases} \left[(1-p)/(1-p+\lambda)\right]^{1-p+\lambda}\left[p/(p-\lambda)\right]^{p-\lambda}, \\ \qquad\qquad \text{for } \lambda \leq p \leq 1 \\ 0 \qquad\qquad \text{otherwise.} \end{cases}$$

This result may be found eg. in Chernoff (1952). The following useful generalisation is due to Sethuraman (1964).

Lemma 5.2.2 Let $\mathfrak{X}$ be a separable Banach space and $\mathfrak{X}_1^*$, the space of all continuous linear functionals x^*, on $\mathfrak{X}$, with norm unity. Let $X_1(\omega), X_2(\omega), \ldots$ be a sequence of random variables, defined on a probability space (Ω, S, P) with values in $\mathfrak{X}$, which are independently and identically distributed with a common distribution $P(\cdot)$. Let

$$(5.2.14)\qquad \int x^*(X(\omega))\, P(d\omega) = 0 \qquad \text{for all } x^* \in \mathfrak{X}_1^*$$

and let

(5.2.15) $$\int \exp\,(t\,\|X(\omega)\|\,)\,P(d\omega) < \infty \qquad \text{for all } t.$$

Then for $\epsilon > 0$,

(5.2.16) $$\frac{1}{n}\log P\left\{\omega : \left\|\frac{X_1(\omega)+\ldots+X_n(\omega)}{n}\right\| \geq \epsilon\right\} \to \log \rho(\mathcal{H}_1^*,\epsilon)$$

where

(5.2.17) $$\rho(\mathcal{H}_1^*,\epsilon) = \sup_{x^* \in \mathcal{H}_1^*} \rho(x^*,\epsilon)$$

and

(5.2.18) $$\rho(x^*,\epsilon) = \text{Max}\left\{\rho_1(x^*,\epsilon) = \min_{t\geq 0} e^{-t\epsilon}\,E(\exp[tx^*(X(\omega))]),\right.$$

$$\left.\rho_2(x^*,\epsilon) = \min_{t\leq 0} e^{t\epsilon}\,E(\exp[t\,x^*(X(\omega))])\right\}.$$

We now state and prove a simple lemma which tells us something about the behaviour of $\rho(\mathcal{H}_1^*,\epsilon)$ defined in (5.2.17), for sufficiently small ϵ. Since $b(\theta)$ is zero for θ in the null hypothesis, it is only the behaviour of $\rho(\mathcal{H}_1^*,\epsilon)$ for ϵ small, that counts in the computation of the limiting efficiencies.

Lemma 5.2.3 For $\epsilon > 0$, sufficiently small

(5.2.19) $$\rho(\mathcal{H}_1^*\,\epsilon) \sim \exp\,[-\epsilon^2/2\sigma^{*2}]$$

where

$$\sigma^{*2} = \sup_{x^* \in \mathcal{X}_1^*} \text{Var } x^*(X(\omega)). \tag{5.2.20}$$

Proof: For any fixed $x^* \in \mathcal{X}_1^*$, we have from Lemma 5.2.2

$$\varphi(x^*, \epsilon) = \max \{ \varphi_1(x^*, \epsilon), \varphi_2(x^*, \epsilon) \}$$

where $\varphi_1(x^*, \epsilon)$ and $\varphi_2(x^*, \epsilon)$ are as given in (5.2.18). Now let $\varphi_1(x^*, \epsilon)$ attain the minimum at the point $t = t_1$. Then, for sufficiently small ϵ, in view of (5.2.14),

$$t_1 = \epsilon/\sigma^2 + o(\epsilon)$$

where $\sigma^2 = \text{var } x^*(X(\omega))$ and $o(\cdot)$ consists of terms in ϵ^2 and higher powers of ϵ, the coefficients of which involve the moments of $x^*(X(\omega))$. But since the moments of $x^*(X(\omega))$ are bounded by the corresponding moments of $||X(\omega)||$, which are all finite by (5.2.15), this $o(\cdot)$ holds uniformly for all x^* in $\mathcal{X}_1^*$. Now for ϵ small, if $\phi(t)$ denotes the moment generating function of $x^*(X(\omega))$,

$$\begin{aligned} \log \varphi_1(x^*, \epsilon) &= -\epsilon t_1 + \log \phi(t_1) \\ &= -\epsilon^2/\sigma^2 + \epsilon^2/2\sigma^2 + o(\epsilon^2) \\ &= -\epsilon^2/2\sigma^2 + o(\epsilon^2) \end{aligned} \tag{5.2.21}$$

where again $o(\cdot)$ is uniform in $x^* \in \mathcal{X}_1^*$ because of the reasons given earlier. An expression, similar to (5.2.21), is valid for the other term, namely $g_2(x^*, \epsilon)$ and hence for their maximum $g(x^*, \epsilon)$. Thus for ϵ sufficiently small

$$\log g(x^*, \epsilon) = -\epsilon^2/2\sigma^2 + o(\epsilon^2). \tag{5.2.22}$$

Now since $o(\cdot)$ is uniform in $x^* \in \mathcal{X}_1^*$, taking supremum over $x^* \in \mathcal{X}_1^*$, on either side of (5.2.22) we have the required result.

5.3 Rayleigh's test

Let $\alpha_1, \ldots, \alpha_n$ be n observations on the circumference of the unit circle. They may be expressed equivalently as unit vectors $u_i = (\cos \alpha_i, \sin \alpha_i) = (x_i, y_i)$, $i = 1, \ldots, n$. The Rayleigh's test for uniformity, as mentioned in Section 3.1, is based on the length R of the vector resultant of these u_i's i.e., on

$$R = \sqrt{\left(\sum_1^n x_i\right)^2 + \left(\sum_1^n y_i\right)^2} = n\sqrt{\bar{x}^2 + \bar{y}^2} \tag{5.3.1}$$

where $\bar{x} = \sum_1^n x_i/n$ and $\bar{y} = \sum_1^n y_i/n$. We reject the hypothesis of uniformity when R is too large. The null distribution of R is given in (2.3.1). We consider

(5.3.2) $$T_n^{(1)} = \sqrt{2/n}\ R$$

as the 'standard sequence' in this case. Let us denote the probabilities under the hypothesis of uniformity H_o, by P_o. Now we require

$$\begin{aligned} P_o\,[T_n^{(1)} \geq \sqrt{n}\,\lambda] &= P_o[R/n \geq \lambda/\sqrt{2}\,] \\ &= P_o[(\bar{x}^2 + \bar{y}^2)^{1/2} \geq \lambda/\sqrt{2}\,] \\ &= P_o[\,\|\textstyle\sum_1^n u_i/n\| \geq \lambda/\sqrt{2}\,] \end{aligned} \tag{5.3.3}$$

where $\|\cdot\|$ denotes the usual Euclidean norm in R^2.

We now appeal to Lemmas 5.2.2 and 5.2.3 where the Banach space is simply the Euclidean space R^2 with the usual norm. From (5.2.16), (5.2.19) and (5.2.20), for λ small, we have

$$\begin{aligned} P_o[T_n^{(1)} \geq \sqrt{n}\,\lambda] &= P_o[R/n \geq \lambda/\sqrt{2}] \\ &= P_o[\,\|\textstyle\sum_1^n u_i/n\,\| > \lambda/\sqrt{2}\,] \\ &= \exp\,[-n(\lambda/\sqrt{2})^2/2\,\sigma_o^2] \end{aligned} \tag{5.3.4}$$

where

$$\sigma_o^2 = \sup_{\{(\ell,m):\ell^2+m^2=1\}} \mathrm{Var}\,(\ell\cos\alpha + m\sin\alpha). \tag{5.3.5}$$

The σ_o^2 (variance under H_o) required here is simply the

maximum latent root of the covariance matrix of the components 'cos α' and 'sin α'. Since the variances are 1/2 each and covariance zero under uniformity, $\sigma_o^2 = 1/2$ and we get

$$(5.3.6) \qquad P_o[T_n^{(1)} \geq \sqrt{n}\,\lambda] = P_o[R/n \geq \lambda/\sqrt{2}\,] \sim \exp[-n\lambda^2/2],$$

for λ small. Under CN alternatives (5.1.1), it is easy to see that

$$(5.3.7) \qquad T_n^{(1)}/\sqrt{n} = \sqrt{2}\,R/n \xrightarrow{p} \sqrt{2}\,\vartheta(k) = b_1(k), \text{ say}$$

where $\vartheta(k)$ is the ratio $[I_1(k)/I_o(k)]$ defined in Chapter II and '$\xrightarrow{p}$' denotes convergence in probability. Thus from (5.3.7) and (5.3.6), the exact Bahadur slope of the Rayleigh statistic, $T_n^{(1)}$ is given by

$$(5.3.8) \qquad s_1(k) = [b_1(k)]^2 = 2[\vartheta(k)]^2.$$

5.4 Kuiper's test

Kuiper (1960) proposed a variant of the Kolmogorov-Smirnow statistic for testing the hypothesis that the observations come from a population with specified d.f. $F(x)$. If $F_n(x)$ denotes the empirical d.f., then his statistic is

$$(5.4.1) \qquad V_n = \sqrt{n}\left\{ \sup_x [F_n(x) - F(x)] - \inf_x [F_n(x) - F(x)] \right\}$$

which is especially suitable for testing goodness of fit on the circle, as the statistic V_n is independent of the origin used for measurement of α. We consider the 'standard' sequence

$$(5.4.2) \qquad T_n^{(2)} = V_n.$$

Kuiper (1960) found the distribution of V_n and we have

$$(5.4.3) \qquad P_0\{T_n^{(2)} \geq z\} = \sum_{m=1}^{\infty} 2(4m^2z^2 - 1)e^{-2m^2z^2} - 8z/3\sqrt{n} \cdot \sum_{m=1}^{\infty} m^2(4m^2z^2 - 3)e^{-2m^2z^2} + O(1/n).$$

In view of the fact that under the alternatives $G(x) \neq F(x)$,

$$(5.4.4) \qquad T_n^{(2)}/\sqrt{n} \xrightarrow{\text{a.s.}} \left\{ \sup_x [G(x)-F(x)] - \inf_x [G(x)-F(x)] \right\}.$$

(with $\xrightarrow{\text{a.s.}}$ denoting convergence almost surely or with probability one) the approximate slope of this sequence of statistics is given by

$$(5.4.5) \qquad s_2^*(G) = 4\left\{ \sup_x [G(x)-F(x)] - \inf_x [G(x)-F(x)] \right\}^2.$$

The exact slope of $\{ V_n \}$ is also not difficult to obtain but this approximate slope is sufficient for our purposes as it has been shown to be reliable (c.f. Abrahamson (1967)), in the sense that the ratio of the approximate slope to the exact one tends to unity for alternatives approaching the hypothesis. Now for the CN alternatives (5.1.1) and the null distribution of uniformity, it is easy to see that

$\sup_{-\pi \leq \alpha < \pi} [G(\alpha) - F(\alpha)]$ is attained at $\alpha = \pi/2$ and the $\inf_{-\pi \leq \alpha < \pi} [G(\alpha) - F(\alpha)]$ at $\alpha = -\pi/2$. Thus, from (5.4.4)

$$T_n^{(2)} / \sqrt{n} \xrightarrow{a.s.} \int_{-\pi}^{\pi/2} [g(\alpha) - 1/2\pi] d\alpha - \int_{-\pi}^{-\pi/2} [g(\alpha) - 1/2\pi] d\alpha$$

(5.4.6)

$$= [2\pi I_o(k)]^{-1} \int_{-\pi/2}^{\pi/2} e^{k \cos \alpha} \, d\alpha - 1/2$$

$$= [P(k) - 1/2]$$

where $P(k)$ is the probability that a CN random variable lies within an arc of length $\pi/2$ from its mean direction. Hence from (5.4.6),

$$s_2^*(k) = [2P(k) - 1]^2. \tag{5.4.7}$$

$P(k)$ may be expressed in the following series

$$P(k) = [2\pi\, I_0(k)]^{-1} \int_{-\pi/2}^{\pi/2} e^{k\cos\alpha}\, d\alpha$$

(5.4.8)

$$= [2\pi\, I_0(k)]^{-1} \int_{-\pi/2}^{\pi/2} \Big[\sum_{r=0}^{\infty} (k^r \cos^r \alpha)/r!\Big]\, d\alpha$$

$$= [2\pi\, I_0(k)]^{-1} \sum_{r=0}^{\infty} k^r/r!\; \beta(r+1/2,\ 1/2).$$

5.5 Watson's test

Watson (1961) proposed the statistic

$$(5.5.1) \qquad W_n = n \int_{-\infty}^{\infty} \Big\{ F_n(x) - F(x) - \int_{-\infty}^{\infty} [F_n(y) - F(y)]\, dF(y) \Big\}^2 dF(x)$$

for testing the hypothesis that the sample, with d.f. $F_n(x)$, was drawn from a population with continuous d.f. $F(x)$. It can be used on the circle, since its value remains independent of the choice of the arbitrary point from which we begin cumulating the probability density or the masses corresponding to the sample points. The asymptotic distribution of W_n is given by (c.f. Watson (1961))

$$(5.5.2) \qquad \lim_{n \to \infty} \text{Prob.}[W_n > v] = 2 \sum_{m=1}^{\infty} (-1)^{m-1} \exp(-2m^2\pi^2 v).$$

We consider

$$(5.5.3) \qquad T_n^{(3)} = \sqrt{W_n}$$

as the 'standard' sequence in this case. Since

$$\log P(W_n > v) = -2\pi^2 v\,(1 + o(1)) \qquad \text{as} \quad v \to \infty$$

the asymptotic distribution of W_n satisfies the condition on the tail, (5.2.8), with $a = 4\pi^2$. Let $G(\alpha)$ denote the cumulative distribution function corresponding to the alternative density (5.1.1). Then in testing for uniformity i.e., when the hypothetical d.f. $F(\alpha)$ is the distribution function corresponding to the uniform density,

$$(5.5.4) \qquad (T_n^{(3)}/\sqrt{n})^2 = W_{n/n} \xrightarrow{a.s.} \int_0^{2\pi} \Big\{ G(\alpha) - \alpha/2\pi - \int_0^{2\pi} [G(\beta) - \beta/2\pi]\, d\beta/2\pi \Big\}^2 \, d\alpha/2\pi .$$

Therefore the approximate slope of the standard sequence $\{ T_n^{(3)} \}$ is given by

$$(5.5.5) \qquad s_3^*(k) = 4\pi^2 \int_0^{2\pi} \Big\{ G(\alpha) - \alpha/2\pi - \int_0^{2\pi} [G(\beta) - \beta/2\pi]\, d\beta/2\pi \Big\}^2 \, d\alpha/2\pi .$$

5.6 Ajne's test, N

Given n observations on the circumference of the unit circle, let $N(\alpha)$ denote the number of points on the half circle $[\alpha, \alpha + \pi)$, taking say the anticlockwise direction as positive. Then Ajne (1966) utilises

(5.6.1) $$N = \sup_{0 \le \alpha < 2\pi} N(\alpha),$$

which is the maximum number of points in any semicircle, for testing for uniformity. We consider, here, the 'standard' sequence

(5.6.2) $$T_n^{(4)} = (N - n/2)/\sqrt{n}$$

$$= \sqrt{n}\,[N/n - 1/2] = N_n^*\ ,\ \text{say.}$$

Then we want the probability

(5.6.3) $$P_o\left\{ T_n^{(4)} \ge \sqrt{n}\,\lambda \right\} = P_o\left\{ (N/n - 1/2) \ge \lambda \right\}.$$

We shall now obtain this by getting upper and lower bounds for this probability. Since, for any fixed α

$$N = \text{Sup } N(\alpha) \ \ge \ N(\alpha),$$

(5.6.4)

$$(1/n) \log P_o\left\{ (N/n - 1/2) \ge \lambda \right\}$$

$$\ge (1/n) \log P_o\left\{ (N(\alpha)/n - 1/2) \ge \lambda \right\}.$$

But $N(\alpha)$, under the hypothesis has a binomial distribution with parameter $1/2$. Therefore, from Lemma 5.2.1,

(5.6.5) $$(1/n) \log P_o\left\{ (N/n - 1/2) \ge \lambda \right\} \ge \log g(1/2, \lambda).$$

In order to get an upper bound for the probability in (5.6.3), for some $\partial > 0$ (to be chosen suitably later on), we divide the whole length of the circumference into $N(\partial) = [2\pi/\partial] + 1$ arcs of length ∂ each and define

(5.6.6) $N(\alpha, \partial)$ = number of observations in the arc $[\alpha, \alpha + \pi + \partial)$.

Then, clearly

(5.6.7) $$N(\alpha, \partial) \geq N(\alpha)$$

for all α. Further since any α lies between $[r\partial, \overline{r+1}\ \partial)$ for some r, the corresponding $N(\alpha) \leq N(r\partial, \partial)$ so that

(5.6.8) $$N = \sup_{\alpha} N(\alpha) \leq \max_{r} N(r\partial, \partial).$$

Thus

$$P_o\{[N/n - 1/2] \geq \lambda\} \leq P_o\{\max_r [N(r\partial,\partial)/n - 1/2] \geq \lambda\}$$

$$= P_o\{[N(r\partial,\partial)/n - 1/2] \geq \lambda, \text{ for at least one } r\}$$

(5.6.9)

$$\leq \sum_{r=1}^{N(\partial)} P_o\{[N(r\partial,\partial)/n - 1/2] \geq \lambda\}.$$

Now $N(\alpha,\partial)$ has a binomial distribution with parameter $(1/2 + \partial)$ and all the terms in the summation (5.6.9) are equal so that

$$P_o\{(N_{/n} - 1/2) \geq \lambda\} \leq N(\partial).P_o\{N(r\partial,\partial)_{/n} - 1_{/2} - \partial \geq \lambda - \partial\}$$

(5.6.10)

$$= N(\partial).\ \varphi^n\ (1_{/2} + \partial,\ \lambda - \partial)$$

where $\varphi(p, \lambda)$ is as defined in Lemma 5.2.1. Now if we choose $\partial = 1/n$, say, then the term on the RHS of (5.6.10) is $2\pi n.\varphi^n(1_{/2} + 1/n,\ \lambda - 1/n)$ so that

$$\overline{\lim}\ (1/n) \log P_o\{[N_{/n} - 1/2] \geq \lambda\}$$

(5.6.11)

$$\leq \overline{\lim} \log \varphi\ (1_{/2} + 1_{/n},\ \lambda - 1/n)$$

$$= \log\ \varphi\ (1_{/2},\ \lambda).$$

Thus from (5.6.9) and (5.6.11), we have

$$(5.6.12) \qquad (1/n) \log P_o\{(N_{/n} - 1/2) \geq \lambda\} = \log \varphi(1/2, \lambda) \sim -2\lambda^2$$

for λ sufficiently small.

Now, in order to get the probability limit to which $T_n^{(3)}/\sqrt{n}$ converges under the CN alternatives, we observe that since $N(\alpha)$ is a binomial sum, for any fixed α,

$$(5.6.13) \qquad N(\alpha)_{/n} \xrightarrow{p} p(\alpha) = P\{\alpha \leq \theta < \alpha + \pi\}$$

where θ is a CN random variable with density (5.1.1).

This stochastic convergence is uniform in α since the random variables involved in the summation $N(\alpha)$ are clearly uniformly bounded in n and α. (See eg. Parzen (1954)). Thus

$$(5.6.14) \qquad N/n = \sup N(\alpha)/n \xrightarrow{p} \sup_{-\pi \leq \alpha < \pi} p(\alpha) .$$

But for the density defined in (5.1.1), this supremum is attained when $\alpha = -\pi/2$ so that

$$(5.6.15) \qquad N/n \xrightarrow{p} [2\pi I_o(k)]^{-1} \int_{-\pi/2}^{\pi/2} e^{k \cos \alpha} \, d\alpha$$

$$= P(k)$$

where $P(k)$ is as defined in (5.4.8). Hence

$$T_n^{(4)} / \sqrt{n} = (N/n - 1/2) \xrightarrow{p} [P(k) - 1/2]$$

and the exact slope of the sequence $\{ T_n^{(4)} \}$ is given by

$$(5.6.16) \qquad s_4(k) = [\, 2 P(k) - 1]^2$$

which is the same as that of Kuiper's test given in (5.4.7).

We remark, in passing, that the statistic N is similar to the Hodges' (1955) bivariate sign-test statistic, where for testing the equality of two bivariate d.f.'s, he proposes the maximum number of vector differences (differences between the observed vectors in the 2 populations) with positive

projections on some line through the origin, as the direction of this line is varied. Similar interesting relations between non-parametric tests in bivariate situations and tests for circular distributions exist. Another case in point is the test proposed recently by Mardia (1967) for bivariate populations and the circular test suggested by Wheeler and Watson (1964).

5.7 Ajne's test A

If $N(\alpha)$ is as defined in Section 5.6, due to Ajne (1966), is also the statistic

$$A_n = \frac{1}{(2\pi n)} \int_0^{2\pi} [N(\alpha) - n/2]^2 \, d\alpha \tag{5.7.1}$$

which can be used for testing uniformity on the circle. Here we take

$$T_n^{(5)} = \sqrt{A_n} \tag{5.7.2}$$

as the 'standard' test sequence. Define, corresponding to any fixed α and the i^{th} observation, α_i,

$$Y_i(\alpha) = \begin{cases} 1/2 & \text{if } \alpha \leq \alpha_i < \alpha + \pi \\ -1/2 & \alpha + \pi \leq \alpha_i < \alpha + 2\pi, \end{cases} \tag{5.7.3}$$

for $i = 1, \ldots, n$. These Y_i's will be treated as random variables (r.v.'s) in $L_2(0, 2\pi)$. We may then write

$$(5.7.4) \qquad (1/n)\,[N(\alpha) - n/2] = (1/n) \sum_{i=1}^{n} Y_i(\alpha) = \bar{Y}_n(\alpha).$$

Therefore

$$(5.7.5) \qquad \begin{aligned} P_0\{T_n^{(5)} \geq \lambda\sqrt{n}\} &= P_0\Big\{\frac{1}{(2\pi n)}\int_0^{2\pi}[N(\alpha) - \tfrac{n}{2}]^2\,d\alpha \geq \lambda^2 n\Big\} \\ &= P_0\Big\{\frac{1}{2\pi}\int_0^{2\pi}[\tfrac{1}{n}(N(\alpha) - \tfrac{n}{2})]^2\,d\alpha \geq \lambda^2\Big\} \\ &= P_0\{\|\bar{Y}_n(\cdot)\|^2 \geq \lambda^2\} \end{aligned}$$

where

$$\|f(\cdot)\|^2 = \frac{1}{2\pi}\int_0^{2\pi} f^2(\alpha)\,d\alpha$$

is the usual norm in $L_2(0, 2\pi)$. This way of looking at $T_n^{(5)}$ allows us to make use of the Lemmas 5.2.2 and 5.2.3, with the Banach space $\mathfrak{X}$ as the space $L_2(0, 2\pi)$ and $\mathfrak{X}_1^*$, the space of all continuous linear functionals on it with norm unity.

For utilising Lemma 5.2.3, we require

$$(5.7.6) \qquad \sigma^{*2} = \sup_{x^* \in \mathfrak{X}_1^*} \mathrm{Var}\,[x^*(Y)]$$

under the hypothesis of uniformity. Corresponding to any element $f(\cdot) \in L_2(0,2\pi)$, define the real-valued random variable

$$(5.7.7) \qquad Z = (f(\cdot), Y(\cdot)) = \frac{1}{2\pi} \int_0^{2\pi} f(\alpha)\, Y(\alpha)\, d\alpha \,.$$

Then, from the definition, (5.7.3), of the random variable $Y(\alpha)$, under the hypothesis of uniformity,

$$(5.7.8) \qquad E(Z) = 0$$

$$\text{Var } (Z) = \frac{1}{4\pi^2} \int_0^{2\pi} \int_0^{2\pi} f(\alpha) f(\beta)\, K(\alpha,\beta)\, d\alpha\, d\beta$$

where

$$K(\alpha,\beta) = \text{Cov } (Y(\alpha), Y(\beta))$$

$$(5.7.9) \qquad = \begin{cases} (\beta + \pi - \alpha)/2\pi - \frac{1}{4} & \text{if } \beta \le \alpha < \beta + \pi \\ (\alpha - \beta - \pi)/2\pi - \frac{1}{4} & \text{if } \beta + \pi \le \alpha < \beta + 2\pi. \end{cases}$$

The σ^{*2} of (5.7.6), then, is the supremum of Var (Z) over all $f(\cdot)$ in $L_2(0, 2\pi)$ such that $\| f(\cdot) \|$ is unity. To obtain this, we use the following standard Fourier expansion methods. For example, it can be checked that the covariance

kernel $K(\alpha,\beta)$, given in (5.7.9), has the eigen functions

$$(5.7.10) \qquad \{ e^{in\beta},\ n = \pm 1, \pm 3, \pm 5, \cdots \}$$

with the corresponding eigen values

$$(5.7.11) \qquad \lambda_{\pm n} = \begin{cases} 0 & \text{if } n \text{ is even} \\ 2/n^2\pi & \text{if } n \text{ is odd.} \end{cases}$$

Therefore, we can write

$$(5.7.12) \qquad Y(\alpha) = \sum_{n\neq 0} X_n\, e^{in\alpha}$$

where

$$X_n = \frac{1}{2\pi} \int_0^{2\pi} Y(\alpha)\, e^{-in\alpha}\, d\alpha$$

is a random variable with mean zero and variance

$$\frac{1}{4\pi^2} \int_0^{2\pi} \int_0^{2\pi} K(\alpha,\beta)\, e^{-in\alpha}\, e^{in\beta}\, d\alpha\, d\beta = \frac{1}{2\pi}\, \lambda_n$$

with λ_n as given in (5.7.11). Further if $f(\alpha)$ has the fourier expansion

$$f(\alpha) = \sum_n c_n\, e^{in\alpha}$$

then from (5.7.12) and (5.7.7),

$$(5.7.13) \qquad Z = \frac{1}{2\pi} \int_0^{2\pi} Y(\alpha)\, f(\alpha)\, d\alpha$$

$$= \sum_{\substack{n \neq 0 \\ n \text{ odd}}} c_n X_n$$

whose variance we can now rewrite as

$$\text{Var}\,(Z) = \frac{1}{2\pi} \sum_{\substack{n \neq 0 \\ n \text{ odd}}} |c_n|^2 \lambda_n \,.$$

$$(5.7.14) \qquad = \sum_{k=-\infty}^{\infty} |c_{2k+1}|^2 \, [1/(2k+1)^2 \pi^2].$$

Thus, we require the supremum of (5.7.14) subject to the restriction

$$(5.7.15) \qquad \| f(\cdot) \|^2 = \sum_{n=-\infty}^{\infty} |c_n|^2 = 1.$$

Clearly this maximisation occurs when $|c_1|^2 = 1$ with all the other Fourier coefficients of $f(\cdot)$ vanishing. Hence

$$(5.7.16) \qquad \sigma^{*2} = \sup_{\|f\| = 1} \text{Var}\ (f(\cdot), Y(\cdot)) = 1/\pi^2.$$

Now, appealing to Lemma 5.2.3,

$$(5.7.17)\qquad (1/n)\log P_0\left\{T_n^{(5)} \geq \lambda\sqrt{n}\right\}$$

$$= (1/n)\log P_0\left\{\|\bar{Y}_n(\cdot)\|^2 \geq \lambda^2\right\}$$

$$\sim -\lambda^2/2\sigma^{*2}$$

$$= -\lambda^2\pi^2/2$$

for λ sufficiently small. It may be noted that this exact slope function agrees with the approximate slope (Equation (6.5)), obtained by Beran (1968), using the asymptotic characteristic function. Beran (1968) also showed that if $\{d_m\}$ be the Fourier transforms of the alternate d.f. $G(\alpha)$ and $\{c_m\}$ those of $[N(\alpha) - n/2]$, then

$$(5.7.18)\qquad A_{n/n} \xrightarrow{a.s.} \sum_{m\neq 0} |c_m|^2|d_m|^2$$

In our case, it can be seen that

$$(5.7.19)\qquad c_m = \begin{cases} 1/\pi i m & \text{if } m \text{ odd} \\ 0 & \text{otherwise} \end{cases}$$

and that the Fourier transforms of the CN distribution given in (5.1.1) are

$$(5.7.20)\qquad d_m = I_m(k)/I_0(k), \quad m = \pm 1, \pm 2, \ldots$$

Hence from (5.7.17), (5.7.18), (5.7.19) and (5.7.20), the exact slope of the sequence $\{ T_n^{(5)} \}$ is given by

$$(5.7.21) \qquad s_5(k) = \pi^2 \sum_{\substack{m \neq 0 \\ m \text{ odd}}} I_m^2(k) / \pi^2 m^2 I_0^2(k)$$

$$= 2 \sum_{j=0}^{\infty} I_{2j+1}^2(k) / (2j+1)^2 I_0^2(k).$$

5.8 The spacings test U_n

If $\{ T_i, i = 1, \ldots, n \}$ denote the lengths of the n arcs made by the sample points, then we considered in Chapter III', the spacings test based on the statistic

$$(5.8.1) \qquad U_n = \sum_{i=1}^{n} \max [T_i - 2\pi/n, 0]$$

$$= \frac{1}{2} \sum_{i=1}^{n} |T_i - 2\pi/n|.$$

In this case we take the test sequence

$$(5.8.2) \qquad T_n^{(6)} = \sqrt{n} [U_n - 2\pi/e] / 2\pi \sqrt{2e^{-1} - 5e^{-2}}$$

as the 'standard' sequence. We have seen that this has an asymptotic normal distribution with mean zero and unit variance,

so that this sequence of test statistics satisfies the condition (5.2.8) on the tail with $a = 1$. Further, from Theorem 3.4.2, under the alternatives (5.1.1),

$$(5.8.3), \qquad U_n \xrightarrow{p} \int_0^{2\pi} e^{-2\pi g(\alpha)}\, d\alpha$$

$$= \int_0^{2\pi} \sum_{j=0}^{\infty} \frac{(-1)^j}{j!} \left(\frac{e^{k \cos \theta}}{I_0(k)}\right)^j$$

$$= 2\pi \sum_{j=0}^{\infty} (-1)^j I_0(jk)/j! I_0^j(k).$$

Thus the approximate slope of this standard sequence, $\{T_n^{(6)}\}$, is given by

$$(5.8.4) \qquad s_6^*(k) = \left[\sum_{j=0}^{\infty} \frac{(-1)^j}{j!} \frac{I_0(jk)}{I_0^j(k)} - e^{-1} \right]^2 / [2e^{-1} - 5e^{-2}].$$

5.9. Comparison of the limiting efficiencies

In this section, we compare the limiting efficiencies of the six test statistics that have been studied in Sections 5.3 to 5.8, on the basis of the slopes given in (5.3.8), (5.4.7), (5.5.5), (5.6.16), (5.7.21) and (5.8.4) respectively.

First we note that Kuiper's statistic V and Ajne's N are equally efficient asymptotically for testing uniformity.

as their slopes are identical. The comparison of the limiting efficiencies is made easier by considering approximations for the slopes when k is small, since in any case, we let k tend to zero for obtaining the limiting efficiencies. (See definition (5.2.7)). Since $\vartheta(k) \doteq k/2$ for k small,

$$s_1(k) = 2[\vartheta(k)]^2 \doteq k^2/2 \tag{6.9.1}$$

for small k. Moreover, for k sufficiently small, the CN density (5.1.1) can be approximated by the density

$$g^*(\alpha) = (1 + k \cos \alpha)/2\pi, \quad -\pi \leq \alpha < \pi \tag{5.9.2}$$

so that

$$P(k) \doteq \left(\frac{1}{2\pi}\right) \int_{-\pi/2}^{\pi/2} (1 + k \cos \alpha) d\alpha \tag{5.9.3}$$

$$= \left(\frac{1}{2} + k/\pi\right).$$

Therefore

$$s_2^*(k) = s_4(k) = [2\,P(k) - 1]^2 \doteq 4\,k^2/\pi^2 . \tag{5.9.4}$$

Using the distribution function

$$G^*(\alpha) = \alpha/2\pi + (k \sin \alpha)/2\pi$$

corresponding to the approximate density $g^*(\alpha)$, we obtain from (5.5.5),

$$s_3^*(k) \triangleq 4\pi^2 \int_0^{2\pi} \left\{ (k \sin \alpha)/2\pi - \int_0^{2\pi} (k \sin y)/2\pi \cdot dy/2\pi \right\}^2 d\alpha/2\pi$$

(5.9.5)

$$= 4\pi^2 \cdot (k^2/8\pi^3) \int_0^{2\pi} \sin^2 \alpha \; d\alpha$$

$$= k^2/2 .$$

Similarly since the Bessel function of imaginary argument $I_n(k)$ has the expansion

$$(5.9.6) \qquad I_n(k) = \frac{(k/2)^n}{\lfloor n+1} \left\{ 1 + \frac{(k/2)^2}{1 \cdot (n+1)} + \frac{(k/2)^4}{1 \cdot 2 (n+1)(n+2)} + \cdots \right\},$$

ignoring terms involving Bessel functions of higher orders in (5.7.21),

$$(5.9.7) \qquad s_5(k) \triangleq 2[I_1(k)/I_0(k)]^2$$

$$\triangleq k^2/2 .$$

And finally, using again the density $g^*(\alpha)$, since

$$\int_0^{2\pi} e^{-2\pi g^*(\alpha)} \, d\alpha = \int_0^{2\pi} e^{-(1+k \cos \alpha)} \, d\alpha$$

we get

$$s_6^*(k) \doteq [I_o(k)e^{-1} - e^{-1}]^2/(2e^{-1} - 5e^{-2})$$

$$= [I_o(k) - 1]^2/(2e - 5)$$

$$\doteq k^4/16(2e - 5)$$

from the expansion (1.2.4) for $I_o(k)$.

The following table gives the limiting efficiencies of the six test statistics studied in this Chapter. Here $L_{X,Y}(0)$ is used to denote the limiting efficiency of the test sequence $\{X\}$ relative to the test sequence $\{Y\}$, when k is allowed to go to zero.

Table 5.1

Table of limiting efficiencies.

$L_{X,Y}(0)$ Y → / X ↓	R	W	A	V	N	U
R	1	1	1	$\pi^2/8$	$\pi^2/8$	∞
W		1	1	$\pi^2/8$	$\pi^2/8$	∞
A			1	$\pi^2/8$	$\pi^2/8$	∞
V				1	1	∞
N					1	∞
U						1

Thus the three tests due to Rayleigh, Watson and Ajne's A turn out to have the same limiting performance for testing uniformity against the CN alternatives. Kuiper's V and Ajne's N have also asymptotically identical performances but they do not fare as good as the first three. If the efficiency of R, W or A is taken to be unity, then V and N have an efficiency of 81%. That the asymptotic efficiency of the spacings test U_n is zero as compared with the other tests against the CN alternatives, is not altogether surprising in view of the results we obtained in Chapter IV. However one need not abandon the symmetric spacings tests because of this, since these limiting Bahadur efficiencies seldom throw sufficient light on the relative powers of the tests in small samples, with which most practical investigations are concerned. Hence it is not always wise to rely on these efficiencies in preferring one test to another. A modest simulation study was, therefore, undertaken to assess the small sample performance of the spacings test U_n as compared to the Rayleigh's test R^2. It may be recalled that the Rayleigh's test based on the length of the resultant, is the likelihood ratio test for testing uniformity in circular normal populations and is by far the best test for the situation.

550 samples of size 10 each were generated from the circular normal distributions (CND's) with concentration parameters $k = 1$ and $k = 3$. The Rayleigh's statistic R^2 and the spacings statistic U_n were computed for each sample and compared with the corresponding 5% and 1% critical points, obtained from Greenwood and Durand (1955) and Table 3.1 of Chapter 3. The proportion of samples, which, these two tests reject at these two levels of significance have been tabulated below.

Table 5.2

Monte Carlo powers of the tests R^2 and U_n.

CND with $k = 1$ (sample size $n = 10$)

Test statistic	Level of significance 5%	1%
Rayleigh's test, R^2	0.4073	0.1891
Spacings test, U_n	0.2636	0.0927

CND with $k = 3$ (sample size $n = 10$):

Test statistic	Level of significance 5%	1%
Rayleigh's test, R^2	0.9854	0.9636
Spacings test, U_n	0.9636	0.8200

From these comparisons, we find that the small sample power of the spacings test statistic U_n is quite satisfactory as compared to that of the Rayleigh's test, even for

observations from the CND. Besides its satisfactory small sample power, the test based on U_n, as we noted elsewhere, detects clustering of any sort and is a valid test for a much wider class of alternatives.

5.10. A simple inequality between Kuiper's V and Ajne's N

As we have remarked in Section 5.6, the slopes of Kuiper's V and Ajne's N are identical. We note in this section, that Kuipers $V_n = T_n^{(2)}$ defined in (5.4.1) is always larger than Ajne's $N_n^* = T_n^{(4)}$ defined in (5.6.2), whatever the sample. If $F_n(\alpha)$ denotes the empirical d.f. as measured from some point, define $F_n^*(\alpha)$ and $F^*(\alpha)$ on the interval $[0, 3\pi)$ as follows.

$$(5.10.1) \qquad F_n^*(\alpha) = \begin{cases} F_n(\alpha) & \text{for } 0 \le \alpha < 2\pi \\ 1 + F_n(\alpha - 2\pi) & \text{for } 2\pi \le \alpha < 3\pi \end{cases}$$

and

$$(5.10.2) \qquad F^*(\alpha) = \alpha/2\pi \qquad \text{for } 0 \le \alpha < 3\pi.$$

Then

$$N(\alpha) = \text{number of observations in } [\alpha, \alpha + \pi)$$
$$= n[F_n^*(\alpha + \pi) - F_n^*(\alpha)], \qquad \text{for } 0 \le \alpha < 2\pi$$

and therefore

$$
\begin{aligned}
N_n^* &= \sqrt{n}\,[N/n - 1/2] \\
&= \sqrt{n}\,\Big\{\sup_{0 \le \alpha < 2\pi} [F_n^*(\alpha + \pi) - F_n^*(\alpha)]\Big\} - \sqrt{n}/2 \\
&= \sqrt{n}\,\Big\{\sup_{0 \le \alpha < 2\pi} [F_n^*(\alpha+\pi) - F^*(\alpha+\pi) + F^*(\alpha) - F_n^*(\alpha)]\Big\} \\
&\le \sqrt{n}\,\Big\{\sup_{0 \le \alpha < 2\pi} [F_n^*(\alpha+\pi) - F^*(\alpha+\pi)] \\
&\qquad\qquad - \inf_{0 \le \alpha < 2\pi} [F_n^*(\alpha) - F^*(\alpha)]\Big\} \\
&= \sqrt{n}\,\Big\{\sup_{0 \le \alpha < 2\pi} [F_n(\alpha) - F(\alpha)] - \inf_{0 \le \alpha < 2\pi} [F_n(\alpha) - F(\alpha)]\Big\} \\
&= V_n.
\end{aligned}
$$

Thus, whatever sample is considered, the value of Ajne's N_n^* can not exceed that of the Kuiper's statistic V_n.

From the equivalence of the Bahadur slopes of V_n and N_n^*, we may conclude that the two statistics have asymptotic distributions, which are of the same exponential order in the tails. But since $V_n \ge N_n^*$, the limiting distributions must be quite dissimilar in their main parts. Thus, in view of the fact that the tails and main parts of the sequence of distributions do not necessarily have the same limiting properties, one might doubt the reliability of the Bahadur comparison between V_n and N_n^*. However, as Abrahamson (1967) puts it,

the Bahadur efficiency concerns itself with how well the null hypothesis explains the sequences of test statistics, when in fact, the hypothesis is false and the statistics are growing roughly in proportion to $\sqrt{n}$. Thus the fact that Kuipers V_n exceeds Ajne's N_n^* in value, assures us in view of the equality of the slopes, that V_n attains a smaller level of significance than does N_n^* and rejects the null hypothesis more often.

CHAPTER VI

SOME TWO-SAMPLE NONPARAMETRIC TESTS FOR THE CIRCLE AND AN EFFICIENCY COMPARISON

6.1 Introduction and Summary

We consider the following two-sample problem on the circle: Given $\alpha_1, \ldots, \alpha_m$ and $\beta_1, \ldots, \beta_n$, two independent samples from distribution functions $F(\alpha)$ and $G(\beta)$ respectively on $[0, 2\pi)$, we wish to test if the two populations are identical i.e., the hypothesis $H_0 : F(\alpha) = G(\alpha)$ for all α. Kuiper (1960) and Watson (1962) gave two-sample extensions of the test statistics V_n and W_n defined in (5.4.1) and (5.5.1) respectively. These extensions are again based on suitable measures of divergence between the empirical distribution functions of the two individual samples. Since there is no natural way of ranking the observations on the circle, one can not make use of the several rank tests that are available on the line with the circular data. However, Schach (1968) defines a class of non-parametric two-sample tests for the circle which are closely related to the class of rank tests for distributions on the line.

Presently we discuss here a two-sample non-parametric test, called the V^2 test, which is analogous to Dixon's (1940)

statistic for the line. It is based on the number of observations, $\{S_i\}$, of one sample in the sample arc-lengths made by the other. The usual run test for the circle can also be expressed in terms of these S_i's and we also compute the asymptotic relative efficiency (ARE) of the run test on the circle as compared with this V^2 test. This ARE has been derived earlier by Blumenthal (1963) but unfortunately his derivations made use of a result of Weiss (1957) which is not quite right, as has been pointed out by Pyke (1965). We utilise the results of our Chapter IV and the conditional approach of Blumenthal (1963) to obtain the distribution of V^2 under the alternatives of interest and show that the ARE of the run test on the circle as compared to the V^2 test is $1/(1+\theta)$ where θ is the constant ratio between the two sample sizes.

6.2 The V^2 test

Let $\alpha_1,\ldots,\alpha_m$ and $\beta_1,\ldots,\beta_n$ be two independent samples from two circular populations with d.f.'s $F(\alpha)$ and $G(\beta)$ on $[0, 2\pi)$, on the basis of which we wish to test the hypothesis $H_o : F \equiv G$. Without loss of generality, we may take $m \geq n$. If $0 \leq \beta_1' \leq \ldots \leq \beta_n' \leq 2\pi$ denote the order statistics from the second sample, define

(6.2.1) S_i = number of α's in between $[\beta'_{i-1}, \beta'_i)$, $i = 2, \ldots, n$

and $S_1 = m - \sum_{2}^{n} S_i$.

These S_i's denote the number of observations from the first sample falling in the i^{th} sample are made by observations from the second sample. For testing the hypothesis H_o, we may consider tests based on these values $\{ S_i, i = 1, \ldots, n \}$. For example, one may use the statistic

(6.2.2) $$V^2 = 1/n \sum_{1}^{n} S_i^2$$

which is equivalent to the statistic

$$c^2 = \sum_{1}^{n} (S_{i/m} - 1/n)^2$$

suggested by Dixon (1940) for the two-sample problem on the line. Following Dixon (1940), we can show that

(6.2.3) $$E(V^2) = m(n+2m-1)/n(n+1)$$

$$Var(V^2) = 4m(m-1)(n-1)(m+n)(m+n+1)/ n^2(n+1)^2(n+2)(n+3).$$

Since the values $\{ S_i \}$ remain invariant under a probability integral transformation on the original observations, it is more convenient to consider

(6.2.4) $$x_i = F(\alpha_i), \quad i = 1,\ldots, m$$
$$y_j = F(\beta_j), \quad j = 1,\ldots, n$$

which are now on the unit interval, with the X's having a uniform distribution on [0,1]. The Y's have the density

(6.2.5) $$g_1(y) = \frac{g(F^{-1}(y))}{f(F^{-1}(y))}, \quad 0 \leq y \leq 1.$$

Blum and Weiss (1957) have shown that the V^2 test is consistent against any alternative $g_1(y)$ such that

$$\int_0^1 g_1^{-1}(y)\, dy > 1.$$

They further demonstrate that this V^2 test is the locally most powerful spacings test against linear alternatives of the form

$$g_\theta(y) = 1 + \theta(y - \tfrac{1}{2}), \quad 0 < |\theta| \leq 2 .$$

These conclusions apply equally well to the V^2 test we have defined in (6.2.2) for the circle, since the distribution of the S_i's i.e., $\{S_i, \ i = 1,\ldots, n\}$ on the circle is clearly the same as those obtained on the line, when there are only $(n-1)$ observations in the second sample. If

(6.2.6) $$m, n \to \infty \text{ such that } \frac{m}{n} \to \theta > 0$$

then under the hypothesis H_o: $F \equiv G$ (or equivalently H_o: $g_1(y) \equiv 1$)

$$\sqrt{n}\,[V^2 - 9 - 2\vartheta^2] \tag{6.2.7}$$

has an asymptotic normal distribution with mean zero and variance $4\vartheta^2(1 + \vartheta)^2$ as can be demonstrated from the results of Blumenthal (1963).

6.3 Runs on the circle

The theory of the run test on the circle is well-known by now (Ref. e.g. David and Barton (1962)). The run statistic on the circle can be expressed in terms of the S_i's defined in the earlier Section (equation (6.2.1)). Let $P(S_1,\ldots,S_n)$ denote the proportion of S_i's which are zero i.e.,

$$P(S_1,\ldots,S_n) = P(S) = \frac{1}{n}\sum_{1}^{n} \vartheta(S_i) \tag{6.3.1}$$

where

$$\vartheta(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6.3.2}$$

Since the number of runs of X's, say R_X, is the same as the number of Y-spacings containing at least one X, we have

(6.3.3) $$R_X = n(1 - P(S)).$$

On the circle we always have an even number of runs, exactly double the number of runs from either sample. Therefore, if R denotes the number of runs on the circle in the combined ordered sample,

(6.3.4) $$R = 2n\,(1 - P(S)).$$

The asymptotic (in the sense of (6.2.6)) normality of the run statistic under the hypothesis H_o, has been established by Wald and Wolfowitz (1940) and we have an asymptotic normal distribution for

(6.3.5) $$\sqrt{n}\,\left(\frac{R}{n} - \frac{2\vartheta}{1+\vartheta}\right)$$

with mean zero and variance $4\vartheta^2/(1+\vartheta)^3$. The consistency of the run test against a large class of alternatives, follows again from the results of Blum and Weiss (1957).

6.4 The ARE of the run test on the circle as compared to the V^2 test

Blumenthal (1963) attempts to derive the distributions of the V^2 test and the run statistic R under general alternatives $g_1(y)$ using an elegant conditional approach. But in obtaining the distribution of V^2, the distribution of $\sum_1^n (DY_i)^2$, where DY_i is the i^{th} Y-spacing, is required under the alternatives, for which he uses a result of Weiss (1957). This result of Weiss (1957), however, is not true in that generality as has been pointed out by Pyke (1965). However, as we remarked earlier in Chapter IV, in computing the ARE, what we need is only the distribution of the statistic under a suitable sequence of alternatives which converge to the hypothesis. In this case again, the relevant alternatives turn out to be of the type

$$g_1(y) = 1 + \lambda(y)/n^{1/4}, \qquad 0 \leq y \leq 1 \tag{6.4.1}$$

which were dealt with in detail in Chapter IV. There, we obtained the distribution of $\sum_1^n (DY_i)^2$ where Y_i's have a distribution of the form (6.4.1).

For the sake of completeness, we briefly outline the conditional approach of Blumenthal (1963) which is summarised

in his Theorem 2.1 and which we reproduce below. Let $H_n(S_1,\ldots,S_n) = H_n(S)$ be any function based on $S_1,\ldots,S_n$ and let $E_n(H_n(S)|Y)$ denote the conditional expectation of $H_n(S)$ given $Y_1,\ldots,Y_n$. Then

Theorem 6.4.1 (Blumenthal): Let $H_n(S)$ and $E_n(H_n(S)|Y)$ be as defined above. If $\sqrt{n}\,[E_n(H_n(S)|Y) - E\,H_n(S)]$, considered as a function of $(Y_1,\ldots,Y_n)$, has a limiting non-degenerate normal distribution, $N(0,c_1)$ and if

$$(6.4.2)\qquad \lim_{n\to\infty} n^{p/2}\, E_n\left\{[H_n(S) - E_n(H_n(S)|Y)]^p|Y\right\} = \begin{cases} (p-1)(p-3)\cdots 3\cdot 1\cdot c_2^{p/2} & \text{if } p \text{ is even} \\ 0 & \text{if } p \text{ is odd} \end{cases}$$

with probability one (where c_2 is some constant), then

$$(6.4.3)\qquad \sqrt{n}\,[H_n(S) - E\,H_n(S)]$$

has an asymptotic normal distribution $N(0,c)$ where $c = c_1 + c_2$.

Now since our V^2 defined in (6.2.2) is

$$= \frac{m}{n} + \frac{2}{n}\sum_1^n \binom{S_i}{2},$$

it has the same limiting distribution as

(6.4.4) $$\vartheta + 2\left[\frac{1}{n}\sum_{1}^{n}\binom{S_i}{2}\right]$$

$$= \vartheta + 2\,H_n(S), \text{ say.}$$

Now consider all the $\binom{m}{2}$ pairs (X_{i_1}, X_{i_2}) of X's. Among these, the number of pairs that fall in the same Y-arc is given by the following two ways

(6.4.5) $$\sum_{1}^{n}\binom{S_i}{2} = n\,H_n(S) \ldots$$

$$= \sum_{1 \le i_1 \le i_2 \le m} t(x_{i_1}, x_{i_2})$$

where

(6.4.6) $$t(x_1,x_2) = \begin{cases} 1 & \text{if } x_1 \text{ and } x_2 \text{ fall in the same Y-arc} \\ 0 & \text{otherwise.} \end{cases}$$

Clearly

$$E(t(x_1,x_2)|Y) = P[t(x_1,x_2) = 1|Y]$$

(6.4.7) $$= \sum_{1}^{n}(DY_i)^2$$

where DY_i is the length of the i^{th} Y-arc as defined earlier. Thus from (6.4.5) and (6.4.7)

(6.4.8) $$E[H_n(S)|Y] = \frac{1}{n} \sum_{i \le i_1 \le i_2 \le m} \left[\sum_{1}^{n}(DY_i)^2\right]$$

$$= \frac{\vartheta^2}{2}\left[n\sum_{1}^{n}(DY_i)^2\right] + \delta_n$$

where $\delta_n \to 0$ stochastically. Hence the asymptotic normality of $E(H_n(S)|Y)$, under the relevant alternatives (6.4.1) is established by our Theorem 4.5.3 of Chapter IV. From the expressions (4.5.19) and (4.5.20) for the asymptotic mean and variance under these alternatives, we have

$$(6.4.9) \qquad \sqrt{n}\,[E(H_n(S)|Y) - \vartheta^2 - \frac{\vartheta^2}{\sqrt{n}}\left(\int_0^1 \lambda^2(p)\,dp\right)]$$

is asymptotically normal with mean zero and variance ϑ^4. Further, for the alternatives (6.4.1) of interest, Blumenthal's Theorem 3.1, in particular, gives

$$(6.4.10) \qquad \lim_{n\to\infty} n^{p/2} E_n\{[H_n(S) - E(H_n(S)|Y)]^p | Y\}$$

$$= \begin{cases} (p-1)(p-3)\cdots 3.1\; c_2^{p/2} & \text{if } p \text{ is even} \\ 0 & \text{if } p \text{ is odd} \end{cases}$$

where $c_2 = \vartheta^2[1+2\vartheta]$. Therefore, appealing now to Theorem (6.4.1), we have

$$(6.4.11) \qquad \sqrt{n}\,[H_n(S) - \vartheta^2 - \frac{\vartheta^2}{\sqrt{n}}\left(\int_0^1 \lambda^2(p)\,dp\right)]$$

is asymptotically normal with mean zero and variance $\vartheta^2(1+\vartheta)^2$

under the alternatives (6.4.1). Hence from the equivalence of the asymptotic distributions of V^2 and $\vartheta + 2H_n(S)$,

$$(6.4.12) \qquad \sqrt{n}\,[V^2 - \vartheta - 2\vartheta^2 - \frac{2\vartheta^2}{\sqrt{n}} \int_0^1 \chi^2(p)\,dp)]$$

is asymptotically normal with mean zero and variance $4\vartheta^2(1+\vartheta)^2$ under the alternatives (6.4.1). Now comparing the null distribution given in (6.2.7) and the distribution under the appropriate alternatives as given in (6.4.12), the 'efficacy' of the V^2 test (See e.g. Section 4.6) is

$$(6.4.13) \qquad e(V^2) = \frac{\vartheta^2}{(1+\vartheta)^2} \left(\int_0^1 \chi^2(p)\,dp\right)^2$$

On the other hand, from the results of Blumenthal (1963) the asymptotic distribution of the run statistic, R, defined in (6.3.4), can be immediately written down for the alternatives of interest, namely (6.4.1). In fact, under these alternatives,

$$(6.4.14) \qquad \sqrt{n}\,\left[\frac{R}{n} - \frac{2\vartheta}{1+\vartheta} + \frac{2\vartheta^2}{(1+\vartheta)^3} \frac{\left(\int_0^1 \chi^2(p)\,dp\right)}{\sqrt{n}}\right]$$

has an asymptotic normal distribution with mean zero and same

variance as under the hypothesis. Therefore the efficacy of the run test for testing H_o, against the alternatives of the type (6.4.1) is

$$(6.4.15) \qquad e(R) = \frac{\vartheta^2}{(1+\vartheta)^3} \left(\int_0^1 \lambda^2 (p)\, dp \right)^2$$

Thus, from (6.4.13) and (6.4.15), the ARE of the run test as compared to the V^2 test is

$$(6.4.16) \qquad e(R, V^2) = \frac{1}{(1+\vartheta)}$$

which shows that the run test is always inferior to the V^2 test. Further the ARE of the V^2 test as compared to the run test increases as the relative proportion of observations $\frac{m}{n}$ in the two samples, increases.

CHAPTER VII

LARGE SAMPLE TESTS FOR HOMOGENEITY OF SEVERAL ANGULAR POPULATIONS

7.1 Introduction

So far we dealt with single sample and two-sample problems for angular populations and discussed some test procedures, parametric as well as nonparametric. In this chapter, we deal with several sample situations and propose two large sample tests based on the 'homogeneity statistic', H of Rao (1965) for testing the equality of polar directions and for the equality of dispersions. The tests given here have already appeared in print. (Ref. Sengupta and Rao (1966)) These tests do not assume any specific circular distribution for the observations and are generally valid provided the samples are not too small. And if, in particular, the data conforms to a CN distribution given in (1.2.3), one can derive as we do in Section 7.5, the standard errors of estimates of the CN parameters through a large sample approach. We show in Section 7.3 that the limiting power of the homogeneity test, for testing equality of polar directions, coincides with that of Watson's approximate analysis of variance test, under the assumptions under which the latter is valid. We also give a numerical example to illustrate the use of these tests.

7.2 Testing for equality of polar vectors

Suppose we have $q(\geq 2)$ populations of angular variables and samples of sizes $n_1,\ldots,n_q$ respectively from these populations. Let us denote the i^{th} sample values by $\{\alpha_{ij},\ j = 1,\ldots, n_i\}$, $i = 1,\ldots, q$.

Let

$$x_{ij} = \cos\alpha_{ij}, \qquad y_{ij} = \sin\alpha_{ij}$$

(7.2.1)
$$\bar{x}_i = \Big(\sum_{j=1}^{n_i} x_{ij}\Big)/n_i, \qquad \bar{y}_i = \Big(\sum_{j=1}^{n_i} y_{ij}\Big)/n_i$$

i.e., $\bar{x}_i$ and $\bar{y}_i$ stand for the means of the 'cos' and 'sin' values respectively from the i^{th} sample. Further let $S_{xx}^{(i)}$, $S_{yy}^{(i)}$ denote the sample variances of the 'cos' and 'sin' values and $S_{xy}^{(i)}$, the sample covariance between the 'cos' and 'sin' values from the i^{th} sample. That is

$$S_{xx}^{(i)} = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2/(n_i - 1)$$

(7.2.2)
$$S_{yy}^{(i)} = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2/(n_i - 1)$$

and
$$S_{xy}^{(i)} = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)/(n_i - 1)$$

where $\bar{x}_i$ and $\bar{y}_i$ are as given in (7.2.1). Now if γ_i

denotes the polar direction of the i^{th} population, we know that a consistent estimator (Ref. Section 2.1) for tan γ_i is provided by the statistic

$$T_i = \bar{y}_i / \bar{x}_i \tag{7.2.3}$$

which has an asymptotic estimated variance

$$s_i^2 = (1/n_i) \left\{ S^{(i)}_{yy/\bar{x}_i^2} + \bar{y}_i^2 \, S^{(i)}_{xx/\bar{x}_i^4} - 2\bar{y}_i \, S^{(i)}_{xy/\bar{x}_i^3} \right\} \tag{7.2.4}$$

in the notations given in (7.2.1) and (7.2.2). This can be easily shown by the method of differencing (See e.g. Rao (1965)). Further the quantity $\sqrt{n_i}\,(T_i - \tan \gamma_i)$ is normally distributed in large samples. Now we consider the null hypothesis

$$H_o: \tan \gamma_1 = \tan \gamma_2 = \dots = \tan \gamma_q. \tag{7.2.5}$$

Since $T_1, \dots, T_q$ are independent, consistent estimators of the same quantity under the hypothesis H_o, based on samples of sizes $n_1, \dots, n_q$ with variances s_i^2 given by (7.2.4), we can use the statistic

$$H = \sum_{i=1}^{q} T_i^2/s_i^2 - \Big(\sum_{1}^{q} T_i/s_i^2 \Big)^2 \Big/ \Big(\sum_{1}^{q} 1/s_i^2 \Big). \tag{7.2.6}$$

This statistic is appropriate for testing H_o since if each T_i does not estimate the same quantity, the differences are reflected in the test-statistic H given above. Under certain very general conditions, H has an asymptotic χ^2 distribution with (q-1) degrees of freedom under the null hypothesis (See Rao (1965)).

Rejection of H_o leads us to conclude that the polar vectors are different. But on the other hand, even if H_o is not rejected, it is possible that the mean directions of the q populations are different since our hypothesis does not distinguish between the 'pole' and the 'antipole', tan γ being the same as tan $(\pi + \gamma)$. But such wide differences in the polar directions can easily be found out by a simple examination of the data. The procedure for applying the homogeneity test is simple and consists in getting T_i from (7.2.3), s_i^2 from (7.2.4) and then computing H using the formula (7.2.6).

7.3 Comparison of the asymptotic powers of the homogeneity test and the analysis of variance test

For testing the equality of polar directions of several CN populations, Watson (1956, 1966) gave an approximate analysis of variance (ANOVA) test which depends on the lengths of the resultants, $R_1, \ldots, R_q$ of the q independent samples

and on R, the length of the overall resultant based on $n = \sum_{1}^{q} n_i$ observations. The test statistic is

$$F = \frac{(\sum_{1}^{q} R_i - R)/(q-1)}{(n - \sum_{1}^{q} R_i)/(n-q)} \tag{7.3.1}$$

which has a F distribution with $(q-1)$ and $(n-q)$ degrees of freedom, provided the q populations have the same concentration parameter say k, which is high. For large k, this corresponds to the linear ANOVA statistic and can be derived essentially from the approximate distributions (2.2.8). In this section, we show that the homogeneity test statistic (7.2.6), besides being valid without any restrictions on the concentrations of the different populations, is asymptotically as efficient as the F test given in (7.3.1), for testing the equality of polar directions.

As the F-test (7.3.1) corresponds to the linear ANOVA test for large k, we show that the homogeneity test (7.2.6), has asymptotically the same power as the linear ANOVA test. It may be recalled that when k is large, the CND becomes essentially normal. We therefore consider the following ANOVA set up: Suppose we have q normal populations and a sample of size n_i from the i^{th} population. Let α_{ij} denote the j^{th} observation from the i^{th} population

$(j = 1,\ldots,n_i,\ i = 1,\ldots,\ q)$. If the i^{th} population is normally distributed with mean γ_i and variance σ^2, then under the hypothesis

$$(7.3.2) \qquad H_o : \gamma_1 = \gamma_2 = \ldots = \gamma_q$$

of equality of means, the statistic

$$(7.3.3) \qquad S = \frac{\sum_i n_i(\bar{\alpha}_{i.} - \bar{\alpha}_{..})^2/(q-1)}{\sum_i \sum_j (\alpha_{ij} - \bar{\alpha}_{i.})^2/(n-q)}$$

has a F-distribution with $(q-1)$ and $(n-q)$ degrees of freedom. (The notations in (7.3.3) are standard and need no explanation). Under the alternative, i.e., when the γ_i s are not all equal, the statistic S, given in (7.3.3), has a noncentral F-distribution, $F(q-1,\ n-q,\ \lambda_1)$ where the noncentrality parameter λ_1 is given by

$$(7.3.4) \qquad \lambda_1 = \sum_i n_i(\gamma_i - \bar{\gamma})^2/\sigma^2 \quad \text{where} \quad \bar{\gamma} = \sum_i n_i \gamma_i/n.$$

However, since our interest lies in making an asymptotic power comparison, we consider the following limiting situation

$$(7.3.5) \qquad n_i, n \to \infty \ \text{such that} \ n_{i/n} \to \pi_i,$$

$$0 < \pi_i < 1,\ i = 1,\ldots,\ q.$$

Now since any reasonable test turns out to be consistent (in the sense, the power tends to unity as the sample size increases indefinitely) against any fixed alternative, we consider the following sequence of alternatives converging to the null hypothesis in the limit.

(7.3.6) Alt.: The i^{th} population mean $\gamma_i = \gamma + \delta_i/\sqrt{n}$, $i = 1, \ldots, q$

where $\gamma, \delta_1, \ldots, \delta_q$ are some fixed constants. Then under the asymptotic situation considered in (7.3.5), the statistic S has a $\chi^2_{q-1}/(q-1)$ distribution under H_o, while it is a non-central $\chi^2_{q-1}(\lambda_1)/(q-1)$ under the alternative (7.3.6). Here, in the limiting situation, this non-centrality parameter λ_1, turns out to be

(7.3.7) $\lambda_1 = \sum_i \pi_i (\delta_i - \bar{\delta})^2/\sigma^2$ where $\bar{\delta} = \sum_i \pi_i \delta_i$.

On the other hand, the homogeneity statistic H of Section 7.2, requires the following simpler set up: $T_1, \ldots, T_q$ are independent consistent estimators of $\gamma_1, \ldots, \gamma_q$ based on samples of sizes $n_1, \ldots, n_q$ and $\sqrt{n_i}\,(T_i - \gamma_i)$ is asymptotically normal with mean zero and variance $\sigma_i^2(\gamma_i)$. Then under H_o given in (7.3.2).

$$(7.3.8) \qquad H = \sum_i n_i (T_i - \hat{\gamma})^2 / \sigma_i^2(T_i)$$

where

$$\hat{\gamma} = \sum_i \frac{n_i T_i}{\sigma_i^2(T_i)} \Big/ \sum_i \frac{n_i}{\sigma_i^2(T_i)}$$

is asymptotically (i.e., in the sense of (7.3.5)) distributed as a χ^2_{q-1} (See Rao (1965)). By exactly similar arguments, it is easy to show that under the alternatives mentioned in (7.3.6), H has again a noncentral $\chi^2_{q-1}(\lambda_2)$ with the non-centrality parameter λ_2, say. Writing

$$(7.3.9) \qquad \gamma^* = \sum_1^q \frac{\pi_i \gamma_i}{\sigma_i^2(\gamma_i)} \Big/ \sum_1^q \frac{\pi_i}{\sigma_i^2(\gamma_i)}$$

$$= \gamma + \Big[\sum_i \frac{\pi_i \delta_i}{\sigma_i^2(\gamma_i)} \Big/ \sum_i \frac{\pi_i}{\sigma_i^2(\gamma_i)} \Big] / \sqrt{n}$$

$$= \gamma + \delta^* / \sqrt{n}, \text{ say}$$

the non-centrality parameter, λ_2, can be seen to be

$$(7.3.10) \qquad \lambda_2 = \sum_i \pi_i (\delta_i - \delta^*)^2 / \sigma_i^2(\gamma).$$

Now, clearly, one should consider the situation where $\sigma_1^2(\gamma) = \ldots = \sigma_q^2(\gamma) = \sigma^2$ (i.e., equal variances in all the populations), in order to compare the asymptotic performance

of H with that of the ANOVA statistic S. In this case, the non-centrality parameter λ_2 given in (7.3.10) is identically the same as λ_1, given in (7.3.7) so that both the tests have asymptotically the same power. Thus when the assumptions of the ANOVA set-up are satisfied, the limiting efficiency of the homogeneity test, (7.2.6), equals that of the ANOVA test (7.3.3) which corresponds to the approximate statistic (7.3.1) for large k. On the other hand, the homogeneity test, (7.2.6), may still be used in large samples even if the conditions of the ANOVA set up do not hold. But however, when dealing with small samples, only Watson's F-test given in (7.3.1), is available for testing the equality of polar directions.

7.4 Testing for equality of dispersions

We shall consider the same set-up as in Section 7.2 and also retain the notations (7.2.1) and (7.2.2). Now consider the statistic

$$U_i = (\bar{x}_i^2 + \bar{y}_i^2) = R_i^2 / n_i^2 \tag{7.4.1}$$

where R_i denotes the length of the resultant for the i^{th} sample of n_i observations. This U_i is a measure of concentration for any unimodal circular population as we have noted in Section 2.2. Large values of U_i i.e., values near

unity show high concentration of the observations and values of U_i near zero indicate a low concentration or a high amount of dispersion in the i^{th} population. It is easy to check that the estimated asymptotic variance of this statistic U_i, in terms of $S_{xx}^{(i)}$, $S_{xy}^{(i)}$ and $S_{yy}^{(i)}$ is

$$(7.4.2) \qquad s_i^{*2} = (4/n_i)\{ \bar{x}_i^2 S_{xx}^{(i)} + \bar{y}_i^2 S_{yy}^{(i)} + 2\bar{x}_i\bar{y}_i S_{xy}^{(i)} \} .$$

If now the null hypothesis stipulates the equality of the concentrations of the q populations, then all these statistics $U_1, \ldots, U_q$ are independent consistent estimators of the same quantity under the null hypothesis. Hence we can use, again, the statistic H for homogeneity and compute

$$(7.4.3) \qquad H^* = \sum_1^q U_i^2 / s_i^{*2} - (\sum_1^q U_i/s_i^{*2})^2/(\sum_1^q 1/s_i^{*2})$$

which is distributed asymptotically as a χ^2 distribution with $(q-1)$ degrees of freedom under the hypothesis of equality of dispersions. Thus a test for the homogeneity regarding the dispersions of the q populations may be made in large samples, using the test-statistic H^*, given in (7.4.3).

7.5 Standard errors of estimates in circular normal populations

In this section, we assume that the basic distribution underlying these observations is CN with density given by (1.2.3) so that the maximum likelihood method of estimating the parameters γ and k as described in Section 2.1 holds. Thus we have the ML estimate

$$\hat{\gamma}_i = \mathrm{Tan}^{-1} T_i \tag{7.5.1}$$

for the polar direction γ_i of the i^{th} population. From this one gets, by the usual differencing method, the standard error (s.e.) of $\hat{\gamma}_i$ as

$$\hat{\text{s.e.}}\ (\hat{\gamma}_i) = s_i/(1+T_i^2) \tag{7.5.2}$$

where s_i is given by (7.2.4). Similarly we can get the s.e. of $\hat{k}_i$, the ML estimate of the concentration parameter k_i of the i^{th} population. Since the ML equation for estimating k_i is

$$I_1(\hat{k}_i)'/I_0(\hat{k}_i) = R_i/n_i = \sqrt{U_i}\ , \tag{7.5.3}$$

differencing method, gives

$$(7.5.4) \qquad \hat{s.e.}(\hat{k}_i) = \frac{s_i^*}{2U_i\left(\frac{1}{\sqrt{U_i}} - \frac{1}{\hat{k}_i} - \sqrt{U_i}\right)}$$

where s_i^* is as given in (7.4.2). Thus, in fairly large samples, one can get the standard errors of the CN parameters of the i^{th} population from equations (7.5.2) and (7.5.4), when a CN distribution can be assumed for the observations.

7.6 Comments on the robustness of the test procedures and a numerical example

The tests described in Sections 7.2 and 7.4 do not assume any specific distribution for the observations and are valid whatever the underlying unimodal distribution. Watson's approximate F-test for testing the equality of the polar vectors, given in (7.3.1) has also been shown to be robust for deviations from circular normality of the observations, as one should expect from the linear case of analysis of variance. But however, the test holds good when the concentrations of the several populations under comparison, are high and equal. The statistic H given in (7.2.6) for testing the equality of polar vectors, is not based on any such assumptions regarding the concentrations and is valid more generally provided the samples are large.

In the following numerical example from Sengupta and Rao (1966), we have three populations referred to as upper, middle and lower Kamthi rock formations. The CN distribution fails to give a good fit to any of these three populations. Further the concentrations are quite low and they are also significantly different as tested by the statistic H^* given in (7.4.3). In spite of these many odds, the homogeneity tests H and H^* for testing the equality of polar directions and concentrations hold good as our samples are quite large. The sample data is presented below in Table 7.1 in a grouped form.

TABLE 7.1

Frequency distributions of cross-bedding azimuths in the three Kamthi units.

Azimuth class interval (in degrees)	Upper Kamthi	Middle Kamthi	Lower Kamthi
0 - 19	75	50	14
20 - 39	75	62	14
40 - 59	15	33	11
60 - 79	25	9	13
80 - 99	7	1	9
100 - 119	3	3	16
120 - 139	3	-	-
140 - 159	-	-	4
160 - 179	-	-	-
180 - 199	-	-	3
200 - 219	21	2	4
220 - 239	8	8	-
240 - 259	24	-	-
260 - 279	16	11	-
280 - 299	36	5	6
300 - 319	75	20	7
320 - 339	90	53	1
340 - 359	107	41	21
	$n_1 = 580$	$n_2 = 298$	$n_3 = 123$

On the basis of this data, we want to test whether the three Kamthi populations have the same dispersion and whether they differ significantly with regard to their mean directions. Some of the computations are given in Table 7.2 below.

TABLE 7.2

Some computational details.

Kamthi formation	n_i	$\bar{x}_i$	$\bar{y}_i$	T_i	s_i	U_i	s_i^*
Upper	580	0.6299	-0.1993	-0.3164	0.0323	0.4365	0.0322
Middle	298	0.7389	0.0119	0.0161	0.0440	0.5462	0.0318
Lower	123	0.4106	0.3090	0.7524	0.1863	0.2640	0.0487

The computed values of H and H* as given by (7.2.6) and (7.4.3) turn out to be very much larger than the 5% critical value, viz. 5.99, thus showing significant differences between the Kamthi populations with regard to both the mean directions and dispersions. Thus the polar azimuthal angles in the three Kamthi rock formations differ significantly and further the dispersions in the three populations can not also be considered equal.

And finally, assuming a CN distribution for the observations, we find the ML estimates $\hat{\gamma}_i$ and $\hat{k}_i$ for the polar direction and concentration of the i^{th} Kamthi unit. We also give below, in Table 7.3, the large sample standard errors of these estimates, obtained from (7.5.2) and (7.5.4).

TABLE 7.3

ML estimates $\hat{\gamma}_i$ and $\hat{k}_i$ and their standard errors.

Kamthi formation	$\hat{\gamma}_i$	$\hat{s.e}(\hat{\gamma}_i)$	$\hat{k}_i$	$\hat{s.e}(\hat{k}_i)$
Upper .	$342^{\circ}27'$	$6^{\circ}49'$	1.7895	0.1254
Middle.	$0^{\circ}55'$	$2^{\circ}31'$	2.2893	0.1643
Lower	$86^{\circ}06'$	$1^{\circ}41'$	1.1910	0.1556

B I B L I O G R A P H Y

Abrahamson, I. G. (1967): Exact Bahadur efficiencies for the Kolmogorov-Smirnov and Kuiper one and two-sample statistics, Ann. Math. Statist., 38, 1475-1490.

Ajne, B. (1966): A simple test for uniformity of a circular distribution, Forskningsrapport, Inst. f. Forsakringsmatemahk O. Matematish Statisk, Stockholm.

Bahadur, R. R. (1960a): On the asymptotic efficiency of tests and estimates, Sankhyā, 22, 229-252.

Bahadur, R. R. (1960b): Stochastic comparison of tests, Ann. Math. Statist., 31, 276-295.

Batschelet, E. (1965): Statistical methods for the analysis of problems in animal orientation and certain biological rhythms: AIBS monograph.

Beran, R. J. W. (1968): The power of some tests for uniformity of a circular distribution, Tech. Report No.78, Johns Hopkins University.

Blum, J. R. and Weiss, L. (1957): Consistency of certain two-sample tests, Ann. Math. Statist., 28, 242-246.

Blumenthal, S. (1963): The asymptotic normality of two test statistics associated with the two-sample problem, Ann. Math. Statist., 34, 1513-1523.

Chentsov, N. N. (1956): Weak convergence of stochastic processes whose trajectories have no discontinuities of the second kind and the heuristic approach to the Kolmogorov-Smirnov tests, Theor. of Prob. and its Appl. 1, 140-144.

Chernoff, H. (1952): A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, Ann. Math. Statist., 23, 493-507.

Cibisov, D. M. (1961): On the tests of fit based on sample spacings, Teoria Vero. i. Prim., 6, 354-358.

Darling, D. A. (1953): On a class of problems related to the random division of an interval, Ann. Math. Statist., 24, 239-253.

David, F. N. and Barton, D. E. (1962): Combinatorial chance, Hafner Publishing Co., New York.

Dixon, W. J. (1940): A criterion for testing the hypothesis that two samples are from the same population, Ann. Math. Statist., 11, 199-204.

Feller, W. (1966): Introduction to probability theory and its applications, Vol. II, John Wiley, New York.

Fisher, R. A. (1929): Tests of significance in harmonic analysis, Proc. Roy. Soc. (Ser. A), 125,54-59.

Fisz, M. (1963): Probability theory and mathematical statistics, Third edition, John Wiley, New York.

Fraser, D. A. S. (1957): Nonparametric methods in Statistics, John Wiley, New York.

Gleser, L. J. (1964): On a measure of test efficiency proposed by R. R. Bahadur, Ann. Math. Statist., 35, 1537-1544.

Goodman, N. R. (1963): Statistical Analysis based on a certain multivariate complex Gaussian distribution (An introduction), Ann. Math. Statist., 34, 152-177.

Greenwood, Major (1946): The statistical study of infectious diseases, Jour. Roy. Stat. Soc., 109, 85-103.

Greenwood, J.A. and Durand, D. (1955): The distribution of the length and components of the sum of n random unit vectors, Ann. Math. Statist., 26 233-246.

Gumbel, E. J., Greenwood, J. A. and Durand, D. (1953): The circular normal distribution: Theory and tables, J. Am. Stat. Assoc., 48, 131-152.

Hodges, J. L. (1955): A bivariate sign test, Ann. Math. Statist., 26, 523-527.

Jackson, O. A. Y. (1967): An analysis of departures from the exponential distribution, Jour. Roy. Stat. Soc. (Ser. B), 29, 540-549.

Kimball, B. F. (1950): On the asymptotic distribution of the sum of powers of unit frequency differences, Ann. Math. Statist., 21, 263-271.

Kuiper, N. H. (1960): Tests concerning random points on a circle, Ned. Akad. Watensch. Proc. (Ser. A), 63, 38-47.

Mardia, K. V. (1967): A non-parametric test for the bivariate two-sample location problem, Jour. Roy. Stat. Soc. (Ser. B), 29, 320-342.

Parzen, E. (1954): On uniform convergence of families of sequences of random variables, Univ. of California Pubs. in statist., 2, 23-54, Univ. of California Press.

Pincus, H. J. (1956): Some vector and arithmetic operations on two-dimensional orientation variates with applications to geological data, J. Geol., 64, 533-557.

Proschan, F. and Pyke, R. (1965): Tests for monotone failure rate, Proc. Fifth Berkeley Symp. Math. Statist. and Prob., Vol. III, 293-312.

Pyke, R. (1965): Spacings, Jour. Roy. Stat. Soc. (Ser. B), 27 395-449.

Rao, C. R. (1942): On the volume of a prismoid in n-space and some problems in continuous probability, Math. Student, 10, 68-74.

---------- (1961): A study of large sample test criteria through properties of efficient estimates, Sankhyā (Ser. A), 23, 25-40.

Rao, C. R. (1965): Linear statistical inference and its applications, John Wiley, New York.

Rao, J. S. and Sengupta, S. (1966): Statistical analysis of cross-bedding azimuths from the Kamthi formation around Bheemaram, Pranhita-Godavari valley, Sankhyā (Ser. B), 28, 165-174.

Rayleigh, Lord (Strutt, J. W.) (1919): On the problem of random vibrations and random flights in one, two and three dimensions, Phil. Mag., 37, 321-347.

Reiche, P. (1938): An analysis of cross-lamination of the Coconino sandstone, J. Geol., 46, 905-932.

Schach, S. (1968): On a class of nonparametric two-sample tests for circular distributions, Tech. Rep. No. 9, Stanford University.

Sethuraman, J. (1964): On the probability of large deviations of families of sample means, Ann. Math. Statist., 35, 1304-1316.

---------- (1965): Limit theorems for stochastic processes, Tech. Report No. 10, Stanford University.

Sherman, B. (1950): A random variable related to the spacing of sample values, Ann. Math. Statist., 21, 339-361.

---------- (1957): Percentiles of the ω_n statistics, Ann. Math. Statist., 28, 259-261.

Skorohod, A. V. (1956): Limit theorems for stochastic processes, Theor. of Prob. and its Appl., 1, 261-290.

Stephens, M. A. (1962): Exact and approximate tests for directions I, Biometrika, 49, 463-477.

Stevens, W. L. (1939): Solution to a geometrical problem in probability, Ann. Eugenics, 9, 315-320.

Uspensky, J. V. (1937): Introduction to mathematical probability, McGraw-Hill, New York.

von Mises, R. (1918): Uber die 'Ganzzahligkeit' der Atomgewichte und verwandte Fragen, Physikal. Z. 19, 490-500.

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

Watson, G. N. (1944): A treatise on the theory of Bessel functions, Cambridge Univ. Press, London.

Watson, G. S. (1956): Analysis of dispersion on a sphere, Monthly notices Roy. Astronomical Soc., Geophys. Suppl., 7, 153-159.

---------- (1961): Goodness of fit tests on a circle, Biometrika, 48, 109-114.

---------- (1962): Goodness of fit tests on a circle II, Biometrika, 49, 57-63.

---------- (1966): The statistics of orientation data, J. Geol., 74, 786-797.

---------- (1967): Another test for the uniformity of a circular distribution, Biometrika, 54, 675-677.

Watson, G. S. and Williams, E. J. (1956): On the construction of significance tests on the circle and the sphere, Biometrika, 43, 344-352.

Weiss, L. (1957): The asymptotic power of certain tests of fit based on sample spacings, Ann. Math. Statist., 28, 783-786.

--------- (1962): Review of Cibisov (1961), Math. Rev., 24, No. A 1776.

----------(1965): On asymptotic sampling theory for distributions approaching the uniform distribution, Z. Wahrscheinlichkeitstheorie Verw. Geb., 4, 217-221.

Wheeler, S. and Watson, G. S. (1964): A distribution-free two-sample test on a circle, Biometrika, 51, 256-257.