RASHI 1 (2) :77-88 (2016)

Goodness-of-fit Tests for the Kappa Statistic

Ayanendranath Basu, Sumit Gupta and [1]Wen-Tao Huang

Applied Statistics Unit, Indian Statistical Institute
203 B. T. Road, Calcutta 700 108, India
[1]Department of Management Sciences and Decision Making,
Tamkang University, Tamsui, New Taipei City, Taiwan, ROC

ABSTRACT

We consider the goodness-of-fit approach for performing tests of significance about the kappa statistic. This approach was suggested by Donner and Eliasziw (1992), who framed the parametric hypothesis testing problem about the kappa statistic (leading to other associated small sample inference) as a multinomial goodness-of-fit testing problem, for which they proposed the use of the Pearson's chi-square statistic. Basu and Basu (1995) considered the use of some other members of the power divergence family of Cressie and Read (1984) for this goodness-of-fit testing. However, when a specific alternative is of interest, all the standard goodness-of-fit tests can be quite poor in comparison with the most powerful test within the above family; in addition the asymptotic chi-square approximation for these statistics may be quite inadequate for small samples. In this paper we present the results of a general investigation based on exact power computations which helps us to search for the most powerful test for each sample size given the appropriate parameters, and identifies the randomized exact critical values. The SPLUS code for the exact power computation is included in the appendix.

***Keywords:** Goodness-of-fit test, kappa statistic, power divergence family.*

Introduction

The kappa statistic introduced by Cohen (1960) is widely used in medical, biological and psychiatric studies for measuring agreement between two raters in the presence or absence of a trait in individuals. Let Π_o denote the probability that the two raters agree on a randomly selected subject; under perfect agreement we will get $\Pi_o = 1$. Suppose Π_e represents the probability that the agreement between the two raters is by chance alone. The kappa statistic, which is defined as

$$\kappa = \frac{\Pi_o - \Pi e}{1 - \Pi e} ,$$

may then be interpreted as the excess observer agreement over that expected by chance. The value of the kappa statistic (henceforth κ) is bounded above by 1. Technically κ can fall below zero when Πo is less than Πe. In most applications that is an inconceivable situation and for our purpose we will consider $0 \leq \kappa \leq 1$.

The popularity of the kappa statistic in applied research as a measure of inter-rater agreement has led to a significant amount of research on the asymptotic properties of κ ; see, for example, Fleiss (1981) for the large sample normal distribution of the sample measure $\hat{\kappa}$ under multinomial sampling. However, the literature related to small sample inference about κ is limited. Donner and Eliasziw (1992) discussed a method for constructing accurate confidence limits and tests for significance for κ in small samples based on the Pearson's chi-square statistic. Basu and Basu (1995) compared the results based on the Pearson's chi-square with the likelihood ratio chi-square and another chi-square measure proposed by Cressie and Read (1984). Altaye, Donner, and Klar (2001) extended the application of the goodness-of-fit testing approach of Donner and Eliasziw (1992) to the case of binary outcome data with multiple raters. Also see Klar *et al.* (2000) and Gonin *et al.* (2000) for some alternative approaches in the context of small sample inference about κ.

In this paper we have restricted ourselves to the simple – but widely realized in practice – scenario, namely the problem of inter-rater agreement for two raters in the context of a binary response. We attempt to provide a comprehensive solution to the hypothesis testing problem with the goodness-of-fit approach; in particular we develop the codes in SPLUS which help determine the exact most powerful goodness-of-fit test against a specific alternative, as well as the small sample critical values of the desired test. Thus when one wishes to have high power against a specific alternative, one is able to determine both the optimal test for the above problem, as well as the critical values to perform that test.

We propose to consider more complicated situations, such as testing of hypothesis about κ in case of multi-

E-mail: ayanbasu@isical.ac.in

rater agreement, in a sequel paper. In the process we hope to extend the work of Altaye, Donner and Eliasziw (2001), and Altaye, Donner, and Klar (2001).

In section 2 we have introduced the model under which the hypothesis testing scenario is developed, and the goodness-of-fit tests are introduced in section 3. section 4 considers testing of hypothesis about κ, introduces the exact tests and discusses the generation of the exact power, and investigates various aspects of the testing problem to gain further insight. The suggested recipe, based on the codes developed is presented in section 5, while section 6 has some concluding remarks. We present the actual codes in the appendix.

The Model

Consider two raters who independently rate n subjects on a binary scale, the ratings being either success or failure. We will let π denote the probability that the rating of the jth subject, $j = 1, 2, ..., n$, by the kth rater, $k = 1, 2$, is a success. Then the probabilities for the joint responses can be expressed as a function of π and κ as

Pr (both ratings are successes)

$$= \pi^2 + \pi(1-\pi)\,\kappa = EP_1$$

Pr (one rating is success and one failure)

$$= 2\pi(1-\pi)(1-\kappa) = \mathrm{EP}_2 \qquad (1)$$

Pr (both ratings are failures)

$$= (1-\pi)^2 + \pi(1-\pi)\,\kappa = EP_3$$

This is the common correlation model for dichotomous data. Let n_1, n_2, and n_3 represent, respectively, the number of subjects rated as successes by both raters, the number of subjects rated as successes by exactly one rater, and the number of subjects rated as failures by both raters, where $n = n_1 + n_2 + n_3$. The vector (n_1, n_2, n_3) is the response vector. Under the common correlation model the maximum likelihood estimator of κ is

$$\hat{\kappa} = 1 - \frac{n_2}{2n\hat{\pi}\left(1-\hat{\pi}\right)} \qquad (2)$$

Where

$$\hat{\pi} = \frac{2n_1 + n_2}{2n} \qquad (3)$$

is the maximum likelihood estimator of π.

Multinomial Goodness-of-fit tests

Suppose that $(n_1, ..., n_K)$ follow a multinomial distribution on K cells with parameters $n = n_1 + ... + n_K$, and probability vector $\pi = (\pi_1, \pi_2, ..., \pi_K)$. The most popular statistic for testing a simple or composite null hypothesis about the probability vector π is the Pearson's chi-square

$$X_P^2 = 2n\sum_{i=1}^{K}\frac{[OP_i - EP_i]^2}{2EP_i}$$

where OP_i represents the observed proportion in the ith cell and EP_i denotes its expected proportion under the null hypothesis. In the specific case of the common correlation model of the kappa statistic the expected proportions EP_i are as in equation (1) with $K = 3$. The expected proportions are replaced by the estimated expected proportions under complex nulls which do not completely specify the probability vector. Another popular test statistic with the same asymptotic distribution under the null hypothesis is the likelihood ratio chi-square, defined by

$$X_L^2 = 2n\sum_{i=1}^{K} OP_i\,[1n(OP_i) - 1n(EP_i)]$$

The power divergence family of Cressie and Read (1984), indexed by a parameter λ is defined as

$$X_\lambda^2 = \frac{2n}{\lambda(1+\lambda)}\sum_{i=1}^{K} OP_i\left[\left(\frac{OP_i}{EP_i}\right)^{\lambda} - 1\right], -\infty < \lambda < \infty \qquad (4)$$

See also Read and Cressie (1988). The Pearson's chi-square is recovered for $\lambda = 1$, while the likelihood ratio chi-square corresponds to $\lambda = 0$ defined via the continuous limit of the quantity in the right hand side of (4) as $\lambda \to 0$. Many other well known goodness of-fit statistics – such as the Neyman's modified chi-square statistic, the modified log likelihood ratio statistic, and the Freeman-Tukey statistic are also members of the power divergence family for different values of λ. All members of the power divergence family have asymptotic chi-square distributions with $K - 1$ degrees of freedom under the simple null hypothesis (where the entire probability vector is specified). For complex nulls the statistics all have the same asymptotic chi-square distributions (under the null) with appropriate degrees of freedom, when the unknown parameters are replaced by first order efficient estimates in the expressions of the expected frequencies. Some members of the power divergence family have certain optimality properties under some specific contexts; see Cressie and Read (1984) for a discussion on the optimality of the likelihood ratio chi-square in terms of the maximal Bahadur efficiency, and the Pearson's chi-square in terms of maximum Pitman asymptotic relative efficiency under certain conditions.

In the case of a symmetric null hypothesis (where the probability of each cell equals $1/K$) the exact power these tests increase with λ for a 'bump' alternative and decrease with λ for a 'dip' alternative (see Cressie and Read, 1984). A bump alternative violates the null through a single cell with a large probability, while the other cells have a common, smaller probability; the dip alternative is the reverse. Even when one leaves aside the question of systematically explaining such phenomena, it demonstrates that the optimal test within the Cressie-Read family for a particular null hypothesis may depend on the specific nature of the alternative.

Testing Hypothesis about κ

One can frame the parametric hypothesis testing problem

$$H_0 : \kappa = \kappa_0 \tag{5}$$

as a goodness-of-fit testing problem as follows. Given the response vector (n_1, n_2, n_3), one can compute the estimated expected proportions EP_i, i = 1, 2, 3, by substituting κ_0 for κ and $\hat{\pi}$ for π in (1). To be specific, we will consider κ_0 to be strictly between 0 and 1. The test statistic for testing the null hypothesis about κ can then be computed for any given value of λ using the formula (4). Donner and Eliasziw (1992) had suggested the use of χ_1^2 (Pearson's chi-square), while Basu and Basu suggested that $\chi_{2/3}^2$ can be a better statistic in certain scenarios. It follows from standard asymptotic results that all the test statistics are asymptotically distributed as chi-squares with one degree of freedom.

Defining the statistics requires that the expected probabilities be strictly positive. This fails when $\hat{\pi}$ = 0 or $\hat{\pi}$ = 1 (*i.e.* when $n = n_1$ or $n = n_3$). In the exact small sample calculations that we perform later, the powers are therefore determined after conditioning on $0 < \hat{\pi} < 1$. In practice this makes little difference since here we only discard the two samples (n, 0, 0) and (0, 0, n), whose probabilities are generally negligible, even for fairly small sample sizes.

Randomized Tests of Exact Size α

Given a sample size n and the null value κ_0, we determine the exact power of the test statistic χ_λ^2 at a given level of significance α and success probability π at specific alternative values of κ as follows. First we enumerate all possible samples (n_1, n_2, n_3) with $n_1 + n_2 + n_3 = n$. There are $\binom{n+2}{n}$ such samples in all. Having enumerated them, we first discard the two samples (n, 0, 0) and (0, 0, n) from the total pool of samples. The probabilities for each remaining sample can then be determined under the null using the probability vector given by (1), conditioning on the reduced sample space after discarding the above two samples. By sorting the test statistics and cumulating their corresponding probabilities over the samples we can easily find the exact small sample critical value for this statistic at any given α; this is done by noting the point where the cumulative probability crosses $1 - \alpha$. Because of the discrete nature of the critical region we have to randomize at the critical value to get tests of exact size α. We then only need to recompute the probabilities under the alternative value of κ to determine the power of the test.

The range of λ values

Prior to any discussion that attempts to find a best test within a certain class, we need to define the class properly. In an absolute sense the power-divergence statistics are indexed by $\lambda \in (-\infty, \infty)$. This interval is clearly inconvenient to perform any search for an optimal test, but can perhaps be reduced following the observation of Cressie and Read (1984), who noted that there is an evident "plateau" effect as $|\lambda|$ increases in that for large $|\lambda|$ there is little change in power as λ varies. All the well known chi-square goodness-of-fit tests correspond to relatively small values of $|\lambda|$. There are other problems for large $|\lambda|$ statistics as well. The chi-square approximation becomes poorer as $|\lambda|$ increases. The process of finding the exact power, which requires complete enumeration of the sample space, must end at a certain reasonable value of the sample size. Beyond this, the exact power computation should give way to the large sample chi-square approximation which would hopefully be fairly accurate by that stage. This becomes difficult for large values of $|\lambda|$ which may require extremely large sample sizes for the above to hold. An additional complication is the presence of empty cells. The observed values show up in the denominator of the statistics for $\lambda < -1$, and because of the presence of the empty cells the moments of the test statistics do not exist for $\lambda \leq -1$. While we can still determine the exact powers and critical values for such values of λ in the spirit of Cressie and Read, other analysis which require moment calculations are not possible with such statistics. As a compromise choice we suggest the interval (–1, 1] for λ over which to look for an optimal test, although the SPLUS codes presented here can be easily modified to perfrom the level α test accurately for any value of λ, provided the critical value for that (λ, α) combination is finite.

For the rest of the paper, when we talk about choosing the best test within the Cressie-Read family, we will implicitly mean the tests within the Cressie-Read family with λ restricted to (–1, 1].

Use of the χ^2 critical values

Donner and Eliasziw (1992), who originally suggested the goodness-of-fit approach for the κ testing problem considered the Pearson's chi-square test only, and based their decisions on the asymptotic critical values of the statistic which correspond to the quantiles of a χ^2 distribution with degrees of freedom 1. Basu and Basu (1995) considered the exact computation of power in the context of some specific members of the Cressie-Read family. In actual calculations, the true small sample critical values of any of these goodness of-fit test statistics within the Cressie-Read family (4) can be widely different from the χ^2 critical values. In table 1 we provide one example of the actual small sample critical values for the χ^2 goodness-of-fit test statistics at different values of κ_0 and for $\lambda = 1$, 0 and – 0.5. The entries in the table give the exact critical values for level $\alpha = 0.05$, for sample sizes $n = 20$, 50 and 100. The numbers correspond to the true value of π being equal to 0.25. For completeness we have included the randomization probabilities (the probability of rejection at the critical value) within parentheses as well. The numbers demonstrate that the exact small sample critical values can be quite far off from the appropriate percentile of the χ^2 distribution (in this case the 95th percentile of a χ^2 distribution with one degrees of freedom equals 3.841). Although here we have only presented the results for three values of χ, a more detailed analysis (not presented here) reveals that statistics with larger negative values of λ appear to be further off from the chi-square approximation in small samples. In the table, the numbers clearly get closer to 3.841 as the sample size increases, but is still not completely adequate even for $n = 100$. However the codes presented here will easily work for sample sizes of, say, 200 or even higher for determining exact powers and critical values. While $n = 20$ requires the enumeration of 231 samples in all, $n = 200$ also requires the enumeration of no more than 20301 samples, a manageable task for a fast computer. Thus, whichever test one wishes to do, one should use the exact critical values determined by our codes at least upto sample sizes of 100, perhaps even larger.

Table 1: Exact small sample critical values and randomization probabilities for some test statistics; $\lambda = 0.25$.

		Value of κ_0				
n	λ	0.25	0.4	0.5	0.7	0.9
20	1	3.9216	4.0761	3.8948	4.3299	4.5885
		(0.8778)	(0.6044)	(0.8041)	(0.6270)	(0.4826)
	0	4.5741	4.5657	4.4483	4.8878	2.7431
		(0.7906)	(0.7454)	(0.4119)	(0.2452)	(0.2906)
	– 0.5	6.7429	5.3711	5.5898	8.0024	3.8893
		(0.8701)	(0.1904)	(0.4275)	(0.1753)	(0.9235)
50	1	3.9175	3.8949	3.9512	3.8662	3.7933
		(0.4910)	(0.0929)	(0.4247)	(0.0002)	(0.9841)
	0	4.1037	4.1706	4.2605	4.1252	4.2503
		(0.6213)	(0.8183)	(0.6070)	(0.0501)	(0.5791)
	– 0.5	4.2454	4.1848	4.2550	4.3306	7.7953
		(0.6475)	(0.8884)	(0.9217)	(0.6780)	(0.2208)
100	1	3.8769	3.9690	3.9542	3.9370	3.8429
		(0.2904)	(0.6590)	(0.1613)	(0.3083)	(0.2475)
	0	3.9703	3.9833	4.0966	4.0193	3.8396
		(0.1732)	(0.6964)	(0.6996)	(0.1914)	(0.1701)
	– 0.5	4.0455	4.1394	4.1455	4.0623	4.2722
		(0.4265)	(0.7786)	(0.5093)	(0.5758)	(0.7344)

Comparison of Some Statistics for a Specific Scenario

To give a flavor of the complex relationships between the tests generated by the different values of λ, we performed some analysis with exact power computations which we present graphically in the following. In figures 1–4 we present the power curves for the testing problems H_0 : κ = 0.25, 0.4, 0.5 and 0.7, respectively, for test statistics corresponding to seven different values of λ (λ = 1, 2 / 3, 0.25, 0, – 0.25, – 0.5 and – 0.9). The nominal level is α = 0.05 and the sample size is *n* = 20. There is no drastic change in the observations we describe below when the sample size is different (say 25 or 50), so we have concentrated on the sample size *n* = 20 to keep a clear focus in our analysis. The exact powers of the statistics are computed for π = 0.25 and graphically presented in the figures for κ in (0, 1). The power curve figures lead to interesting observations. No single power curve dominates all the rest in either of the four figures. In fact a statistic which is more powerful than another for an alternative on one side of the null is generally (though not necessarily) less powerful when the situation reverses. Almost all the tests are biased, their power curves dipping below the level of the test in the vicinity of the null. It appears that for the case H_0 : κ = 0.25, the tests corresponding to large positive values of λ perform better when the alternative value of kappa *exceeds* the null. On the other hand large negative values of λ are superior (although marginally) in this case when the alternative is *smaller* than the null. The situation, however, reverses when the null value of kappa equals 0.7. Here large negative values of λ provide more powerful tests for alternatives *larger* than the null on the average, while large positive values are preferred for the alternatives *below* the null. For κ = 0.5 the situation is less clear cut, but closer to the H_0 : κ = 0.7 case. But for the case H_0 : κ = 0.4, general observations in the nature of the above do not appear to hold.

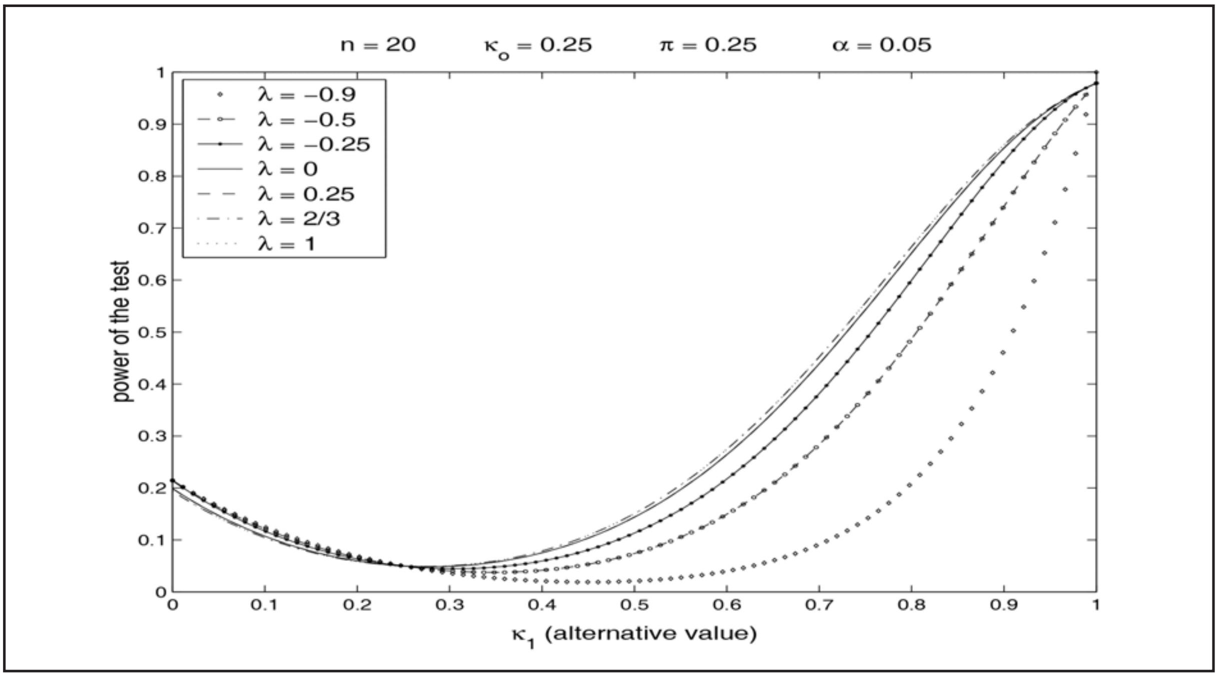

Fig. 1: Power Curves for different values of λ for testing H_0 : κ = 0.25

In Figures 5–8, which again correspond to H_0 : κ = 0.25, 0.4, 0.5 and 0.7, respectively, we plot the power functions from a different perspective. In each of these graphs, we plot the power functions at five, fixed, alternative values of *κ* (κ_1 = 0.25, 0.4, 0.5, 0.7 and 0.9), but for all values of λ in (–1, 1]; the computations are at the nominal level α = 0.05. On the whole they confirm what we have observed in figures 1–4. For example, figure 5 reiterates that power is an increasing function of λ at the alternatives κ_1 = 0.4, 0.5, 0.7 and 0.9 when testing H_0 : κ = 0.25. Figure 8 confirms that power is an increasing function of λ at κ_1 = 0.25, 0.4, 0.5, but a decreasing function at κ_1 = 0.9 when testing H_0 : κ = 0.7. The curves in figure 6 seem have a general trend in either the upward or downward direction although they are not monotone; Figure 5 appears to indicate that the powers at each of the alternatives are higher for central values of λ. Thus looking for a pattern in the power functions is very difficult for the H_0 : κ = 0.4 or H_0 : κ = 0.5 cases.

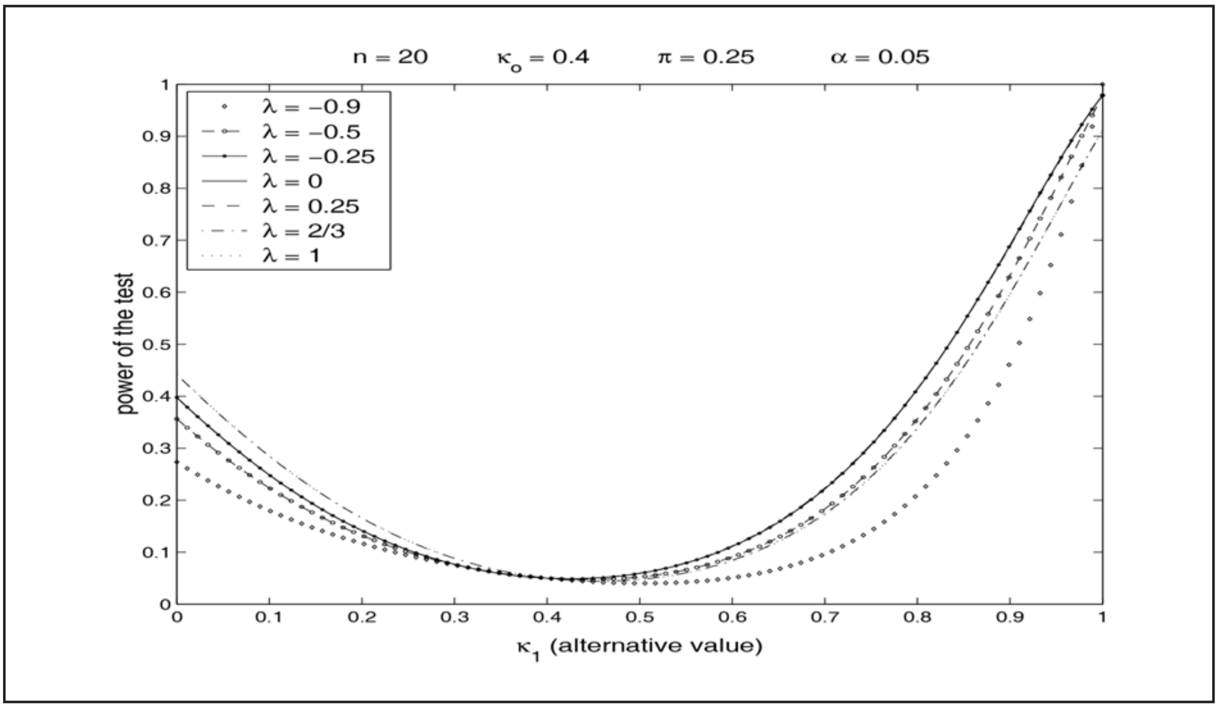

Fig. 2: Power Curves for different values of λ for testing $H_0 : \kappa = 0.4$

It is clear that it can lead to very weak inference about κ if we choose a single, fixed, test statistic for performing hypothesis tests about κ through the goodness-of-fit approach. Figures 1–8 demonstrate that for each of the null hypotheses considered here there are certain alternatives where the use of the Pearson's chi-square will lead to disaster in terms of attained power. For example if one is interested in testing $H_0 : \kappa = 0.7$, and if it is suspected that the true value of κ is greater than the null value, the test based on χ_1^2 will be practically powerless to detect the failure of the null. This is not a specialty of the Pearsonian chi-square alone; each individual test within the Cressie-Read family appears to have its own pitfall at certain alternatives when testing for any null value of κ. In short, the choice of any single, specific test statistic will lead to substantial loss of power at several (null, alternative) combinations.

Thus one can clearly do better by choosing the test as a function of the (null, alternative) combination. Yet, there do not appear to be general guidelines for choosing the most powerful statistic for a suspected alternative at a given null value of κ. The message of the graphs in figures 1–8 are sufficiently muddled; effective general recommendations for choosing the most powerful test at given (null, alternative) combinations based on the above are difficult. Coupled with the variations which arise due to the change in the sample size *n* and success probability π (not presented here for brevity), it is virtually impossible to design a general rule to pick out the most powerful test within the Cressie-Read family for a given situation.

The computational route and the suggested recipe

We have observed that the asymptotic limit is inadequate for the distribution of these test statistics in small samples. Choosing the optimal test by observing the patterns of the power functions appears to be extremely difficult. However, with a fairly simple computational effort, we can let the computer choose the optimal test for us. In the appendix we have added the SPLUS codes for the two programs we have used for this purpose. The first program simply enumerates all the samples given a particular sample size. The second program calculates the power of the test for a given null value, given value of the alternative, given level α, and an underlying value of π. By simply running the second program over a loop of a sequence of λ values one can construct the power curve as a function of λ at any (null, alternative) combination. In practise π will be unknown, and in actual implementation, given the response vector (n_1, n_2, n_3), we will calculate $\hat{\pi}$, the MLE of π, substitute that in the expression of the cell probabilities given in (1), run the two programs successively (the second one over a loop), and can come up with a very good idea about the optimal test that will generate the maximum power against a specified alternative.

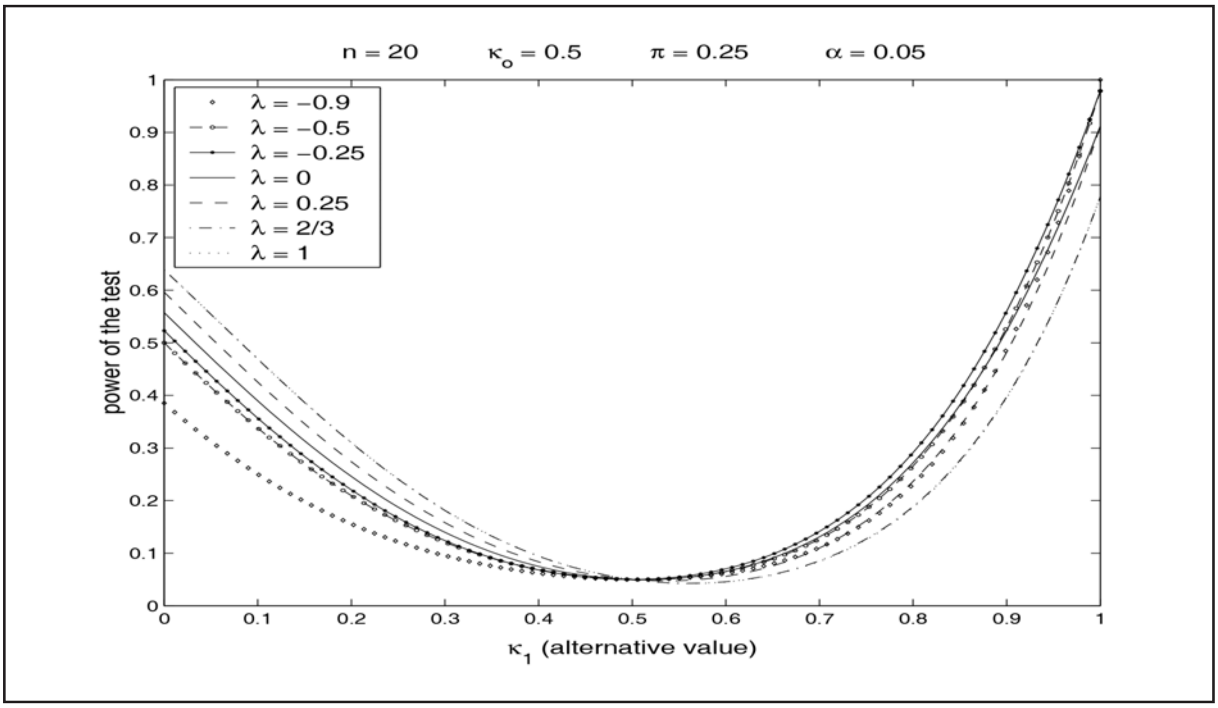

Fig. 3: Power Curves for different values of λ for testing H_0 : κ = 0.5

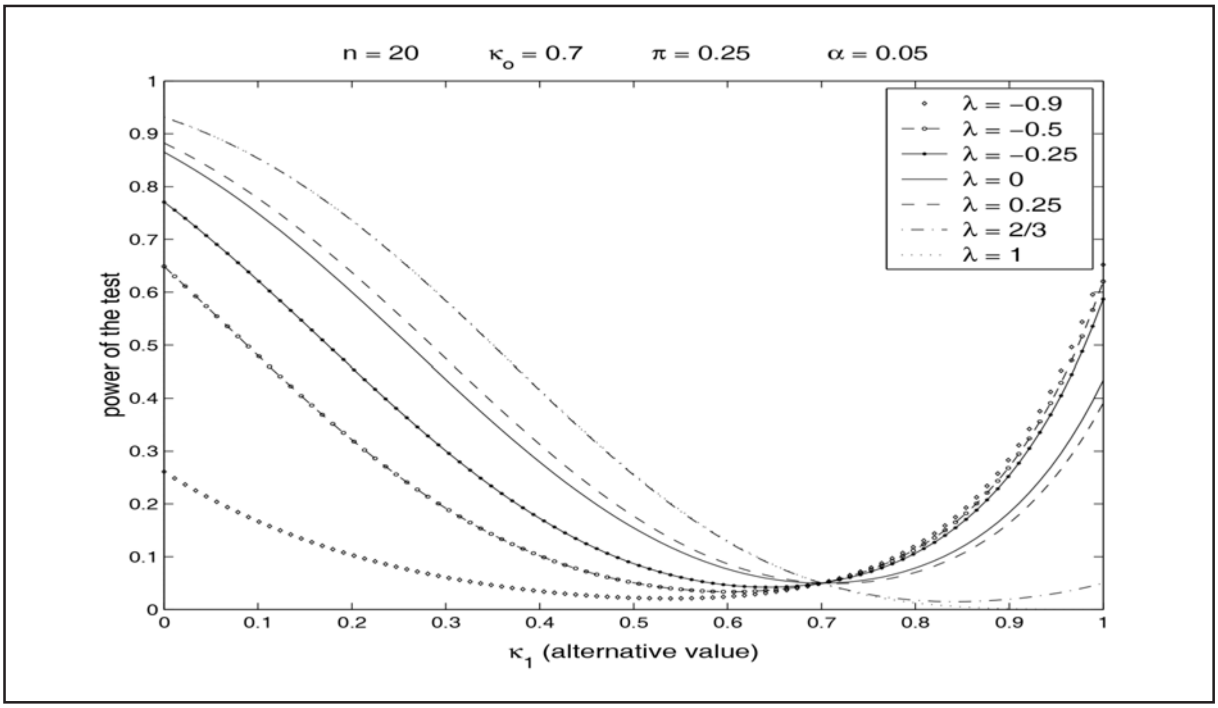

Fig. 4: Power Curves for different values of λ for testing H_0 : κ = 0.7

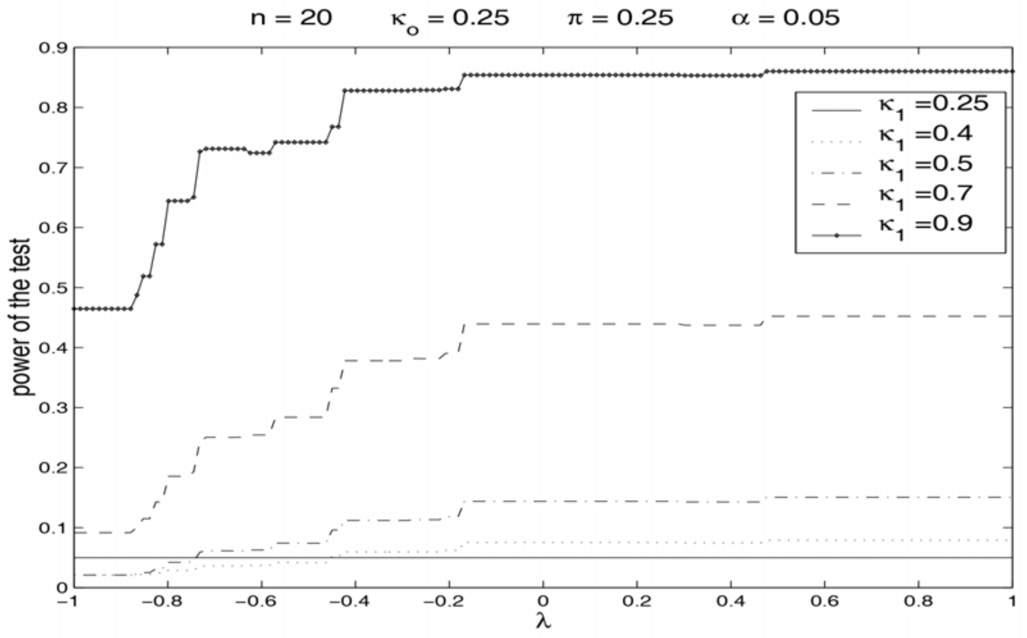

Fig. 5: Power curves for different κ_1 (alternative values of κ) for testing H_0 : κ = 0.25

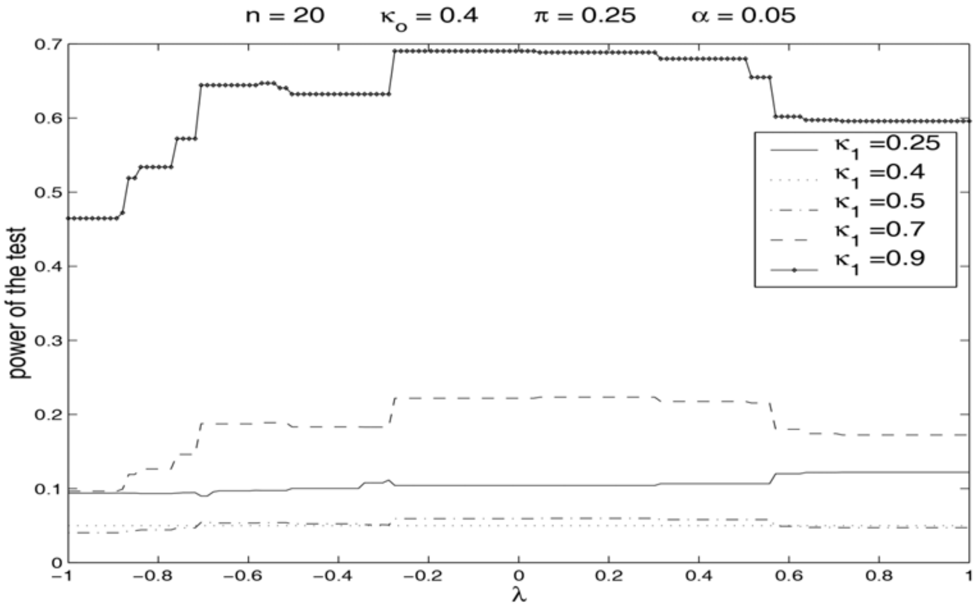

Fig. 6: Power curves for different κ_1 (alternative values of κ) for testing H_0 : κ = 0.4

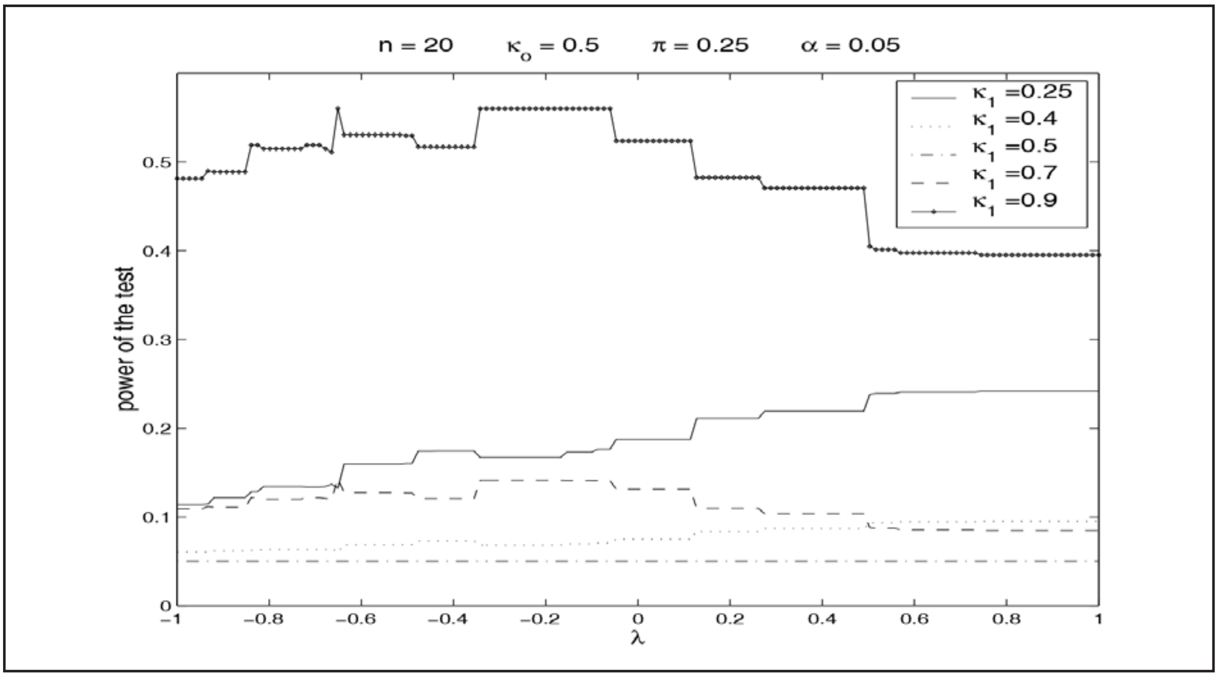

Fig. 7: Power curves for different κ_1 (alternative values of κ) for testing $H_0 : \kappa = 0.5$

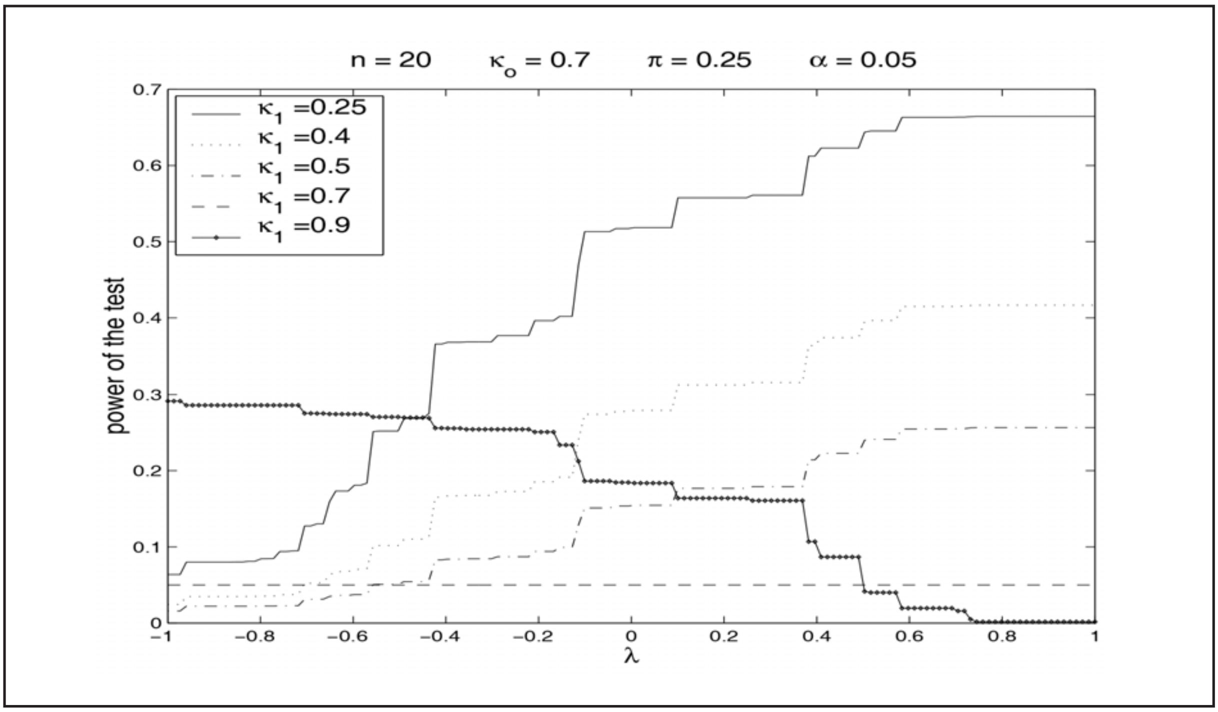

Fig. 8: Power curves for different κ_1 (alternative values of κ) for testing $H_0 : \kappa = 0.7$

An Example

We present here an example extracted from Shoukri (2004). Senior undergraduate students at the Ontario Veterinary College were asked to identiy foals with cervical vertebral malformation from x-rays. The x-rays were classified as either affected or normal. Two students (students A and B in Shoukri's example; pp. 49-50) classified 20 individual cases, and denoting the affected cases as successes, the realized data vector (n_1, n_2, n_3) equals (10, 5, 5). This leads to estimated values $\hat{\pi}$ = 0.625 and $\hat{\kappa}$ = 0.467 for π and κ. We assume π to be 0.625, α to be 0.05, and want to test $H_0 : \kappa = 0.4$ so that we have the maximum possible power at $\kappa_1 = 0.6$. The most powerful test in this scenario is generated by λ around – 0.34, with the corresponding power being 0.1567 at $\alpha = 0.05$, where the critical value and the randomization probabilities are 4.0991 and 0.0832 respectively. A test with the Pearsonian chi-square would only generate a power of 0.0794 with this combination of hypotheses. Actual implementional leads to a failure to reject the null hypothesis $H_0 : \kappa = 0.4$ when testing with $\lambda = -0.34$; all the other tests fail to reject the null with this data as well. For the same π and α, the $\chi^2_{-0.34}$ test does, however, reject the null for the observed data (10, 2, 8), but the Pearson's chi-square as well as many other tests within the Cressie-Read family fail to do so.

6. Concluding Remarks

In this paper we have attempted to choose the most powerful test for the kappa statistic at specific alternatives by extending the previous goodness-of-fit approaches. We have restricted ourselves to the simplest case, but as this simple scenario is fairly often encountered in practice we trust our methods will be quite useful. For each combination of parameter values the software provided helps to generate the optimal test within the power divergence family.

The following generalizations to this work are necessary and would be useful, and we hope to undertake them in the near future. Firstly, it is necessary to extend this to the case of multi-rater agreement, and to the case where the response is categorical with more than two categories. Often such response are ordinal, which adds another dimension to this problem. It is relatively straightforward to extend the approach of the present paper to some of these scenarios, but can be quite nontrivial for some of the others. The other problem that has to be addressed is the development of compromise test statistics which consider testing a null value of κ against a composite alternative, and has reasonable (not necessarily optimal) power at all (or most) of the alternatives. This is another problem which we propose to handle in the future by extending the approach of Basu et al. (2001).

REFERENCES :

Altaye, M., Donner, A. and Eliasziw, M. (2001). A general goodness-of-fit approach for inference procedures concerning the kappa statistic, *Stat. Medicine*, **20** : 2479–88.

Altaye, M., Donner, A. and Klar, N. (2001). Inference procedures for assessing interobserver agreement among multiple raters, *Biometrics*, **57** : 584–88.

Basu, A., Ray, S., Park, C. and Basu, S. (2002). Improved power in multinomial goodness-of-fit tests, *Statistician*, **51** : 381–93.

Basu, S. and Basu, A. (1995). Comparison of several goodness-of-fit tests for the kappa statistic based on exact power and coverage probability, *Stat. Medicine*, **14** : 347–56.

Cohen, J. (1960). A coefficient of agreement of nominal scales, *Educational and Psycho- logical Measurement*, **20** : 37–46.

Cressie, N. and Read, T.R.C. (1984). Multinomial goodness-of-fit tests, *J. Roy. Statist. Assoc.*, **B**, **46** : 440–64.

Donner, A. and Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures of the kappa statistic; confidence inerval construction, significance-testing and sample size estimation, *Stat. Med.*, **11** : 1511–19.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Ed., Wiley, New York.

Gonin, R., Lipsitz, S.R., Fitzmaurice, G.M., and Molenberghs, G. (2000). Regression modelling of weighted kappa by using generalized estimating equations, *Applied Statistics*, **49** : 1–18.

Klar, N., Lipsitz, S.R. and Ibrahim, J. (2000). An estimation equations aproach for medelling kappa, *Biometrical Journal*, **42** : 45-58.

Read, T. R. C. and Cressie, N. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data.* Springer-Verlag, New York.

Shoukri, M. M. (2004). *Measures of interobserver agreement*,

Appendix

Program 1: This program enumerates all the samples (*n*1*; n*2*; n*3) satisfying *n*1+*n*2+*n*3 = *n* given a particular sample size *n*

```
# ********** Program 1 **********
# Two rater kappa with binomial response
# To generate multinomial sample
n_25
# n is the sample size
# (n+g-1) choose (g-1) is the number of ways in which n objects
# —— cont —— can be divided into g groups
# Here number of groups is g=3
g_3
m_((n+2)*(n+1))/2
# m is the number of different samples in the sample space
sample.vec_matrix(0,m,g)
# The sample vec matrix will store the different samples along
# —— cont —— its rows
index_0
index1_0
repeat{#index1 loop starts
index2_0
repeat{#index2 loop starts
index3_0
repeat{#index3 loop starts
x_c(index1,index2,index3)
if(sum(x) == n)
{
index_index+1
#print(paste(index,"index"))
sample.vec[index,]_x
#print(x)
}
else if(sum(x) > n)
break
index3_index3 + 1
if(index3 > n) break
}#End of index3 loop
index2_index2 + 1
if(index2 > n)break
}#End of index2 loop
index1_index1+1
if(index1 > n)break
}#End of index1 loop
print(paste("The number of groups g = ",g))
print(paste("The sample size n = ",n))
print(paste("number of samples m = ",m))
# ********** End Program 1 **********
```

Program 2: This program calculates the exact small sample critical value and the randomization probability for a given sample size *n*, success probability π, null value κ_0, and parameter index λ. The program also calculates the exact small sample power at a given alternative value of κ.

```
# ********** Program 2 **********
# Testing Program for the Kappa statistic
# Before running this file, run "sample.in" to
# —— cont —— generate the samples
n_25
g_3
p1_0.25
# p1 represents the pi parameter
alpha_0.05
# alpha represents the level of the test
alpha1_1-alpha
knull_0.90
kalt_0.50
lambda_0
# knull represents the null kappa parameter
# kalt represents the alternative kappa parameter
m_(n+2)*(n+1)/2
# Finding cell probabilities under the null
null.1_p1*p1+p1*(1-p1)*knull
null.2_2*p1*(1-p1)*(1-knull)
null.3_(1-p1)*(1-p1)+p1*(1-p1)*knull
null.vec_c(null.1,null.2,null.3)
# null.vec is the vector of cell probabilities under the null
# Finding cell probabilities under the alternative
alt.1_p1*p1+p1*(1-p1)*kalt
alt.2_2*p1*(1-p1)*(1-kalt)
alt.3_(1-p1)*(1-p1)+p1*(1-p1)*kalt
alt.vec_c(alt.1,alt.2,alt.3)
# alt.vec is the vector of cell probabilities under the alternative
sampnull.vec_rep(0,m)
# sampnull.vec stores the probabilities of the m samples under the null
sampalt.vec_rep(0,m)
# sampalt.vec stores the probabilites of the m samples under the alt
num_prod(c(1:n))
for(index in 1:m){
x_sample.vec[index,]
x[x==0]_1
den_prod(c(1:x[1])) * prod(c(1:x[2])) * prod(c(1:x[3]))
sampnull.vec[index]_(num/den)*prod (null.vec ^ sample.vec[index,])
sampalt.vec[index]_(num/den) *prod(alt.vec^ sample.vec[index,])
}
mminus1_m-1
m1_m-2
```

```
sample1.vec_sample.vec[2:mminus1,]
sampnull1.vec_sampnull.vec[2:mminus1]
sampnull1.vec_sampnull1.vec/(1-sampnull.vec[1]-sampnull.vec[m])
sampalt1.vec_sampalt.vec[2:mminus1]
sampalt1.vec_sampalt1.vec/(1-sampalt.vec[1]-sampalt.vec[m])
test.stat_rep(0,m1)
# sample1.vec, sampnull1.vec, sampalt1.vec are the modificed vectors
# —— cont —— after removing the two samples for which the mle of
# —— cont —— pi (pihat) is zero
for (index in 1:m1){
xx_sample1.vec[index,]
xx1_xx[xx>0]
# Calculating the estimated expected cell probabilities under the
# —— cont —— null for this sample
# First need pihat (the mle of pi)
pihat_(2*xx[1]+xx[2])/(2*n)
nullest.1_pihat*pihat+pihat*(1-pihat)*knull
nullest.2_2*pihat*(1-pihat)*(1-knull)
nullest.3_(1-pihat)*(1-pihat)+pihat*(1-pihat)*knull
nullest.vec_c(nullest.1,nullest.2,nullest.3)
expt.vec_n*nullest.vec
# manipulating to handle observed zero frequency cells
ratio_sample1.vec[index,]/expt.vec
ratio1_ratio[ratio>0]
###########if(lambda==0)################
if(lambda==0){
test.stat[index]_2 * sum(xx1 * log(ratio1))
}
##########if(lambda==-1)####################
else if(lambda==-1){
if(length(ratio[ratio==0]) > 0) test.stat[index]_50000
else if (length(ratio[ratio==0]) == 0)
test.stat[index]_2 * sum(expt.vec * log(1/ratio))
}
#######if(lambda < -1)#############
else if (lambda < -1)
{
if(length(ratio[ratio==0]) > 0 ) test.stat[index]_50000
else if(length(ratio[ratio==0]) == 0 )
{
test.stat[index]_2 * sum((xx * (ratio^lambda -1)))
test.stat[index]_test.stat[index]/(lambda * (1+lambda))
}
}
#######if(lambda>-1 and < 0 )#################
else if(lambda > -1 && lambda < 0){
test.stat[index]_2*(sum((xx^(lambda+1))/expt. vec^lambda)-n)
test.stat[index]_test.stat[index]/(lambda * (1+lambda))
}
#######if(lambda > 0 )#############
else if(lambda > 0){
test.stat[index]_2 * sum((xx * (ratio^lambda -1)))
test.stat[index]_test.stat[index]/(lambda * (1+lambda))
}
}#End of for loop
temp_cbind(test.stat,sampnull1.vec,sampalt1.vec)
temp_temp[sort.list(test.stat),]
cum.sampnull1.vec_cumsum(temp[,2])
cum.sampalt1.vec_cumsum(temp[,3])
temp_cbind(temp,cum.sampnull1.vec,cum.sampalt1.vec)
# Finding the small sample critical value and
# —— cont —— the randomization probability
tindex0_1
tindex1_1
tindex2_1
found_0
for (i in 1:m) {
if (temp[,4][i] > alpha1) found_1
if (found==1) tindex0_i
if (found==1) break}
found_0
for (i in tindex0:1) {
if (temp[,1][tindex0] - temp[,1][i] > 0.00001) found_1
if (found==1) tindex1_i
if (found==1) break}
found_0
for (i in tindex0:m) {
if (temp[,1][i] - temp[,1][tindex0] > 0.000001) found_1
if (found==1) tindex2_i-1
if (found==1) break}
print(paste(“tindex0 =”, tindex0, “ tindex1 = “, tindex1, “ tindex2= “,tindex2))
a1_temp[,4][tindex1]
a2_temp[,4][tindex2]
b1_temp[,5][tindex1]
b2_temp[,5][tindex2]
crit.val_temp[,1][tindex2]
rand.prob_(alpha1-a1)/(a2-a1)
print(paste(“critical value = “,crit.val,”randomization probability = “,rand.prob))
power_1-(b1+(b2-b1)*(alpha1-a1)/(a2-a1))
print(paste(“power = “,power))
print(paste(“Result for Complex Test for n=”,n,” pi =”,p1))
print(paste(“Null Kappa knull = “,knull))
print(paste(“Alternative Kappa kalt = “,kalt))
print(paste(“Value of lambda = “,lambda))
# ********** End Program 2 **********
```