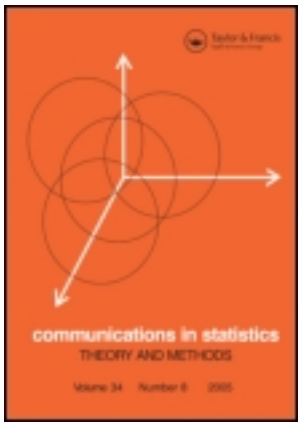

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

Testing Supremacy or Inferiority of Multinomial Cell Probabilities with Application to Biting Preferences of Loggerhead Marine Turtles

Alessio Farcomeni [a]
[a] Sapienza - University of Rome, Rome, Italy

Available online: 14 Dec 2011

To cite this article: Alessio Farcomeni (2012): Testing Supremacy or Inferiority of Multinomial Cell Probabilities with Application to Biting Preferences of Loggerhead Marine Turtles, Communications in Statistics - Theory and Methods, 41:1, 34-45

To link to this article: http://dx.doi.org/10.1080/03610926.2010.513786

Communications in Statistics—Theory and Methods, 41: 34–45, 2012

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610926.2010.513786

Taylor & Francis
Taylor & Francis Group

Testing Supremacy or Inferiority of Multinomial Cell Probabilities with Application to Biting Preferences of Loggerhead Marine Turtles

ALESSIO FARCOMENI

Sapienza - University of Rome, Rome, Italy

We propose exact tests for uniform superiority or uniform inferiority of a multinomial cell probability. We extend the approach to testing supremacy or inferiority in multivariate settings. We also show how to perform superiority or inferiority tests in presence of covariates, and discuss the different interpretations of the proposed tests. The tests are illustrated on an original data set on biting behavior of loggerhead marine turtles.

Keywords Exact test; Inferiority; Intersection-union test; Logit; Mid-P value; Multinomial; Permutation; Supremacy; Union-intersection test.

Mathematics Subject Classification Primary 62F03; Secondary 62Q05.

1. Introduction

When a multinomial response is observed, a question of interest may be if a cell probability is uniformly larger (*supremacy*) or uniformly smaller (*inferiority*) than all others. For instance, a new drug can be tested together with already available drugs, and its efficacy rate tested for superiority or its adverse events rate tested for inferiority. A panel of customers can be presented with different products, and a company may be interested in discovering whether their new product is preferred to the others. In these and other scenarios a question of interest is whether a multinomial cell probability, chosen prior to data collection, is uniformly larger or uniformly smaller than the other cell probabilities.

Tests for supremacy have been recently developed and investigated both in simulations and theoretically by Nettleton (2009). The Likelihood Ratio Test (LRT) of Nettleton (2009) is equivalent to the test of Berry (2001), who developed it for testing for the general existence of a unique most probable cell (with unknown identity).

Received December 4, 2009; Accepted August 3, 2010
Address correspondence to Alessio Farcomeni, Sapienza - University of Rome, Piazzale Aldo Moro, 5, Rome, 00186, Italy; E-mail: alessio.farcomeni@uniroma1.it

A related but different problem is testing whether a cell corresponds to the least probable event, that is, inferiority of a cell probability. Alam and Thompson (1972) discussed this latter problem from a design point of view.

To the best of our knowledge, testing for supremacy or for inferiority of probability of a multinomial event has been considered so far only in the univariate case. Categorical covariates may define strata, and interest may lie in the supremacy within strata. For instance, a new drug can be superior to all others in a subset of patients (e.g., without diabetes) but not for another subset (e.g., patients with diabetes).

We extend in this article tests for supremacy or inferiority to joint and conditional multivariate cases, and when general covariates are recorded. We discuss the different interpretations of the proposed settings.

Our motivating example regards an original study on loggerhead marine turtles. One of the questions related to this experiment was whether turtles have a marked preference for certain visual characteristics of their prey, e.g., the color. Each turtle was placed in a test pool filled with sea water, with three baits equal in all characteristics but their color: one bait was colored blue, one red, and another yellow. Chemical cues could be added to the baits so to stimulate olfaction of the animal. A preliminary analysis and a thorough discussion of the experimental setting and findings can be found in Piovano et al. (2009). A question of interest was whether red baits were least preferred by the turtles for their first bite. We will see in this article that we can positively answer this question only if we consider the chemical cues, i.e., the red baits are inferior to all others only when chemical cues are absent.

The rest of the article is as follows. In Sec. 2, we introduce our tests statistics and permutation strategy to exactly evaluate significance. We extend the approach to joint and conditional multivariate settings in Sec. 3, and to multinomial models with covariates in Sec. 4. A brief simulation study is used in Sec. 5 to illustrate the validity of the permutation approach. We then apply the tests to our motivating example in Sec. 6. Finally, in Sec. 7 we give some concluding remarks and summarize the proposed tests, their applicability, and interpretation.

2. Intersection Union Tests for Supremacy and Inferiority

Suppose Y_i, $i = 1, \ldots, n$ are i.i.d. categorical response variables with k categories, such that $\Pr(Y_i = j) = \pi_j$, $j = 1, \ldots, k$; $\pi_j > 0$ and $\sum_j \pi_j = 1$. The (minimal) sufficient statistics for this problem are given by the counts $(n_1, \ldots, n_k)$ of subjects observed in each category, with $\sum n_i = n$.

Without loss of generality, we can assume the target category to be the kth, so that we can formulate a set of hypotheses for supremacy as

$$\begin{cases} H_0 : \pi_k \leq \max\{\pi_1, \ldots, \pi_{k-1}\} \\ H_1 : \pi_k > \max\{\pi_1, \ldots, \pi_{k-1}\} \end{cases} \tag{1}$$

and for inferiority as

$$\begin{cases} H_0 : \pi_k \geq \min\{\pi_1, \ldots, \pi_{k-1}\} \\ H_1 : \pi_k < \min\{\pi_1, \ldots, \pi_{k-1}\} \end{cases} \tag{2}$$

It is easily seen that the null hypothesis in (1) can be reformulated as $H_0 : \bigcup_{j=1}^{k-1} \frac{\pi_k}{\pi_j} \leq 1$, while the null hypothesis in (2) can be reformulated as $H_0 : \bigcup_{j=1}^{k-1} \frac{\pi_k}{\pi_j} \geq 1$.

Since the null hypothesis is a union of sets, we can use the intersection-union (IU) principle (see for instance, Casella and Berger, 2002, Ch. 8; Silvapulle and Sen, 2004). Berger and Hsu (1996) studied IU tests in the related context of bioequivalence trials.

The IU test for supremacy (inferiority) reduces to separately computing $k-1$ p-values, $p_1, \ldots, p_{k-1}$, one for each null of the kind $H_{0j} : \frac{\pi_k}{\pi_j} \leq (\geq) 1$, and then setting $p = \max\{p_1, \ldots, p_{k-1}\}$. Nettleton (2009) showed that using a LRT for H_{0j} and the IU principle is equivalent to the LRT for H_0.

In this article, we use three ways of testing H_{0j}, or an appropriate formulation related to the setting. One is still based on the LRT, the second is based on the Wald statistic, and the third is based on a permutation strategy (Pesarin, 2001), and shall be used to evaluate exact significance levels for H_{0j} when the contingency table is too sparse or too imbalanced to make asymptotic approximations reliable. In the univariate case this is seldom the case, but the exact approach will be useful for the generalizations to the multivariate case. It is proved in Theorem 2.1 that the p-value for testing H_{0j} can be evaluated exactly through a binomial test (Conover, 1998). In this article, we also propose a correction to the binomial test, obtaining the mid-P value (Lancanster, 1961; Agresti, 2002).

More formally, letting $M = \max_{j=1}^{k-1} n_j$, the p-value for testing (1) can be evaluated as

$$p = \sum_{x=n_k}^{n_k+M} \binom{M+n_k}{x} 2^{-M-n_k}. \tag{3}$$

When testing inferiority, the sum goes from zero to n_k and $M = \min_{j=1}^{k-1} n_j$. Due to the discrete nature of the problem, the exact test may be somehow conservative and the exact level may be sensibly smaller than α. A continuity correction is obtained for instance through the mid-P value, which is less conservative, and is defined as

$$p = \binom{M+n_k}{n_k} 2^{-M-n_k-1} + \sum_{x=n_k+1}^{n_k+M} \binom{M+n_k}{x} 2^{-M-n_k}. \tag{4}$$

We now show that tests based on computing the significance level as (3) or as (4) are α-level.

Theorem 2.1. *Under the null hypothesis, p-values evaluated through* (3) *or* (4) *are below* α *with probability at most* α.

Proof. When testing supremacy, the p-value for testing $H_{0j} : \frac{\pi_k}{\pi_j} \leq 1$ is formally defined as $p_j = \sup_{\pi_k \leq \pi_j} \Pr(T_j \geq T_j^{(o)} \mid \pi) = \Pr(T_j \geq T_j^{(o)} \mid \pi_k = \pi_j)$. Many test statistics lead to the binomial test we are proposing. We can use for instance $T_j^{(o)} = n_k/(n_j + n_k)$.

We can set up a permutation strategy, and condition upon the marginal count $n_j + n_k$. Under the hypothesis that $\pi_j = \pi_k$, this reduces to testing that n_k arises from a binomial distribution with parameters $(n_j + n_k)$ and 0.5, leading to the exact binomial test.

Note that the counts are differently distributed in H_{0j} and H_{1j} since their distribution depends on the vector π. The proportions $n_j/(n_j + n_k)$ and $n_k/(n_j + n_k)$ are differently distributed in H_{0j} and H_{1j} also conditionally on $n_j + n_k$ since their conditional distributions depend on $p_j/(p_j + p_k)$.

The p-value for the exact binomial test of H_{0j} is defined as $p_j = \sum_{x=n_k}^{n_k+n_j} \binom{n_j+n_k}{x} 2^{-n_j-n_k}$; for instance, see Conover (1998).

Since we have a IU test, the null hypothesis H_0 is rejected only if $H_{0j} : \frac{\pi_k}{\pi_j} \leq 1$ is rejected for all $j = 1, \ldots, k-1$, that is, if $p_j < \alpha$ for all $j = 1, \ldots, k-1$. Consequently, $p = \max_j p_j$, and we get (3).

The number of permuted test statistics which are exactly as extreme as the observed test statistic is $\binom{n_j+n_k}{n_k}$. One can divide this number by a factor of two without violating the nominal α level (Agresti, 2002; Lancanster, 1961), obtaining the mid-P value, whose expression is given by (4).

The same reasoning applies to inferiority tests.

It is interesting to note that the exact binomial test corresponds to a permutation test in which all possible permutations are enumerated. Each permutation p-value is computed as the fraction of permuted test statistics which are as extreme or more extreme than the observed test statistic. The p-values are consequently evaluated exactly.

Due to the fact that we are in a categorical data situation, there are many available choices for working with ties; for instance, refer to Good (2000). One strategy would be to use randomization, but with the undesireable feature that different runs on the same data may lead to different conclusions. A widely used solution is given by using the mid-P value, which gives half weight to the ties, as we do in this article. The mid-P value is equivalent to a repeated randomization.

3. Conditional and Joint Multivariate Tests for Supremacy or Inferiority

In this section we discuss testing for supremacy or inferiority in the multivariate case. There are basically three different situations of interest: joint testing, universal conditional testing, and omnibus conditional testing.

The first situation is formalized as testing for supremacy or inferiority on a joint multivariate distribution: for each observation $i = 1, \ldots, n$ we observe a *vector* of R categorical response variables $Y_{i1}, \ldots, Y_{iR}$, each with k_r, $r = 1, \ldots, R$, categories. This multivariate setting applies when $Y_{i1}, \ldots, Y_{iR}$ are different in nature (for instance, for each subject we observe a categorical score related to a different health aspect), but also when there are repeated measurements over time (i.e., longitudinal categorical data).

Define $\pi_{j_1,\ldots,j_R} = \Pr(Y_{i1} = j_1, \ldots, Y_{iR} = j_R)$. The null hypothesis for supremacy on the *joint* probability can be formulated, without loss of generality, as

$$H_0 : \pi_{k_1,\ldots,k_R} \leq \max\{\pi_{j_1,\ldots,j_R}; j_1 = 1, \ldots, k_1 - 1; \ldots; j_R = 1, \ldots, k_R - 1\}.$$

The null hypothesis for inferiority can be formulated similarly.

The minimal sufficient statistics in this case are the counts $n_{j_1,\ldots,j_r}$, for all possible configurations.

It is recognized readily the equivalence with a univariate setting with a response $\widetilde{Y}_i$ with $\prod_{r=1}^{R} k_r$ categories, one for each possible combination of the outcomes

$Y_{i1}, \ldots, Y_{iR}$. Hence, while there is no theoretical difference with the univariate case, in practice many counts are often equal to zero and the asymptotic distributional results for the LRT and Wald statistics are not reliable. Hence, one shall use the permutation approach of (3) or (4) to evaluate significance in the multivariate case.

A different situation arises when interest lies in testing on the *conditional* probabilities. For simplicity, we suppose there is a single outcome of interest Y_i, $i = 1, \ldots, n$, and S strata defined by a categorical covariate X_i. This setting readily generalizes to R outcomes and many categorical covariates, as for instance strata can be defined also by combinations of the levels of more than one categorical covariate.

Define $\pi_{j|s} = \Pr(Y_i = j \mid X_i = s)$; $j = 1, \ldots, k$; $s = 1, \ldots, S$. The univariate tests of the previous section correspond to testing that $\pi_{k|s}$ is the largest (smallest) when averaged over s with respect to an opportune (and unknown) discrete probability measure for X_i.

There are two additional kinds of conditional tests for supremacy which may be of interest. The first can be used to verify supremacy of the kth cell in all strata, i.e., a *universal* supremacy test whose null hypothesis is defined as

$$H_0 : \exists\ s = 1, \ldots, S\ :\ \pi_{k|s} \leq \max\{\pi_{1|s}, \ldots, \pi_{k-1|s}\}.$$

The universal supremacy null hypothesis can be still tested using the IU principle. In order to do so, we shall note that H_0 is the union of S null hypotheses H_{0s} : $\pi_{k|s} \leq \max\{\pi_{1|s}, \ldots, \pi_{k-1|s}\}$. We can compute a significance level p_s for each H_{0s} using a univariate test for supremacy (with the LRT test if the counts $n_{1s}, \ldots, n_{ks}$ are large enough, or with binomial tests if asymptotic approximations are likely not reliable) and then compute the significance level p corresponding to H_0 as $p = \max\{p_1, \ldots, p_S\}$. The null hypothesis for universal inferiority can be formulated and tested similarly.

The second kind of conditional approach may be used to verify supremacy of the kth cell in at least one stratum, i.e., an *omnibus* supremacy test whose null hypothesis is defined as

$$H_0 : \pi_{k|s} \leq \max\{\pi_{1|s}, \ldots, \pi_{k-1|s}\}, \quad \forall s = 1, \ldots, S.$$

Omnibus supremacy can be tested using a union-intersection (UI) principle. In fact, $H_0 : \bigcap_s H_{0s}$, where H_{0s} is defined as above. Consequently, by the UI principle, we shall compute a significance level p_s for each H_{0s} and compute the significance level p corresponding to H_0 as $p = \min\{p_1, \ldots, p_S\}$.

If the omnibus test is rejected, we can claim there is at least one stratum in which the kth cell probability is the largest. One could then proceed with multiple comparisons performing S univariate tests. The stratum-specific significance levels would have already been computed for the omnibus test, and would only need being adjusted for multiplicity (e.g., Hochberg and Tamhane, 1987). One could simply use a Bonferroni correction, or a more powerful correction like Holm or Sidak. Information by dependence among strata could be exploited by combining the p-values. The null hypothesis for omnibus inferiority can be formulated and tested similarly.

Note finally that we can easily generalize to testing supremacy for stratum-specific cells $\pi_{k_s|s}$, $s = 1, \ldots, S$, both in the omnibus and universal testing approaches.

4. Accounting for Covariates

Suppose now for each subject we have a vector of covariates X_i, so that $\Pr(Y_i = j \mid X_i) = \pi_{ij}$.

We start by specifying, for $i = 1, \ldots, n$, a multinomial logit model of the kind:

$$\begin{cases} \log\left(\dfrac{\pi_{i1}}{\pi_{ik}}\right) = \gamma_1 + \beta_1' X_i \\ \log\left(\dfrac{\pi_{i2}}{\pi_{ik}}\right) = \gamma_2 + \beta_2' X_i \\ \qquad\qquad . \\ \log\left(\dfrac{\pi_{i(k-1)}}{\pi_{ik}}\right) = \gamma_{k-1} + \beta_{k-1}' X_i, \end{cases} \tag{5}$$

where the local logit reparameterization is conveniently chosen using the kth category as a baseline, γ_j denotes a category specific intercept, and β_j a category specific vector of logistic regression parameters. Details on such a model can be found, for instance, in Agresti (2002).

There are many available alternative parameterizations, which do not modify the approach we are about to propose. A parsimonious alternative for instance is given by setting $\beta_j = \beta$ for all $j = 1, \ldots, k-1$. This latter parameterization follows a standard practice in regression models for ordinal variables (for instance, see McCullagh, 1980).

The kth cell probability is uniformly larger if $\gamma_j + \beta_j' X_i$ is always less than or equal to zero, i.e., if it is the largest for any configuration of the covariates. This imples supremacy uniformly with respect to the values of the covariates, and reduces to the universal conditional approach of the previous section with categorical covariates.

In the majority of cases in which we are interested in testing supremacy (or inferiority), covariates are deemed as confounders. We then propose to remove the effects of covariates, and test the following supremacy (inferiority) null hypothesis: $H_0 : \bigcup_{j=1}^{k-1} \gamma_j \geq (\leq) 0$. The intercept γ_j in fact summarizes the ratio between the jth and the kth multinomial cell probabilities after having adjusted for individual covariates.

Formally, the null hypothesis corresponds to testing for supremacy (inferiority) of the kth cell probability for a zero configuration of the covariates. We would like to stress once again that if the null is false the kth cell probability may still not be the largest for different covariate configurations, i.e., $\pi_{ki} > \pi_{ji}$ only when $X_i = 0$.

The interpretation of the test restricted to the intercepts depends on the structure of the design matrix. With continuous covariates, we are testing supremacy for the probability associated with a baseline zero level for the covariates. Covariates can be centered on means, medians, or on specific values of interest in order to make the test meaningful.

With categorical covariates, the interpretation depends on the parameterization. With a corner point parameterization (i.e., if we use a dummy variable for each category), we are testing supremacy for the probability associated with the single stratum corresponding to a value of zeros for all dummies. We refer to this test as the specific conditional supremacy testing, since it corresponds to testing supremacy or inferiority only in a given stratum.

When a sum-to-zero parameterization is used, we test supremacy of the geometric average of the probabilities across strata. To see this, suppose we have a single covariate X_i which can take values $1, \ldots, S$. A sum-to-zero parameterization leads to the use of S parameters $\beta_{j1}, \ldots, \beta_{jS}$ with the restriction that $\sum_s \beta_{js} = 0$. Due to (5) we have that

$$\exp(\gamma_j + \beta_{js}) = \frac{\pi_{j|s}}{\pi_{k|s}}.$$

Consequently, note that

$$\prod_s \exp(\gamma_j + \beta_{js}) = \prod_s \frac{\pi_{j|s}}{\pi_{k|s}}, \tag{6}$$

i.e., the ratio of the geometric average for each conditional cell probability, and that the left-hand side of (6) simplifies as

$$\prod_s \exp(\gamma_j + \beta_{js}) = \exp(S\gamma_j + \sum_s \beta_{js}) = \exp(S\gamma_j). \tag{7}$$

Hence, claiming that $\gamma_j \leq 0$ with a sum-to-zero parameterization implies that the geometric average of the kth cell probabilities is above the geometric average of the jth cell probabilities.

Consequently, a corner point parameterization corresponds to a specific H_{0s} in the notation of the previous section. On the other hand, a sum-to-zero parameterization yields a supremacy (inferiority) test which is more stringent than the omnibus conditional test and less stringent than the universal conditional test. We refer to this latter situation as *adjusted* testing. The two parameterizations can of course be mixed when using more than one categorical covariate.

Setting up a permutation approach for tests on the intercepts of (5) is cumbersome, since variability of the Pearson residuals depends on the fitted value and they are not exchangeable anymore. This problem is still unresolved in the literature about permutation methods. Also, the LRT would be cumbersome, since we would have to set up a Newton-Raphson type algorithm in the restricted parameter space defined by the null hypothesis. On the other hand, a IU Wald test is straightforward and requires only testing $H_{0j} : \gamma_j \geq (\leq) 0$ through computing a p-value p_j based on a Wald statistic. The Wald statistic relies on the (unrestricted) maximum likelihood estimate $\hat{\gamma}_j$ and its standard error, following the usual definition. The test is one-sided. By the IU principle, once again $p = \max_j p_j$.

Note that in a multinomial logit model with only intercepts, the proposed test reduces to an IU Wald test for (1) or (2).

5. Simulations

In this section we report a brief simulation study, which is used to check the validity of the binomial tests. We simulate only tests for supremacy without covariates.

We use the same simulation setting of Nettleton (2009), and compare our permutation approach with the LRT test of Nettleton (2009) and a IU Wald test, which in the simulations of Nettleton (2009) was found to be the most powerful, in different settings, among the available competitors. We will report the results for

the other tests directly from Nettleton (2009), who used a very large number of replications giving highly reliable results. We fix the $\alpha = 0.05$, and set different n, π, and k. For each simulation setting, we use 100,000 replications.

Table 1 shows the estimated Type I error rates. In almost all settings the error rates are at or below (often much smaller than) the nominal error rate. The tests are conservative when more than two cells are tied. As noted also by Nettleton (2009), when the nominal error rate is exceeded the actual error rate is not alarmingly large. For large n, the difference between the exact binomial approach and the mid-P value becomes negligible, as could be expected. All tests should be then deemed valid from a practical point of view.

Table 1

Type I error rates for nominal level $\alpha = 0.05$ for permutation testing for supremacy, LRT, and Wald IU tests, for different choices of k, cell probabilities π, and sample size n. Estimates are based on 100,000 replications

p_1	p_2	p_3	p_4	p_5	n	mid-P	Binomial	LRT	Wald IUT
0	0	0	1/2	1/2	50	0.059	0.034	0.059	0.059
0	0	1/3	1/3	1/3		0.010	0.008	0.009	0.013
0	0	1/5	2/5	2/5		0.044	0.033	0.045	0.050
0	1/4	1/4	1/4	1/4		0.004	0.002	0.004	0.005
0	1/7	2/7	2/7	2/7		0.010	0.007	0.010	0.012
0	1/6	1/6	1/3	1/3		0.040	0.028	0.039	0.044
1/5	1/5	1/5	1/5	1/5		0.002	0.001	0.002	0.002
1/9	2/9	2/9	2/9	2/9		0.004	0.002	0.004	0.005
1/8	1/8	1/4	1/4	1/4		0.009	0.006	0.010	0.011
1/7	1/7	1/7	2/7	2/7		0.034	0.023	0.033	0.038
0	0	0	1/2	1/2	200	0.052	0.038	0.052	0.052
0	0	1/3	1/3	1/3		0.012	0.009	0.013	0.013
0	0	1/5	2/5	2/5		0.048	0.042	0.051	0.052
0	1/4	1/4	1/4	1/4		0.004	0.003	0.005	0.005
0	1/7	2/7	2/7	2/7		0.011	0.009	0.012	0.012
0	1/6	1/6	1/3	1/3		0.050	0.042	0.051	0.052
1/5	1/5	1/5	1/5	1/5		0.002	0.002	0.002	0.002
1/9	2/9	2/9	2/9	2/9		0.004	0.003	0.004	0.004
1/8	1/8	1/4	1/4	1/4		0.011	0.008	0.012	0.012
1/7	1/7	1/7	2/7	2/7		0.049	0.041	0.050	0.050
0	0	0	1/2	1/2	1000	0.048	0.048	0.046	0.053
0	0	1/3	1/3	1/3		0.012	0.011	0.012	0.012
0	0	1/5	2/5	2/5		0.050	0.047	0.050	0.051
0	1/4	1/4	1/4	1/4		0.004	0.004	0.005	0.005
0	1/7	2/7	2/7	2/7		0.011	0.010	0.012	0.012
0	1/6	1/6	1/3	1/3		0.049	0.046	0.050	0.050
1/5	1/5	1/5	1/5	1/5		0.002	0.002	0.002	0.002
1/9	2/9	2/9	2/9	2/9		0.005	0.005	0.005	0.005
1/8	1/8	1/4	1/4	1/4		0.012	0.011	0.012	0.012
1/7	1/7	1/7	2/7	2/7		0.050	0.046	0.050	0.051

Table 2
Power of 0.05 level permutation, LRT, and Wald IU tests for supremacy, for different choices of k, cell probabilities π, and sample size n. Estimates are based on 100,000 replications

p_1	p_2	p_3	p_4	p_5	n	mid-P	Binomial	LRT	Wald IUT
0	0	0	0.45	0.55	50	0.200	0.129	0.199	0.199
0	0	0	0.35	0.65		0.726	0.622	0.725	0.725
0	0	0.30	0.30	0.40		0.085	0.069	0.085	0.102
0	0	0.20	0.20	0.60		0.917	0.892	0.919	0.926
0	0.10	0.10	0.35	0.45		0.196	0.160	0.196	0.206
0	0.10	0.10	0.25	0.55		0.783	0.737	0.784	0.795
0	0.10	0.10	0.15	0.65		0.996	0.994	0.996	0.997
0.10	0.10	0.10	0.30	0.40		0.209	0.169	0.209	0.223
0.10	0.10	0.10	0.20	0.50		0.828	0.784	0.827	0.841
0.10	0.10	0.10	0.10	0.60		0.997	0.995	0.997	0.997
0	0	0	0.45	0.55	200	0.416	0.363	0.418	0.418
0	0	0	0.35	0.65		0.997	0.995	0.997	0.997
0	0	0.30	0.30	0.40		0.354	0.322	0.363	0.363
0	0	0.20	0.20	0.60		1.000	1.000	1.000	1.000
0	0.10	0.10	0.35	0.45		0.476	0.447	0.477	0.480
0	0.10	0.10	0.25	0.55		1.000	1.000	0.999	0.999
0	0.10	0.10	0.15	0.65		1.000	1.000	1.000	1.000
0.10	0.10	0.10	0.30	0.40		0.518	0.486	0.521	0.524
0.10	0.10	0.10	0.20	0.50		1.000	1.000	1.000	1.000
0.10	0.10	0.10	0.10	0.60		1.000	1.000	1.000	1.000
0	0	0	0.45	0.55	1000	0.932	0.932	0.932	0.939
0	0	0	0.35	0.65		1.000	1.000	1.000	1.000
0	0	0.30	0.30	0.40		0.971	0.968	0.972	0.972
0	0	0.20	0.20	0.60		1.000	1.000	1.000	1.000
0	0.10	0.10	0.35	0.45		0.970	0.968	0.971	0.972
0	0.10	0.10	0.25	0.55		1.000	1.000	1.000	1.000
0	0.10	0.10	0.15	0.65		1.000	1.000	1.000	1.000
0.10	0.10	0.10	0.30	0.40		0.984	0.983	0.984	0.984
0.10	0.10	0.10	0.20	0.50		1.000	1.000	1.000	1.000
0.10	0.10	0.10	0.10	0.60		1.000	1.000	1.000	1.000

In Table 2, we report the estimated power. As expected, power increases with the sample size n and the distance of p_k from the other cell probabilities. The exact binomial approach in (3) is a little less powerful, but the differences in terms of power become negligible as n increases.

6. Turtle Data

We now turn to our motivating example. Loggerhead sea turtles *Caretta caretta* were incidentally caught during fishing activities in the Mediterranian sea and rehabilitated in rescue centres, where they experienced different periods of captivity.

Just before being released, at the end of rehabilitation, each turtle underwent our experiment. At each run, the turtle was placed in a test tank filled with sea water. Three baits were placed in the pool (one red, one yellow, one blue), and the run was stopped if the turtle reached one of the baits for eating (not if the bait was just hit), or after 10 min. It was recorded whether the turtle had eaten a bait before the 10th min, the color of the bait that was eaten (if one had been), and the experimental setting.

We restrict to the subset of the whole data related to the first bait eaten by each turtle, finally having a total of $n = 26$ first bites. We thus have no repeated measurements in the subset we presently use. The experimental question is whether the red bait is the least likely to be bit as a first choice.

We recorded $n_1 = 12$ turtles biting a blue bait, $n_2 = 10$ biting a yellow bait, and $n_3 = 4$ turtles biting a red bait. Hence, in this example, $k = 3$.

If we ignore covariates, we get a mid-P value of $p = 0.0592$, failing to reject the hypothesis of inferiority of the cell corresponding to the red bait at the $\alpha = 0.05$ level. Also, the IU Wald and LRT tests fail to reject the null hypothesis.

We have two covariates. The first covariate is an indicator of the presence of chemical cues, introduced by the experimenter, in the water. The second covariate is the length of captivity. The covariates are known confounders for visual preferences of biting behaviour of turtles. The presence of chemical cues may obviously change biting behaviour, for instance stimulating hunger of the turtle and making it give less importance to color of the bait. Length of captivity is a confounder since, during captivity, turtles are fed somewhat differently from how they would get food in nature, so that they may develop a different sensitivity to preys.

If we include the chemical cues in the model and test for inferiority of the third cell, we finally get $p = 0.002$ for an omnibus conditional test. A universal test instead is not rejected since $p = 0.758$. The two groups are clearly different: without chemical cues, there is an effect of the color of the bait, and the red bait is the least preferred for the first bite with a Bonferroni adjusted $p = 0.004$ (in fact, it is never chosen while the other colors are chosen eight times each). With chemical cues, color does not seem to matter anymore and the red baits are chosen four times (as many as the blue and twice as the yellow).

We now use the multinomial logit formulation (5), with a sum-to-zero parameterization, thus performing an adjusted test. We obtain a p-value of 0.003 with category specific regression parameters and 0.002 with common regression parameters. The geometric average of the probability of biting a red bait conditional on the presence or absence of chemical cues can be claimed to be inferior to the other geometric averages. The estimated geometric average is actually zero, against 0.45 for blue and 0.32 for yellow baits.

We now consider the numerical covariate, which is respectively centered at zero (no captivity), 6 months, its mean of 3.5 months, and its median of 1 month. In all cases, with and without restricting to cell specific regression parameters, we get p-values well above $\alpha = 0.05$.

We finally consider using both the numerical and categorical covariates. We restrict to the parameterization with equal regression parameters across categories, and center at zero the length of captivity. We use a corner point parameterization and reach the same conclusions about biting preferences without chemical cues, with $p = 0.002$; and $p = 0.52$ when we reverse the corner point parameterization. With a sum-to-zero constraint we get $p = 0.013$. We essentially reach the same conclusions no matter where we center the length of captivity.

We conclude that red baits can be claimed to be significantly uniformly less preferred than blue or yellow baits by loggerhead turtles when there are no chemical cues, and for any tested value of the length of captivity. Conditionally on the presence of chemical cues, we can not claim inferiority of the red bait and the color does not actually seem to matter.

7. Conclusions

We proposed a family of IU tests for testing supremacy or inferiority, also after adjusting for effects of covariates in a joint or conditional multivariate setting. The method is widely applicable.

We would like to conclude stressing that we have set up different settings for testing supremacy or inferiority hypotheses. These settings are different in nature and interpretation. We summarize the six settings considered.

- Univariate case. We have a single categorical response, we would like to verify supremacy or inferiority of a pre-specified cell probability.
- Joint multivariate case. There are R response variables, and it is of interest to prove that one cell probability in the R-dimensional contingency table is the largest (or the smallest). In this case, due to likely sparsity of the contingency table, the permutation approach should be used.
- Specific conditional case. We want to test that a multinomial cell probability is the largest (or the smallest) conditionally on a given value for one or more covariates. This reduces to testing on the intercepts of a model as (5), with a corner point parameterization for categorical covariates and an opportune centering of continuous covariates.
- Omnibus conditional case. The response variable is stratified within levels of one or more categorical covariates. We are interested in showing that a cell probability is the largest (or the smallest) in *at least one* stratum. If the test is rejected, one could perform multiple comparisons to detect the strata for which the cell probability of interest is the largest (or the smallest)
- Adjusted conditional case. The response variable is stratified within levels of one or more categorical covariates. We are interested in showing that the geometric average (across strata) of a cell probability is the largest (or the smallest) of the geometric average (across strata) of each cell probability. This reduces to testing on the intercepts of a model as (5), with a sum-to-zero parameterization.
- Universal conditional case. The response variable is stratified within levels of one or more categorical covariates. We are interested in showing that a cell probability is the largest (or the smallest) within *all* strata.

Acknowledgments

The author is grateful to the Acquiring Editor and two anonymous referees for many constructive comments.

References

Agresti, A. (2002). *Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons.

Alam, K., Thompson, J. R. (1972). On selecting the least probable multinomial event. *Ann. Mathemat. Statist.* 43:1981–1990.

Berger, R. L., Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statist. Sci.* 11:283–302.

Berry, J. C. (2001). On the existence of a unique most probable category. *J. Statist. Plann. Infer.* 99:175–182.

Casella, G., Berger, R. L. (2002). *Statistical Inference*. Pacific Grove, CA: Duxbury.

Conover, W. J. (1998). *Practical Nonparametric Statistics*. New York: Wiley.

Good, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer.

Hochberg, Y., Tamhane, A. C. (1987). *Multiple Comparisons Procedures*. New York: Wiley.

Lancanster, H. O. (1961). Significance tests in discrete distributions. *J. Amer. Statist. Assoc.* 56:223–234.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *J. Roy. Statist. Soc. Ser. B* 42:109–142.

Nettleton, D. (2009). Testing for the supremacy of a multinomial cell probability. *J. Amer. Statist. Assoc.* 104:1052–1059.

Pesarin, F. (2001). *Multivariate Permutation Tests with Applications to Biostatistics*. New York: Wiley.

Piovano, S., Farcomeni, A., Giacoma, C. (2009). Chemical cues elicite biting feeding behaviour in loggerhead sea turtles. http://afarcome.interfree.it/ piov.pdf. Submitted.

Silvapulle, M. J., Sen, P. K. (2004). *Constrained Statistical Inference*. New York: Wiley.