*Quaderni di Statistica*
*Vol. 3, 2001*

# A comparison between two asymptotic tests for analysing preferences

Angela D'Elia
*Dipartimento di Scienze Statistiche, Università di Napoli Federico II*
*E-mail: angdelia@unina.it*

*Summary*: In this paper we focus on tests of hypotheses for the parameter of a model for ranks. In particular, we consider the situation of rankings $m$ items and we are interested in testing the existence of an indifference feeling towards a given item. Both a likelihood ratio test and a Wald test are developed, and their performances in finite samples are compared through a simulation study. These two asymptotic tests are also applied on a real dataset, originated from a marketing survey on the consumption of olive-oil.

*Key words*: Ranks, Preference parameter, Asymptotic tests.

## *1. Introduction*

Several methods exist for the statistical analysis of the preference data expressed as ranks of $m$ items (Marden, 1995; Taplin, 1997), but none of them is based on an explicit probabilistic model of the ranks themselves.

In this paper we review a statistical model we proposed for the ranks (D'Elia, 1999, 2000), and we investigate some inferential issues about its parameter, which could be considered as a *liking measure* for a given item (D'Elia, 2001b).

In particular, we focus the discussion on two asymptotic tests (likelihood ratio and Wald tests), in order to develop a tool for assessing the presence of an indifference feeling toward a given item.

The paper is organised as follows. In section 2 we briefly introduce the probabilistic model and we derive some inferential results about its parameter; then we propose a generalized linear model (GLM) for ranks data. In section 3 we develop both the likelihood ratio test and the Wald test for the parameter, while in section 4 we compare the performance of the two asymptotic tests in finite samples, running a Monte Carlo simulation study. Section 5 is devoted to the results of a survey about the Italian consumers preferences towards different kind of olive-oils. A discussion concerning further developments and possible extensions of the basic model ends the paper.

## *2. A statistical model for ranks*

In this section we briefly discuss: *i*) the probabilistic model for the ranks, *ii*) some inferential results, and *iii*) the corresponding GLM framework.

### *2.1 The probabilistic model*

Let $\mathcal{O}_1, ..., \mathcal{O}_j, ..., \mathcal{O}_m$ be a set of $m$ items (car or food brands, political parties, professions, colours, etc.) and let $r_1, ..., r_j, ..., r_m$ be the corresponding observed ranks. Besides, let us assume that $r_j = 1$ means the best, while $r_j = m$ means the worst in the raters' opinion.

If we consider a single item at a time, say $\mathcal{O}$ (dropping out the index $j$), then the corresponding observed rank $r$ can be thought as of the realization of a particular[1] Inverse Hypergeometric random variable (r.v.): $R \sim IHG(\theta, m), \theta \in \Theta$, whose probability mass function is:

$$Pr(R{=}r) = \begin{cases} \theta, & r = 1, \\ c_r\, \theta(1-\theta)^{r-1} \prod_{s=1}^{r-1}(m - s - 1 + s\theta)^{-1}, & r = 2, \ldots, m, \end{cases}$$

where $c_r = \prod_{s=1}^{r-1}(m-s) = (m-1)!/(m-r)!,\ r = 2, \ldots, m$, and whose

[1]The peculiarity of "our" Inverse Hypergeometric r.v. with respect to the standard one (Wilks, 1963; Guenther, 1975) is that we allow for a continuous parameter space $\Theta = \{\theta : 0 < \theta \leq 1\}$, while in the current literature the probabilistic model is related to a parameter $B = 1, 2...$, which is discrete.

expectation and variance, respectively, result:

$$\mathbb{E}(R) = \frac{m-\theta}{1+\theta(m-2)}, \qquad Var(R) = \frac{(m-1)^2\theta(1-\theta)(m-\theta)}{[1+(m-2)\theta]^2[2+(m-3)\theta]}.$$

Since $\theta = Pr(R=1)$, and moreover $\mathbb{E}(R) \to 1$ when $\theta \to 1$, while $\mathbb{E}(R) \to m$ when $\theta \to 0$, we can consider the parameter $\theta$ as a *liking measure* for the item $\mathcal{O}$.

Then, making inference on $\theta$ means to understand how much the population does prefer the considered item.

### *2.2 The estimation of* $\theta$

For a given the item $\mathcal{O}$, let **r**=$(r_1, r_2, ..., r_n)'$ be the observed ranks that a sample of $n$ raters have assigned to it. From the previous probabilistic model for $R$, the log-likelihood function results:

$$l(\theta, \mathbf{r}) \propto n\, log(\theta) + (S_n - n) log(1-\theta) - \sum_{i=1}^{n} \delta(r_i) \sum_{s=1}^{r_i-1} log(m - s - 1 + s\theta),$$

where $S_n = \sum_{i=1}^{n} r_i$ and

$$\delta(r_i) = \begin{cases} 1, & r_i > 1, \\ 0, & \text{elsewhere.} \end{cases}$$

It is worth noticing that the amount of information provided by **r** (the observed rank) is the same that the one given by the observed frequencies of $r$=1, ..., $r$=$m$, that is $n_1, ..., n_m$, respectively, where $\sum_{r=1}^{m} n_r = n$. Thus, for inferential purposes, we can use alike the following expression for the log-likelihood function:

$$l(\theta; n_1, ..., n_m) \propto n\, log(\theta) + (S_n - n) log(1-\theta) - n \sum_{r=1}^{m-1} h_r(\theta)(1 - F(r)),$$

where $h_r(\theta) = log(m - r - 1 + r\theta)$, and $F(r) = \sum_{k=1}^{r}(n_k/n)$, $r = 1, 2, ..., m$, is the empirical distribution function of the observed ranks. Of course, $S_n = \sum_{i=1}^{n} r_i = \sum_{r=1}^{m} r n_r$.

If we let $l(\theta; n_1, ..., n_m) = \sum_{r=1}^{m} n_r log(p_r(\theta))$, where

$$p_r(\theta) = Pr(R = r), \qquad r = 1, ..., m,$$

then the maximum likelihood estimator $T_n$ for $\theta$ is the solution of:

$$V(\theta) = \frac{\partial l(\theta; n_1, ..., n_m)}{\partial \theta} = \sum_{r=1}^{m} n_r \frac{p'_r(\theta)}{p_r(\theta)} = 0,$$

where $\frac{p'_r(\theta)}{p_r(\theta)} = \frac{\partial log\{p_r(\theta)\}}{\partial \theta}$ is the ratio of two polynomials in $\theta$. It is easy to show that the solutions of $V(\theta) = 0$ are the roots of an $(m-1)$-degree polynomial in $\theta$.

Moreover, it can be shown (D'Elia, 2001b) that $V(\theta) = 0$ has always a single real root in the admissible interval $(0, 1]$.

An important feature of this model is that, while an explicit expression for $T_n$ can be derived only in specific cases, we are able to obtain the asymptotic expression for the variance of the maximum likelihood estimator $T_n$. Thus, we can exploit this result for constructing asymptotic confidence intervals and for developing a Wald test for the parameter $\theta$.

Indeed, the observed frequencies $n_1, ..., n_m$ are realizations of a Multinomial r.v. $(N_1, ..., N_m) \sim MN(n, \mathbf{p})$, where $\mathbf{p} = (p_1(\theta), p_2(\theta), ..., p_m(\theta))'$. Thus, we have

$$\mathbb{E}\left(\frac{-\partial^2 l(\theta)}{\partial \theta^2}\right) = -\sum_{r=1}^{m} \frac{p''_r(\theta)p_r(\theta) - \{p'_r(\theta)\}^2}{\{p_r(\theta)\}^2}\mathbb{E}(N_r) = n\sum_{r=1}^{m} \frac{\{p'_r(\theta)\}^2}{p_r(\theta)},$$

since $\mathbb{E}(N_r) = np_r(\theta)$ and $\sum_{r=1}^{m} p''_r(\theta) = 0$. Finally, it follows that:

$$var(T_n) \simeq \frac{1}{n}\left(\sum_{r=1}^{m} \frac{\{p'_r(\theta)\}^2}{p_r(\theta)}\right)^{-1}.$$

The quantity $\sum_{r=1}^{m} \frac{\{p'_r(\theta)\}^2}{p_r(\theta)}$ can be easily obtained by means of an algorithm which exploits a recursive relation for computing $p_r(\theta)$, as we show in the Appendix.

### *2.3 The GLM framework*

In order to study how preferences, and then ranks, change with covariates values representing the main characteristics of the raters, it is useful to exploit our model in a GLM framework (D'Elia, 1999).

Let $\mathbf{X}$ be the design matrix ($n \times (p+1)$), where the first column is $\mathbf{1} = (1, 1, ..., 1)'$ and $x_{i,h}$ is the observed value of the $h$-th covariate in the $i$-th unit ($h = 2, ..., p+1; i = 1, 2, ..., n$), and let $\beta = (\beta_0, \beta_1, ..., \beta_p)'$ be the vector of unknown covariates coefficients.

Since $Pr(R = 1) = \theta$, then $\frac{\theta}{1-\theta}$ are the odds of $(R = 1)$ versus $(R \neq 1)$, and we let

$$\theta = \frac{1}{1 + e^{-\mathbf{X}\beta}} = \frac{e^{\mathbf{X}\beta}}{1 + e^{\mathbf{X}\beta}},$$

so that for $\mathbf{X}\beta \in (-\infty, \infty)$, $\theta \in [0, 1]$ as it is proper for a probability.

After a simple algebra, it follows that

$$\mathbb{E}(R) = 1 + \frac{m-1}{1 + (m-1)e^{\mathbf{X}\beta}},$$

pointing out that the expected rank changes inversely with respect to the predictor value $\mathbf{X}\beta$.

Exploiting the previous relations, the log-likelihood function for the GLM is:

$$l(\beta; \mathbf{r}, \mathbf{X}) = -\sum_{i=1}^{n} \sum_{j=1}^{r_i} log[(m-1)e^{\mathbf{x}_i'\beta} + m - j] + log(m) + \mathbf{x}_i'\beta,$$

that can be numerically optimized in order to get the maximum likelihood estimates of the $\beta$'s.

### *3. Two asymptotic tests*

In the statistical analysis of preferences data it is often interesting to check if the population feels indifference toward a given item. For example, in a marketing setting such finding would mean that a greater effort

must be done in order to increase the evaluation in the consumers and/or customers of the fixed item $\mathcal{O}$.

Formally, using the model shown in the previous section, this is equivalent, for a given item $\mathcal{O}$, to test that $\theta = 1/m$. It must be noticed that if $\theta = 1/m$ then the probability mass function results:

$$Pr(R = r) = \frac{1}{m}, \qquad r = 1, 2, ..., m,$$

that is the discrete Rectangular distribution, where all the ranks (1,2,...,$m$) have equal probability, and $\mathbb{E}(R) = (m+1)/2$.

In the following, we develop two asymptotic tests for the null hypothesis $H_0 : \theta = 1/m$ versus $H_1 : \theta \neq 1/m$.

### *3.1 The likelihood ratio test*

Let $l_{H_0}$ and $l_{ML}$ be the maximum of the log-likelihood function under $H_0$ and the maximum of the log-likelihood in $\Theta$, respectively. The test statistic results (D'Elia, 2001b):

$$\begin{aligned} -2\Lambda(\mathbf{r}) &= -2(l_{H_0} - l_{ML}) = \\ &-2 \left\{ \sum_{i=1}^{n} \delta(r_i) \sum_{s=1}^{r_i - 1} log \left( \frac{m - s - 1 + sT_n}{m - s - 1 + s/m} \right) + \right. \\ &\left. + S_n log \left( \frac{1 - 1/m}{1 - T_n} \right) + n \, log \left( \frac{(1 - T_n)}{T_n(m-1)} \right) \right\} \end{aligned}$$

or considering the observed frequencies $(n_1, ..., n_m)'$ and the empirical distribution function $F_n(r)$, $r = 1, \ldots, m$,

$$\begin{aligned} -2\Lambda(\mathbf{n}) &= \\ &-2n \left\{ \sum_{r=1}^{m-1} (1 - F_n(r)) \{ h_r(T_n) - h_r(1/m) \} + \right. \\ &\left. + \bar{R}_n \, log \left( \frac{1 - 1/m}{1 - T_n} \right) + log \left( \frac{(1 - T_n)}{T_n(m-1)} \right) \right\} \end{aligned}$$

where $T_n$ is the ML estimator of $\theta$ and $\bar{R}_n = S_n/n$ is the rank average.

It is well known, on the basis of the Wilks theorem, that the critical region of size $\alpha$ for rejecting $H_0$ results: $C_0(\alpha) = \{\mathbf{n} : -2\Lambda(\mathbf{n}) \geq \chi^2_{1,\alpha}\}$.

As far as concerns the power function, considering alternatives that are local, that is sequences: $\{\theta_n\} = \theta_0 + \frac{c}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$, such that $\theta_n$ converges to $\theta_0 = 1/m$, as $n \to \infty$, it can be shown (Cox and Hinkley, 1974, p. 318) that $-2\Lambda(\mathbf{n})$ asymptotically converges to a non-central $\chi^2_1$ distribution, with non-centrality parameter $\xi_n = ni(\theta_0)(\theta_n - \theta_0)^2$, where $ni(\theta_0) = \mathbb{E}\left(\frac{-\partial^2 l(\theta)}{\partial \theta^2}\right)_{|\theta=\theta_0}$.

Moreover, $-2\Lambda(\mathbf{n})$ can be approximated by $\{Z + (\xi_n)^{1/2}\}^2$, where $Z \sim N(0,1)$ is a standard Normal r.v. Consequently, the asymptotic power function can be obtained:

$$\gamma(\theta_n) \simeq 2 - \Phi(z_{\alpha/2} - \xi_n^{1/2}) - \Phi(z_{\alpha/2} + \xi_n^{1/2}),$$

where $z_{\alpha/2}$ is such that $\Phi(z_{\alpha/2}) = 1 - \alpha/2$, and $\Phi(z)$ is the distribution function of $Z \sim N(0,1)$.

Since in our case $\xi_n = n(\theta_n - 1/m)^2 \sum_{r=1}^{m} \frac{\{p'_r(\theta_0)\}^2}{p_r(\theta_0)}$, the power function results:

$$\gamma(\theta_n) \simeq 2 - \Phi\left\{ z_{\alpha/2} - (\theta_n - 1/m)\left[ n \sum_{r=1}^{m} \frac{\{p'_r(\theta_0)\}^2}{p_r(\theta_0)} \right]^{1/2} \right\} +$$

$$-\Phi\left\{ z_{\alpha/2} + (\theta_n - 1/m)\left[ n \sum_{r=1}^{m} \frac{\{p'_r(\theta_0)\}^2}{p_r(\theta_0)} \right]^{1/2} \right\}.$$

Obviously, all the above results are still valid for any simple hypothesis $H_0 = \theta = \theta_0$, but letting $\theta_0 = 1/m$ seems to us the most interesting case for real applications.

### *3.2 The Wald test*

The Wald test has an easier structure than the previous likelihood ratio test. Indeed, this is due to the fact that we were able to obtain an exact expression for the Fisher's information of the parameter $\theta$, as shown in section 2.

Then, in order to test $H_0 : \theta = \theta_0 = 1/m$ versus $H_1 : \theta \neq 1/m$, the Wald statistic $W = (T_n - \theta_0)^2 n i(T_n)$ becomes

$$W(\mathbf{r}) = n(T_n - 1/m)^2 \left( \sum_{r=1}^{m} \frac{\{p_r'(T_n)\}^2}{p_r(T_n)} \right).$$

Assuming the usual regularity conditions, under $H_0$, $W(\mathbf{r}) \xrightarrow{d} \chi_1^2$, so that the critical region of size $\alpha$ for rejecting $H_0$ results: $C_0(\alpha) = \{\mathbf{r} : W(\mathbf{r}) \geq \chi_{1,\alpha}^2\}$.

As far as concerns the asymptotic power of the Wald test, again considering alternatives that are local, that is sequences $\{\theta_n\}$ such that $\theta_n$ converges to $\theta_0 = 1/m$, and the fact that $i(T_n) \to i(\theta_0)$, as $n \to \infty$, it can be shown (Lehmann, 1999, p.160) that

$$\gamma(\theta_n) \simeq 2 - \Phi\left( z_{\alpha/2} - \frac{\sqrt{n}(\theta_n - \theta_0)}{\sqrt{i(\theta_0)}} \right) - \Phi\left( z_{\alpha/2} + \frac{\sqrt{n}(\theta_n - \theta_0)}{\sqrt{i(\theta_0)}} \right).$$

Then, substituting $\theta_0 = 1/m$ and $i(\theta_0) = \sum_{r=1}^{m} \frac{\{p_r'(\theta_0)\}^2}{p_r(\theta_0)}$ we get the same asymptotic expression obtained for the power function of the likelihood ratio test.

Of course, the asymptotic equivalence between the likelihood ratio test and the Wald test is a well known result (see for example: Lehmann, 1999, pp. 530-531), but their performances in finite sample may be quite different, and should be investigated by means of simulation studies, as we do in the next section.

### *4. A comparison of the tests in finite samples*

In this section we present the results we obtained from a Monte Carlo simulation study, that was performed in order to compare the powers of the likelihood ratio and of the Wald tests for the hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, in finite samples.

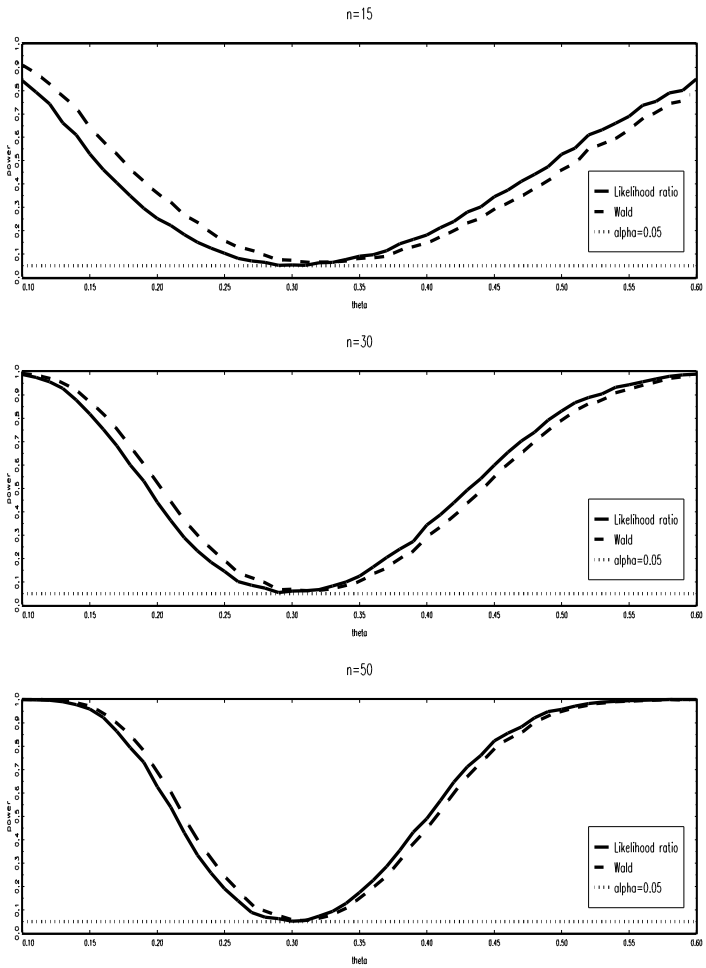


*Figure 1. Power functions of the likelihood ratio test and Wald test for different sample sizes.*

The simulation study was so planned:

- we generated[2] 5000 samples of size $n = 15$, from an $IHG(\theta, m)$ r.v. with $m = 3$;
- for a given $\alpha = 0.05$, we let $\theta_0 = 0.3$ and, varying the value of $\theta \in [0.1, 0.6]$, we computed the empirical powers of the likelihood ratio and of the Wald tests with respect to each alternative value;
- the previous steps were repeated for the cases of $n = 30$ and $n = 50$, respectively.

The plots of the power functions for both the tests, for $n = 15, 30, 50$, are shown in Figure 1, with a fixed baseline at $\alpha = 0.05$.

Since the alternative hypothesis is bidirectional, there is not a uniformly most powerful (UMP) test: indeed, as it can be noticed, whatever the sample size, the Wald test is more powerful for $\theta < 0.3\,(= \theta_0)$, while the likelihood ratio test is more powerful for $\theta > 0.3\,(= \theta_0)$. Of course, these results were confirmed also for different values of $\theta_0$, in other simulation trials.

Anyway, for increasing values of $n$, the two empirical power functions become closer and closer, as it was expected since they should be asymptotically equivalent.

Moreover, being both the tests consistent, when $n = 50$ their the power functions increase steeply to 1 in correspondence of moderate departures from the value $\theta_0$=0.3.

## *5. Preferences in the Italian olive-oil market*

One of the main goal for the firms operating in the food market is to understand which aspects of a product the consumers consider more important, and then what kind of product they do prefer. In particular, such a knowledge can help the firms in deciding their marketing strategies, and in the development of more efficient product valorization policies.

[2] A description of the algorithm for generating pseudo-random numbers from an $IHG$ r.v. is given by D'Elia (2001b).

In the Italian food market a product that more than others seems to represent the "made in Italy style" is the extra-vergin olive-oil. For this reason, during these last years a great attention has been devoted to the knowledge of the preferences of consumers towards different kinds of olive-oils.

In this section we illustrate some results from a survey on the olive-oil consumers. A pilote study was conducted during autumn 2000 (D'Elia, 2001b), while a greater survey was performed during spring 2001. The results we discuss here refer only to the latter.

The survey was conducted on 300 consumers, that were interviewed in Milan, Rome and Naples (100 from each city). Each consumer was asked to make a ranking of 5 different kinds of olive-oil and we assumed that rank = 1 means the best, and rank = 5 means the worst. No ties were allowed in order to not encourage lazy behaviours of the raters.

The main features of olive-oils considered were the following:

- A: good taste,
- B: high nourishing values,
- C: famous brand
- D: quality certification,
- E: known geographical origin.

In Table 1, the estimates of the parameter $\theta$, with the corresponding standard errors, are shown on the whole sample of consumers and for each city separately.

As it can be noticed on the whole sample, all the consumers give great importance to the "good taste" ($\hat{\theta} = \widehat{Pr}(R = 1) = 0.577$), while they seem not at all interested in the circumstance that the product has got a quality certification ($\hat{\theta} = 0.111$).

These feelings are confirmed also considering the three samples separetely, but it emerges that the Northern citizens (Milan) have the same

preferences between olive-oils C and D, that is famous brand and quality certification, while the Southern consumers (Naples) give more importance to the reputation of the brand than to the presence of a quality certificate.

*Table 1. Preference parameter estimates*

| Olive-oil | A | B | C | D | E |
|---|---|---|---|---|---|
| $\hat{\theta}$ | 0.577 | 0.225 | 0.199 | 0.111 | 0.136 |
| $s.e.$ | *0.023* | *0.013* | *0.012* | *0.008* | *0.009* |
| | **Milan** | | | | |
| $\hat{\theta}$ | 0.566 | 0.221 | 0.143 | 0.137 | 0.156 |
| $s.e.$ | *0.040* | *0.023* | *0.017* | *0.016* | *0.179* |
| | **Rome** | | | | |
| $\hat{\theta}$ | 0.521 | 0.248 | 0.210 | 0.129 | 0.108 |
| $s.e.$ | *0.039* | *0.025* | *0.022* | *0.016* | *0.014* |
| | **Naples** | | | | |
| $\hat{\theta}$ | 0.659 | 0.211 | 0.262 | 0.074 | 0.149 |
| $s.e.$ | *0.041* | *0.021* | *0.026* | *0.011* | *0.166* |

We have, then, investigate the presence of an indifference feeling for the different kinds of olive-oil, by means of the likelihood ratio test ($-2\Lambda$) and of the Wald test (W) developed in the previous sections. The results are shown in Table 2.

Since the sample size is quite big ($n = 300$ for the whole sample and $n = 100$ for each city), both the tests lead to same conclusions about the hypothesis of indifference almost in every case.

In particular, in the whole sample we cannot reject the hypothesis of indifference for the olive-oil C (famous brand), while for the olive-oil B (high nourishing) the two test gives slight different conclusions: rejecting indifference by likelihood ratio test (*p-value*=0.045) and not rejecting by Wald (*p-value*=0.060). We observe the same results also for Rome: indeed, consumers of Rome feel indifference for the olive-oil C, while for the olive-oil B the hypothesis of indifference is rejected using the likelihood ratio test (*p-value*=0.040), but it is not on the basis of Wald test (*p-value*=0.056).

These results confirm, at least for these datasets, what we noticed in the simulation study (Section 4): that is, the likelihood ratio test is more powerful than the Wald test for values of $\theta > \theta_0 \, (= 0.2 = 1/5$, in our case): indeed, in the whole sample and in Rome subsample we have $\hat{\theta} = 0.225$ and $\hat{\theta} = 0.248$, respectively (Table 1).

Finally, with respect to both the consumers of Milan and of Naples, the hypothesis of indifference is not rejected for the olive-oil B (high nourishing) and the two tests agree in all the cases.

*Table 2. Test of indifference*

| Olive-oil | A | B | C | D | E |
|---|---|---|---|---|---|
| $-2\Lambda$ | 320.999 | 4.015 | 0.014 | 79.916 | 37.194 |
| $p-value$ | $< 10^{-72}$ | *0.045* | *0.906* | $< 10^{-19}$ | $< 10^{-9}$ |
| $W$ | 269.308 | 3.530 | 0.014 | 112.711 | 44.239 |
| $p-value$ | $< 10^{-60}$ | *0.060* | *0.906* | $< 10^{-26}$ | $< 10^{-11}$ |
| | **Milan** | | | | |
| $-2\Lambda$ | 103.689 | 0.899 | 8.897 | 11.790 | 5.260 |
| $p-value$ | $< 10^{-24}$ | *0.343* | *0.003* | *0.0006* | *0.022* |
| $W$ | 85.066 | 0.808 | 11.127 | 14.066 | 5.910 |
| $p-value$ | $< 10^{-20}$ | *0.369* | *0.0008* | *0.0002* | *0.015* |
| | **Rome** | | | | |
| $-2\Lambda$ | 86.372 | 4.197 | 0.199 | 15.621 | 29.457 |
| $p-value$ | $< 10^{-20}$ | *0.040* | *0.655* | $< 10^{-5}$ | $< 10^{-8}$ |
| $W$ | 69.002 | 3.663 | 0.187 | 19.606 | 42.029 |
| $p-value$ | $< 10^{-17}$ | *0.056* | *0.665* | $< 10^{-6}$ | $< 10^{-11}$ |
| | **Naples** | | | | |
| $-2\Lambda$ | 136.798 | 0.277 | 6.381 | 65.703 | 7.968 |
| $p-value$ | $< 10^{-31}$ | *0.598* | *0.012* | $< 10^{-16}$ | *0.005* |
| $W$ | 129.496 | 0.220 | 5.650 | 128.751 | 8.478 |
| $p-value$ | $< 10^{-30}$ | *0.624* | *0.017* | $< 10^{-20}$ | *0.004* |

### *6. Further developments*

In this paper we have shown the main results concerning a model for ranks to be used in order to study preferences data.

With respect to this issue, it seems to us that there is the need of a greater development of the GLM framework of this model. In particular, it should be interesting to develop GLM Mixed models for ranks.

Indeed, in several situations it can happen that the raters belong to separated clusters, so that the effect of belonging to a group can modify the expressed ranking, and this circumstance should be taken into account. This is the case, for example, of longitudinal marketing survey, when we must consider the correlation among the ranks expressed by the same rater during the time, or also the case of surveys on different groups of homogeneous typology of consumers (e.g discount stores buyers, specialized stores buyers, etc.).

In these situations, we suggest to develop a model whose linear predictor contains random effects and/or autocorrelation terms, too. Then, for example, exploiting the relations obtained in section 2, if we let $\theta = (1 + e^{-(\mathbf{X}\beta+\mathbf{Zu})})^{-1}$, a random effects model for the ranks might be specified in the following way:

$$\mathbb{E}(R) = 1 + \frac{m-1}{1+(m-1)e^{\mathbf{X}\beta+\mathbf{Zu}}},$$

where **u** is the random effect vector and **Z** is the design matrix for **u**.

Moreover, we aim also at proposing adequate measures of goodness of fit for this kind of models, together with suitable diagnostic tools.

Another extension of the model is related to a multivariate approach, in order to analyze at the same time the whole ranking of the $m$ items, as proposed in D'Elia (2001a). Indeed, it should be taken into account that any choice process is intrinsically multivariate, and, since a ranking is a permutation of the first $m$ integers, conditional relations among its components (the ranks) should be considered.

*Acknowledgments*: This work was financially supported by funds of the Dipartimento di Scienze Statistiche, Università di Napoli Federico II. The author is really grateful to Prof. Francesco de Stefano and Dr. Teresa Del Giudice from the Dipartimento di Economia e Politica Agraria, Università di Napoli Federico II, for making available the olive-oil dataset.

## *References*

Cox D.R. and Hinkley D. W. (1974) *Theoretical Statistics*, Chapman & Hall, London.

D'Elia A. (1999) A Proposal for Ranks Statistical Modelling, *Statistical Modelling* (Friedl H., Berghold A., Kauermann G. eds.), Graz - Austria, 468-471.

D'Elia A. (2000) Un Modello Lineare Generalizzato per i Ranghi: Aspetti Statistici, Problemi Computazionali e Verifiche Empiriche, *Italian Journal of Applied Statistics*, 12, 205-227.

D'Elia A. (2001a) A Multivariate Model for Studying Preference Data, *New Trends in Statistical Modelling*, (Klein B. and Korsholm L. eds) Odense - Denmark, 425-428.

D'Elia A. (2001b) A Statistical Model for Studying Preferences, *Preliminary Report, submitted for publication*.

Guenther W. C. (1975) The Inverse Hypergeometric - a Useful Model, *Statistica Neerlandica*, 29, 129-144.

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

Marden J. I. (1995) *Analyzing and Modeling Rank Data*, Chapman & Hall, London.

Taplin R. H. (1997) The Statistical Analysis of Preference Data, *Applied Statistics*, 46, 493-512.

Wilks S. S. (1963) *Mathematical Statistics*, J. Wiley & Sons, London.

***Appendix***

Here, we show how to compute in a fast way the quantity

$$i(\theta) = \sum_{r=1}^{m} \frac{\{p_r'(\theta)\}^2}{p_r(\theta)}.$$

For the $IHG(\theta, m)$ r.v. the following recursive formulas apply:

$$\begin{aligned} p_1(\theta) &= \theta, \\ p_{r+1}(\theta) &= p_r(\theta)a(\theta), \end{aligned}$$

where $a(\theta) = (1-\theta)\frac{m-r}{m-r-1+r\theta}$. Then,

$$\begin{aligned} q_1(\theta) = \ p_1'(\theta) &= 1, \\ q_{r+1}(\theta) = \ p_{r+1}'(\theta) &= q_r(\theta)a(\theta) + p_r(\theta)a'(\theta), \end{aligned}$$

where $a'(\theta) = \frac{(m-r)(1-m)}{(m-r-1+r\theta)^2} = b(\theta)$.

Of course, letting $c(\theta) = \frac{m-r}{m-r-1+r\theta}$, we have:

$$a(\theta) = c(\theta)(1-\theta), \qquad b(\theta) = c(\theta)\frac{1-m}{(m-r-1+r\theta)}.$$

Then, if we let:

$$\mathbf{v}_r = \begin{bmatrix} p_r(\theta) \\ q_r(\theta) \end{bmatrix}, \qquad \mathbf{D}_r = \begin{bmatrix} a(\theta) & 0 \\ b(\theta) & a(\theta) \end{bmatrix}, \qquad r = 1, 2, ..., m,$$

we get:

$$\mathbf{v}_1 = (\theta, 1)', \qquad \mathbf{v}_{r+1} = \mathbf{D}_r\mathbf{v}_r, \qquad r = 1, 2, ..., m-1.$$

The following procedure (written in Gauss language) can be used for an effective computation of the quantity $i(\theta)$:

```
PROC fishinf(th,m);
LOCAL p, q, r, c, a, b, infor;

p=zeros(m,1); q=p;
p[1]=th; q[1]=1;
r=1;
do while r<=m-1;
        den=m-r-1+r*th;
        c=(m-r)/den;
        a=c*(1-th);
        b=c*(1-m)/den;
        p[r+1]=p[r]*a;
        q[r+1]=q[r]*a+p[r]*b;
        r=r+1;
endo;
infor=(sumc((q^2)./p));

RETP(infor);
ENDP;
```