Quaderni di Statistica
Vol. 9, 2007

A general approach for modelling individual choices

Domenico Piccolo
Dipartimento di Scienze Statistiche, Università di Napoli Federico II
E-mail: domenico.piccolo@unina.it

Summary: In this article, we generalize the approach for deriving the Inverse HyperGeometric (IHG) random variable proposed by Ridout (1999). After reviewing some preliminary results, the paper focuses on a general relationship between a statement on the sequential odds of the choices and the probability distribution of preferences. Finally, some statistical consequences of these findings are discussed.

Keywords: Odds and probability, IHG random variable, Preferences distributions.

1. Introduction

The statistical approach to ordinal data is mainly based on Generalized Linear Models (GLM) proposed by McCullagh (1980) and discussed by Agresti (2002). In that context, for model specification, a relationship among log-odds of cumulative probability and a linear function of covariates is assumed.

From a different viewpoint, some models have been introduced in literature order to explain the behaviour of respondents when faced to multiple choices. In this vein, among others, D'Elia and Piccolo (2005) and Piccolo and D'Elia (2007) proposed a new class of models, MUB and CUB models, respectively, which proved to be useful in several fields of applications.

When ranks data are characterized by a unique mode, the Inverse HyperGeometric (IHG, henceforth) random variable is, instead, the proba-

bilistic model to be preferred. The properties of this random variable have been discussed by several Authors (Wilks, 1963; Guenther, 1975; Johnson *et al.*, 1992) as a model describing the discrete waiting time when drawings of balls from an urn are made without replacement. Several applications of IHG random variables in sampling theory (Guenther, 1969) and in capture-recapture methods (Bailey, 1951) have been discussed.

In different contexts, this proposal has been judged as adequate for modelling the choice mechanism among m alternatives, as in searching strategies based on memory (Ridout, 1999), learning theory (Hutchinson, 2001) and preference data modelling (D'Elia, 1999; 2003; Piccolo, 2000). In the last case, the IHG distribution has been used in ranking and rating studies where a strong adversion/liking is present. Specifically, for a marketing research area, Del Giudice and D'Elia (2001) confirmed that the IHG random variable is able to reproduce adequately the consumers' choices even if subjects' covariates are included in the model.

Indeed, the paper by Ridout (1999) is the starting point for our discussion. He considered the problem raised by Jolliffe and Jolliffe (1977): several coal tits were asked to locate a food source when released in a room that contained four feeders. The experiment consisted in registering the number of times the birds looked for the correct feeder. This problem is related to the learning theory and to "Answer-Until-Correct" test. Then, Ridout showed that an IHG model improves the fitting of data and allows for a neat and simple interpretation of the memory behavior for replicated choices.

This paper is organized as follows: in the next section we formally introduce the IHG random variable and briefly discuss some of its properties. Then, in section 3 we derive the general relationship among hypotheses on the odds of sequential choices and the probability of ranks and, in section 4, we specify the structure in some cases. Some concluding remarks, focusing on the statistical consequences of these results, end the paper.

2. The IHG random variable

We suppose that a set of m objects (items/services/etc.) has been defined, and r is the rank assigned to it by a single rater. Thus, our analysis concerns the *stated preference* towards a single object.

The number $m > 1$ of objects to be compared is generally known and fixed. Then, r is the observed value of a discrete random variable R defined on the support $\{1, 2, \ldots, m\}$. We assume that $r = 1$ means "most preferred" while $r = m$ means "least preferred".

The experiment can be parameterized in many ways: from an urn that contains β white balls and $m-1$ not-white balls, a ball is randomly drawn *without replacement*[1] until a white ball occurs. The random variable R represents the number of drawings needed for th eoccurrence of a "success". As long as β is large with respect to m, we expect low values for R, and thus we may relate this mechanism to the ordered choice problem.

In general, we let $\beta > 0$ any positive real; thus,

$$Pr\,(R = 1) = \frac{\beta}{\beta + m - 1} = \theta$$

is the probability of a "success", that is the probability to select a white ball in the first drawing.

The parameter θ is a *measure of liking/preference* in the ranking problem. The IHG random variable can be well parameterized either by $\beta \in (0, +\infty)$ or by $\theta \in (0, 1)$ as these parameters are related by the one-to-one mapping:

$$\theta = \frac{\beta}{\beta + m - 1} \quad \Longleftrightarrow \quad \beta = (m - 1)\,\frac{\theta}{1 - \theta}\,.$$

For any given $m > 1$, we denote this random variable as $R \sim IHG(\beta)$ or $R \sim IHG(\theta)$. Then, it is immediate to show that its probability mass function is defined by:

$$Pr\,(R = r) = \frac{\beta}{m} \prod_{s=1}^{r} \frac{m - s + 1}{m - s + \beta}\,, \qquad r = 1, 2, \ldots, m.$$

[1] In general, Wilks (1963) and Guenther (1973) define this random variable as the number of drawings until $k \geq 1$ white balls are selected. Notice that if the drawings are made *with replacement* R is a negative Binomial random variable.

From a computational point of view, in order to obtain the probability distribution of R, it is more efficient to exploit the recursive relationships:

$$\begin{cases} P_r\left(R=1\right) & = & \dfrac{\beta}{\beta+m-1}\,; \\ P_r\left(R=r+1\right) & = & P_r\left(R=r\right)\dfrac{m-r}{m-r-1+\beta}\,, r=1,2,\ldots,m-1. \end{cases}$$

If W is a Beta-Binomial random variable with parameters $\alpha=1$ and β, and we let $n=m-1$, then it can be shown that $W+1\sim IHG(\beta)$.

It is worth noticing that if $\beta=1$ (or $\theta=1/m$), then the IHG random variable is the Uniform discrete distribution over the support of the first m integers. Moreover, when $\beta=2$ (or $\theta=2/(m+1)$) the probability distribution of R is the linear decreasing function: $P_r\left(R=r\right)=2/m-r\,2/[m(m+1)]\,r$, $r=1,2,\ldots,m$. Thus, it decreases from $P_r\left(R=1\right)=2/(m+1)$ to $P_r\left(R=m\right)=2/[m(m+1)]$.

The IHG random variable has only the first m factorials moments $\mu_{(k)}$ different from 0:

$$\mu_{(k)}=\frac{\beta+m}{m}\,k!\prod_{s=1}^{k}\frac{1+m-s}{\beta+s}\,,\quad k=1,2,\ldots,m.$$

Specifically,

$$\mathbb{E}\left(R\right)=\frac{\beta+m}{\beta+1}=\frac{m-\theta}{1+\theta\left(m-2\right)}.$$

Some features of IHG distributions for varying β (or θ) and a given $m=9$ are illustrated in Figure 1. They confirm the following general result: for $\beta\neq1$ (or $\theta\neq1/m$) the mode is located at $R=1$ or $R=m$, depending on whether $\beta\gtrless1$ (or $\theta\gtrless1/m$).

From an inferential point of view, given a sample of n observed ranks $(r_1,r_2,\ldots,r_n)'$, all the relevant information for the parameter are contained in the vector of the observed frequencies $(n_1,n_2,\ldots,n_m)'$, where n_j is the number of subjects expressing the rank $(R=j)$, $j=1,2,\ldots,m$. As a matter of fact, the log-likelihood function is

$$\ell(\theta)=\sum_{r=1}^{m}n_r\,\log\,P_r\left(R=r\mid\theta\right).$$

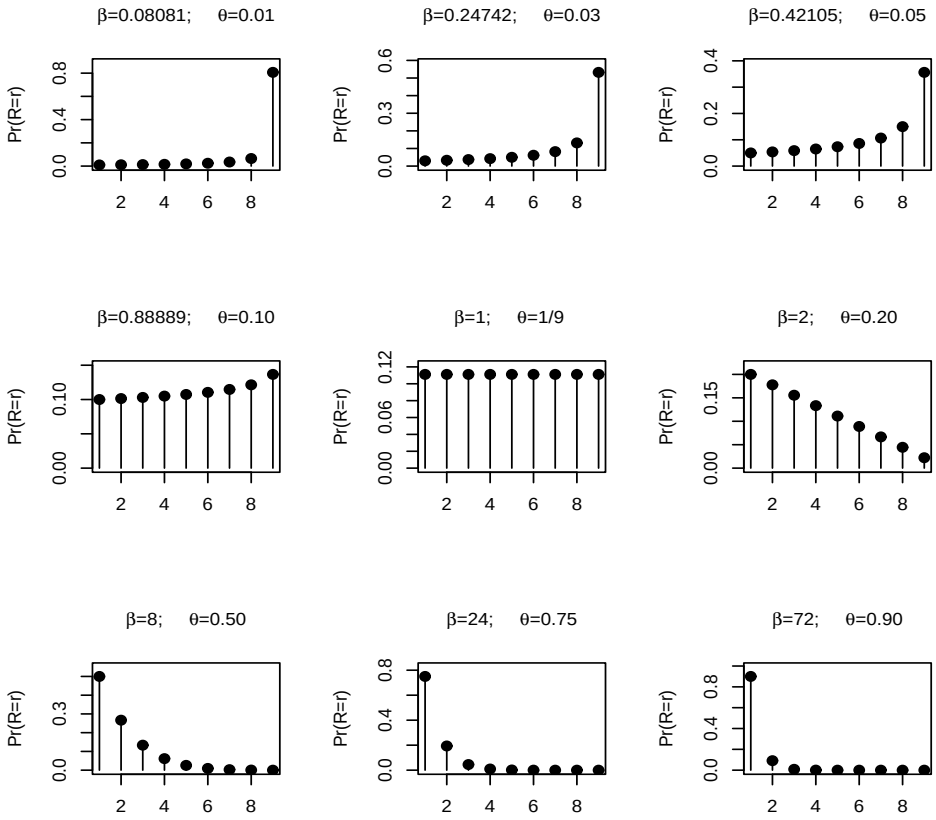

Figure 1. Inverse Hypergeometric probability mass functions $(m = 9)$.

Then, the sufficient statistic for θ is any subset of dimension $(m-1)$ of the vector of observed frequencies.

Several inferential aspects of IHG random variables are discussed by Guenther (1975) and more recently by Piccolo (2000, 2001) and D'Elia (2003). Moreover, denoting with $\overline{R}_n$ the sample mean of the ranks, D'Elia and Piccolo (2005) showed that the moment estimator of θ:

$$T_n = \frac{m - \overline{R}_n}{1 + (m-2)\,\overline{R}_n}\,,$$

is consistent and asymptotically Normal. In addition, it achieve a quite high relative efficiency with respect to the maximum likelihood estimator.

A final remark on the IHG distribution concerns the interpretation of θ: indeed, this parameter is a probability, thus it is scale invariant. However, its relevance in term of preference/liking strictly depends on m. Figure 1 shows that for $\theta > 1/9$ a positive feeling towards the object is expected whereas for $\theta < 1/9$ a certain degree of aversion is present. Consequently, if the estimate of θ is 0.2, say we cannot express any judgement on the preference without referring to m; indeed, only for $m > 5$ this estimate implies a liking.

3. Sequential odds and ranks probabilities

A consistent strategy for selecting an object from a set of different alternatives is a sequential choice. Thus, if we let $p_r = P_r(R = r)$, $r = 1, 2, \ldots, m$, the odds that the rank is r given that the object was not selected previously (rank is not less than r) is defined by:

$$\phi_r = \frac{p_r}{\sum_{k=r+1}^{m} p_k} = \frac{p_r}{1 - \sum_{k=1}^{r} p_k}, \; r = 1, 2, \ldots, m-1; \qquad \phi_m = 1. \quad (1)$$

Hereafter, $\phi_r, r = 1, 2, \ldots, m$ the *sequential odds* denotes of an ordinal choice. In this regard, with reference to the coal tits experiment, Ridout (1999, p.661) explains that: "*the correct feeder will be found at the r-th attempt, conditional on it not having been found at any previous attempt*". In fact, sequential odds are a discrete version of the well known *hazard function* for continuous random variable.

Since $\phi_m = 1$ and $p_m = 1 - p_1 - \cdots - p_{m-1}$, it is worth considering only the first $(m-1)$ odds and probabilities, respectively. Afterwards, we will show how these quantities are relater each other. However, sequential odds can also be defined by means of the distribution function $F(r) = P_r(R \leq r)$:

$$\phi_r = \frac{F(r) - F(r-1)}{1 - F(r)}, \; r = 1, 2, \ldots, m-1; \qquad \phi_m = F(m).$$

Of course, $F(0) = 0$ and $F(m) = 1$. By iterative substitution in this formula, after some algebra, the inverse relationships are deduced:

$$F(r) = \sum_{k=1}^{r} \frac{\phi_{r+1-k}}{\prod_{j=k+1-r}^{r} (1+\phi_j)},\ r = 1, 2, \ldots, m-1; \qquad F(m) = \phi_m.$$

Formula (1) shows that any assumption on a probability distribution implies well defined sequential odds. In this respect, it is interesting to solve the inverse problem by deriving the probability distribution which is implied by a stated assumption on the sequential odds. The following theorem shows the inverse mapping of (1) by computing the probability from the odds.

■ ***Theorem***

Given a set of sequential odds $\{\phi_r,\ r = 1, 2, \ldots, m-1\}$, the corresponding probabilities are obtained by:

$$p_r = \frac{\phi_r}{\prod_{k=1}^{r} (1+\phi_k)},\ r = 1, 2, \ldots, m-1; \qquad p_m = \frac{1}{\prod_{k=1}^{m-1} (1+\phi_k)}. \quad (2)$$

Proof

We prove the theorem by a repeated application of the induction principle on the algebraic formulation, although some induction on the corresponding matrix formulation of (1) and (2) is also possible.

By simple substitution, it is immediate to derive (2) for $r = 1, 2$. Then, suppose that (2) holds for some r, we have to prove (2) is true for $r + 1$. In other words, we have to prove that:

$$p_{r+1} = \frac{\phi_{r+1}}{\prod_{k=1}^{r+1} (1+\phi_k)} = \frac{\phi_{r+1}}{1+\phi_{r+1}} \frac{p_r}{\phi_r}.$$

Now, by letting $S_r = 1 - p_1 - \cdots - p_r$, the definition of the odds implies:

$$\phi_r = \frac{p_r}{S_r}; \qquad \phi_{r+1} = \frac{p_{r+1}}{S_r - p_{r+1}};$$

and, exploiting the second and the first identities, respectively, we get:

$$p_{r+1} = \frac{\phi_{r+1}}{1+\phi_{r+1}} S_r = \frac{\phi_{r+1}}{1+\phi_{r+1}} \frac{p_r}{\phi_r} .$$

The result shows the validity of the first part of (2) for $r = 1, 2, \ldots, m-1$.

Instead, in order to prove the formula for p_m, the induction will be based on a recursive formula for S_r, $r = 1, 2, \ldots, m-1$.

In fact, $p_m = S_{m-1} = 1 - p_1 - \cdots - p_{m-1}$. Then, a simple algebra shows that:

$$\begin{aligned} S_1 &= 1 - p_1 = 1 - \frac{\phi_1}{1+\phi_1} = \frac{1}{1+\phi_1}; \\ S_2 &= S_1 - p_1 = \frac{1}{1+\phi_1} - \frac{\phi_2}{(1+\phi_1)(1+\phi_2)} = \frac{1}{(1+\phi_1)(1+\phi_2)} . \end{aligned}$$

Now, assuming that, for some r, the following statement is true:

$$S_r = \frac{1}{\prod_{k=1}^{r} (1+\phi_k)} ,$$

we need to prove it for $r+1$. Specifically,

$$\begin{aligned} S_{r+1} &= S_r - p_{r+1} \\ &= \frac{1}{\prod_{k=1}^{r} (1+\phi_k)} - \frac{\phi_{r+1}}{\prod_{k=1}^{r+1} (1+\phi_k)} = \frac{1}{\prod_{k=1}^{r+1} (1+\phi_k)} ; \end{aligned}$$

and by replacing $r = m-2$ in the last expression, we get the final result.

4. Some sequential odds specifications

In this section, we discuss the consequences on the probability distribution of the ranks when some restrictions on the sequential odds are specificied. The assumptions that we are making are relevant in different

contexts and are obviously related to the number and nature of the objects to be selected.

In order to simplify the comparisons, Figures illustrated in the following subsections are plotted assuming $m = 9$.

4.1. Constant sequential odds

First of all, suppose that *sequential odds are constants*, that is:

$$\phi_r = c, \quad r = 1, 2, \ldots, m-1.$$

Then, from theorem of section 3, we get:

$$p_r = \frac{c}{(1+c)^r}, \quad r = 1, 2, \ldots, m-1; \qquad p_m = \frac{1}{(1+c)^{m-1}}.$$

If we let $c = \delta/(1+\delta)$, and then $\delta = c/(1+c)$, we find that constant odds imply the following probability distribution:

$$p_r = \delta\,(1-\delta)^{r-1}, \quad r = 1, 2, \ldots, m-1; \qquad p_m = (1-\delta)^{m-1}.$$

This is a *geometric distribution* over the finite support $\{1, 2, \ldots, m-1\}$ with a constrained p_m value in order to preserve the unit sum.

Figure 2 shows some distributions for varying values of c and δ. It is interesting to consider that the bump in the shape of the implied probability distributions would remain at the extreme values of R, for any choice of δ. Specifically, mode is at $R = 1$ for $c \leq 1$ (that is, $\delta \leq 0.5$), while mode is at $R = m$ for $c > 1$ (that is, $\delta > 0.5$). Finally, if $c = 1$, we have e regularly decaying geometric distribution as $p_r = (0.5)^r$, $r = 1, 2, \ldots, m$.

4.2. Hyperbolic sequential odds

Assume that the *odds* of choosing an object, given it was not selected previously, *increase hyperbolically* at any trial, proportionally to some constant $\beta > 0$, that is:

$$\phi_r = \frac{\beta}{m-r}, \quad r = 1, 2, \ldots, m-1.$$

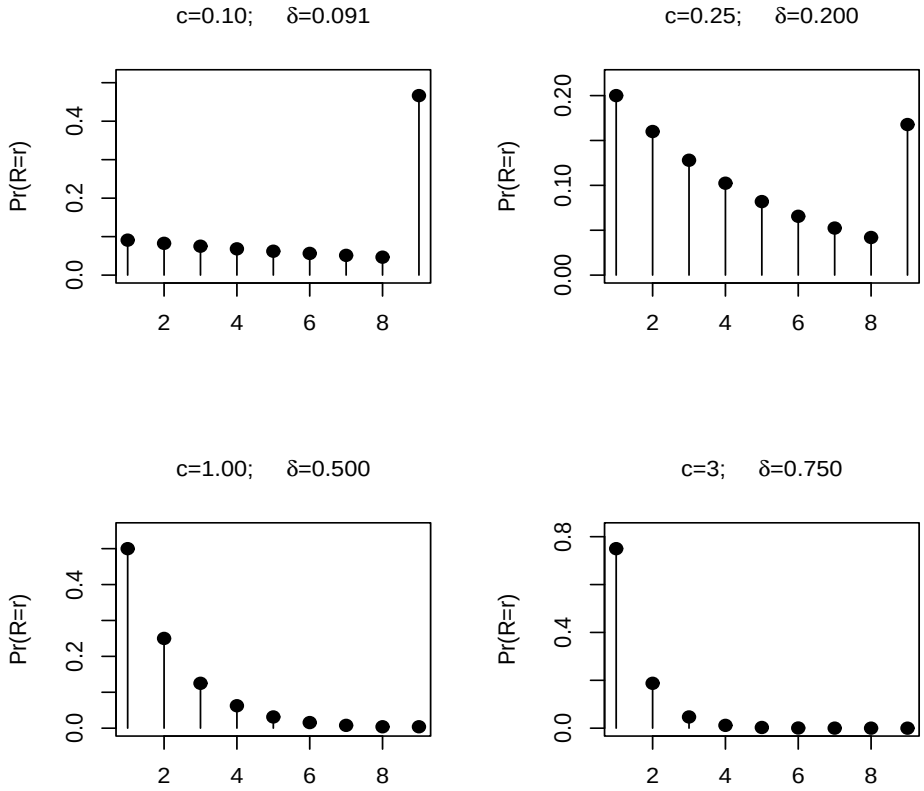

Figure 2. Probability distributions for constant sequential odds.

By theorem of section 3, this choice implies:

$$\begin{aligned} p_r &= \frac{\dfrac{\beta}{m-r}}{\displaystyle\prod_{k=1}^{r}\left(1+\frac{\beta}{m-k}\right)} = \frac{\beta}{m-r}\prod_{k=1}^{r}\frac{m-k}{\beta+m-k} \\ &= \frac{\beta}{m}\prod_{k=1}^{r}\frac{1+m-k}{\beta+m-k}\,, \quad r=1,2,\ldots,m\,. \end{aligned}$$

This is the IHG distribution with parameter β, as discussed in section 2. We observe that, under this assumption, the formula for p_r is correct also when $r = m$.

In this regard, we notice that if the remaining $(m - r)$ choices are randomly selected, then $\beta = 1$ and IHG distribution reduces to Uniform on the support $\{r = 1, 2, \ldots, m\}$.

Of course, if one parameterizes the distribution by θ, the equivalent assumptions on the sequential odds are:

$$\phi_r = \frac{m-1}{m-r}\,\frac{\theta}{1-\theta}, \quad r = 1, 2, \ldots, m-1.$$

For varying values of β, Figure 3 shows the sequential odds that correspond to IHG probability distributions reported in Figure 1. Odds are regularly increasing with r with a unique maximum at $R = m$ when $\beta < 1$, two maxima at $R = m - 1$ and $R = m$ when $\beta = 1$ and a unique maximum at $R = m - 1$ when $\beta > 1$.

4.3. Linear sequential odds

Assume that *odds increase linearly*, with a relative change $\alpha > 1$, that is:

$$\phi_r = \alpha\, r - 1, \quad r = 1, 2, \ldots, m-1.$$

Then, theorem 3 asserts that the implied probability distribution is defined by:

$$p_r = \frac{\alpha\, r - 1}{\alpha^r\, r!}, \quad r = 1, 2, \ldots, m-1; \qquad p_m = \frac{1}{\alpha^{m-1}\,(m-1)!}.$$

This is a sort of linear-exponential distribution and the relationships among the parameter α and the shape of the corresponding probability distributions are enhanced in Figure 4.

It turns out that the shape of the distribution is monotonically decaying as long as $\alpha > 1 + 2^{-1/2} \simeq 1.707$; otherwise, we observe a single mode at $R = 2$. Finally, when $\alpha = 1 + 2^{-1/2}$ there are two modes at $R = 1, 2$.

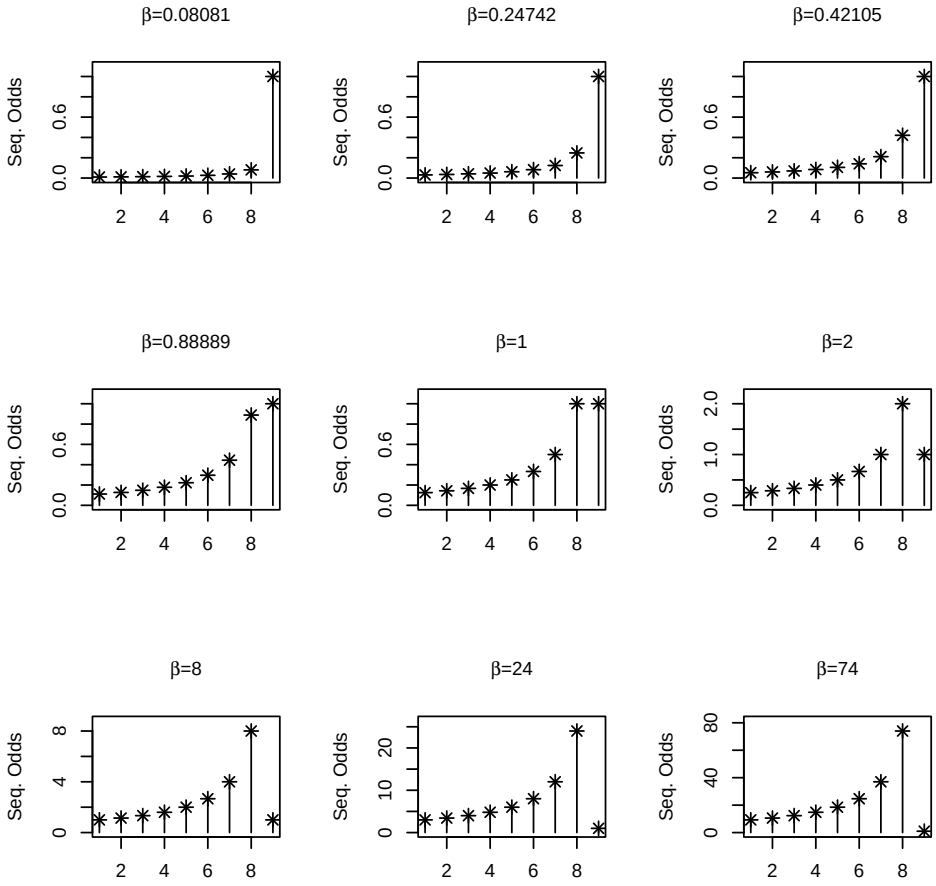

Figure 3. Hyperbolic sequential odds of IHG distributions of Figure 1.

4.4. J-shaped sequential odds

Assume that the *odds behave as a J-shaped function*, defined by:

$$\phi_r = \frac{r}{m-r}, \quad r = 1, 2, \ldots, m-1.$$

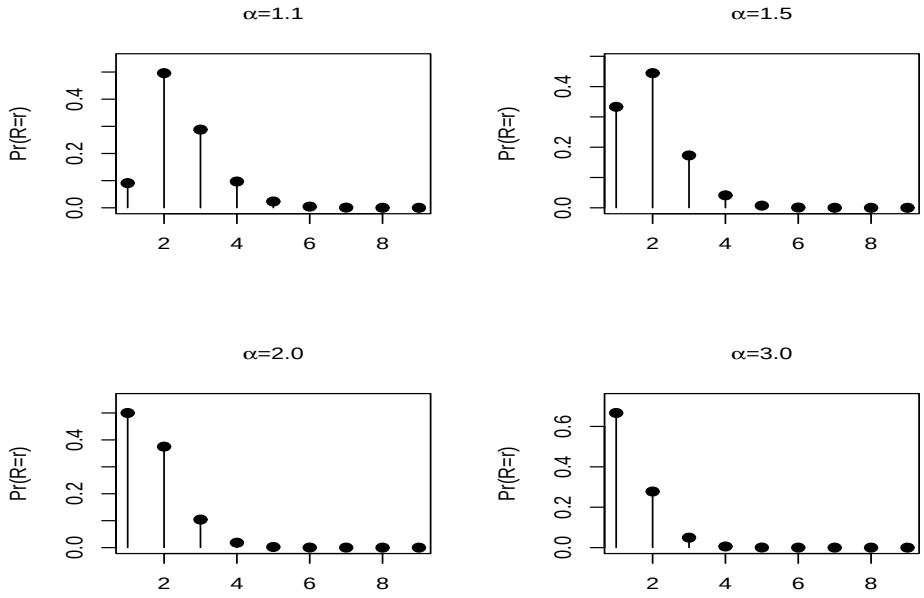

Figure 4. Probability distributions for linear sequential odds.

After some algebra, it is immediate to show that the implied probability distribution of ranks is:

$$p_r = r \binom{m-1}{r-1} \left(\frac{1}{m}\right)^m, \quad r = 1, 2, \ldots, m-1; \qquad p_m = \frac{(m-1)!}{m^{m-1}}.$$

This random variable is unique (as there are no parameters in its formulation) and it is a sort of Binomial-type distribution. Figure 5 shows its behaviour for some selected values of m. We observe that, for any m, the unique mode is at $R = 2$.

4.5. U-shaped sequential odds

Finally, we assume again that $m = 9$ and study the implications of *U-shaped sequential odds*, defined in Table 1 for different specifications,

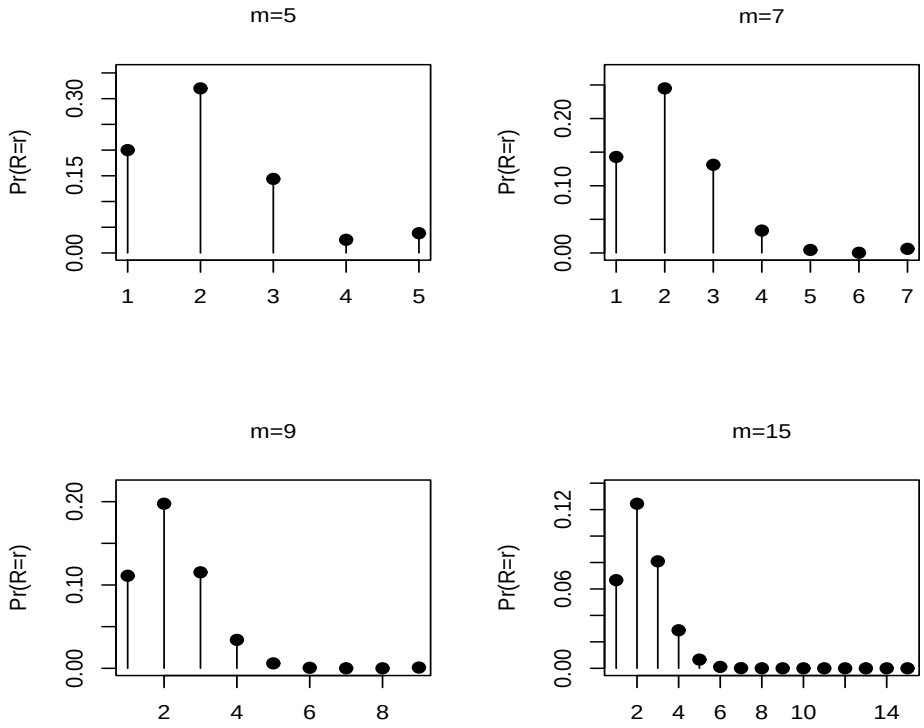

Figure 5. Probability distributions for J-shaped sequential odds.

denoted as A, B, C, D, respectively. Then, in Figure 6 we show the corresponding probability distributions of ranks.

Table 1. U-shaped sequential odds (numerically specified).

	$r \rightarrow$	1	2	3	4	5	6	7	8	9
A:	$\phi_r \rightarrow$	5	4	3	2	1	2	3	4	1
B:	$\phi_r \rightarrow$	50	40	30	20	10	20	30	40	1
C:	$\phi_r \rightarrow$	1	1	0.7	0.5	0.3	0.5	0.7	1	1
D:	$\phi_r \rightarrow$	3	2	1.5	1	0.5	0.1	0.5	1	1

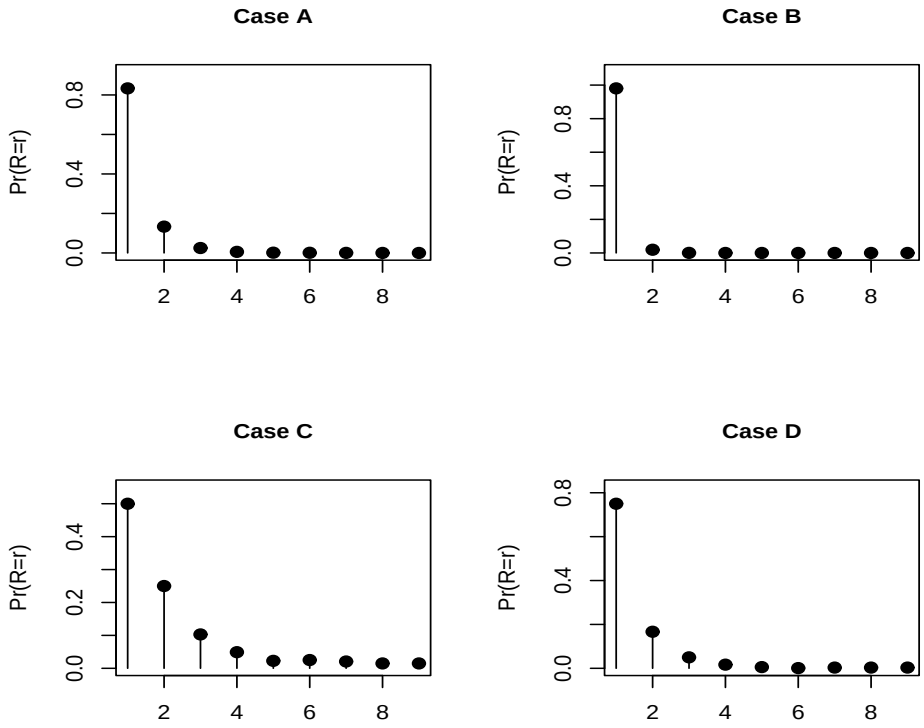

Figure 6. Probability distributions for U-shaped sequential odds.

We conclude that a sharp mode at $R = 1$ is always implied by these assumptions except than in case C, where sequential odds are more homogeneous. In this last case, the probability distribution is similar to that implied by constant sequential odds (as depicted in the left bottom panel of Figure 2).

5. Concluding remarks

In this paper we focused on some relationships among sequential odds and probability distributions for ordinal choices. The results confirm that

any specification about the behaviour of subjects making discrete choices implies a well defined probability distribution; in some cases, simple hypotheses may also induce sharp distributions.

In particular, the IHG distribution has been derived by imposing an hyperbolic increasing odds. Even if it seems a reasonable assumption, it strictly implies an extreme mode, thus this model is only convenient when a well defined liking/disliking is evident from data.

Further studies can be pursued with reference to inferential issues derived bt the stated assumptions about sequential odds. Specifically,we have to check the following suggestions:

1. Sequential odds may be consistently estimated by:

$$\hat{\phi}_r = \frac{n_r}{n - \sum_{k=1}^{r} n_k},\ r = 1, 2, \ldots, m-1.$$

   Then, if stated odds $\phi_r(\boldsymbol{\theta})$ are functions of a parameter vector $\boldsymbol{\theta}$, it may be estimated by introducing some distance criterion as:

$$min! \sum_{j=1}^{m-1} \left[\hat{\phi}_r - \phi_r(\boldsymbol{\theta})\right]^2 .$$

2. A more direct approach expresses the log-likelihood function as a function of sequential odds $\phi_r(\boldsymbol{\theta})$:

$$\begin{aligned} \ell(\boldsymbol{\theta}) \ &= \ \sum_{r=1}^{m-1} n_r \log \phi_r(\boldsymbol{\theta}) - \sum_{r=1}^{m-1} n_r \sum_{k=1}^{r} \log(1 + \phi_k(\boldsymbol{\theta})) \\ &\quad -n_m \sum_{r=1}^{m-1} \log(1 + \phi_k(\boldsymbol{\theta})). \end{aligned}$$

Finally, subjects' covariates may be introduced in the specification of sequential odds in the same logic of GLM approach.

Acknowledgements: The work has been supported by PRIN-2006 research project: "Stima e verifica di modelli statistici per l'analisi della soddisfazione degli studenti universitari" and benefited from research structures of CFEPSR, Portici. Critical suggestions by the Editor and a referee are gratefully acknowledged.

References

Agresti A. (2002), *Categorical Data Analysis*, 2^{nd} edition, J. Wiley & Sons, New York.

Bailey N.T. (1951), On estimating the size of mobile populations from recapture data, *Biometrika*, 38, 293-306.

Del Giudice T., D'Elia A. (2001), Valorizzazione dell'olio extra-vergine di oliva meridionale: una proposta metodologica per l'analisi delle preferenze, *Rivista di Economia Agraria*, LVI, 571-609.

D'Elia A. (1999), A proposal for ranks statistical modelling, in: H. Friedl, A. Berghold and G. Kauermann, eds., *Statistical Modelling*. Proceedings of 14th IWSM. University of Graz, Austria, 468-471.

D'Elia A. (2003), Modelling ranks using the Inverse Hypergeometric distribution, *Statistical Modelling: an International Journal*, 3, 65-78.

D'Elia A., Piccolo D. (2005), A mixture model for preference data analysis, *Computational Statistics & Data Analysis*, 49, 917-934.

Guenther W.C. (1969), Modified sampling, Binomial and Hypergeometric cases, *Technometrics*, 11, 639-647.

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

Hutchinson T.P. (2001), Partial knowledge and answer-until-correct tasks in birds and humans, *Biometrics*, 57, 1251-1252.

Johnson N.L., Kotz S. and Kemp A.W. (1992), *Univariate Discrete Distributions*, 2^{nd} edition, J. Wiley & Sons, New York.

Jolliffe and Jolliffe (1997), Modelling memory in coal tits: an illustration of the EM algorithm, *Biometrics*, 53, 1136-1142

McCullagh P. (1980), Regression models for ordinal data (with discussion), *Journal of the Royal Statistical Society, Series B*, 42, 109-142.

Piccolo D. (2000), Analisi statistica di un modello per le preferenze nel caso di tre alternative, *Quaderni di Statistica*, 2, 241-267.

Piccolo D. (2001), Some Approximations for the Asymptotic Variance of the

Maximum Likelihood Estimator of the Parameter in the Inverse Hypergeometric Random Variable, *Quaderni di Statistica*, 3, pp.215-229; corrected reprint in: *Quaderni di Statistica*, (2002), 4, pp.199-213.

Piccolo D., D'Elia A. (2007), A new approach for modelling consumers' preferences, *Food and Quality Preference*, doi:10.1016/j.foodqual.2007.07.002, in press.

Ridout M.S. (1999), Memory in coal tits: an alternative model, *Biometrics*, 55, 660-662.

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