Dipartimento di
Economia, Statistica e Diritto
Università di Pavia

Serie Statistica

n. 6/2011

Paola Cerchiello

On the distribution of functionals of discrete ordinal variables

QUADERNI DEL DIPARTIMENTO DI ECONOMIA, STATISTICA E DIRITTO
UNIVERSITÀ DI PAVIA

I **QUADERNI DEL DIPARTIMENTO DI ECONOMIA, STATISTICA E DIRITTO** hanno lo scopo di favorire la tempestiva divulgazione, in forma provvisoria o definitiva, di ricerche scientifiche originali. La pubblicazione di lavori nella collana è soggetta a referaggio e all'approvazione del Comitato Scientifico.
Questa nuova edizione dei **QUADERNI** rappresenta la continuazione di tre serie di pubblicazioni pre-esistenti: Quaderni del Dipartimento di Economia Pubblica e Territoriale, Quaderni di ricerca del Dipartimento di Statistica ed Economia Applicate "L. Lenti" e Osservatorio dei contratti della P.A.

On the distribution of functionals of discrete ordinal variables☆

Paola Cerchiello[1,]

Department of Economics, Statistics and Law, University of Pavia

---

1. Introduction

Arnold et al. (1992) and David and Nagaraja (2003) considered the fundamental theory of order statistics.
Khatri (1962) investigated the probability mass function and the cumulative distribution function of a single order statistic and the joint probability mass function and cumulative distribution function of any two or three order statistics from a discrete parent.
We focus our attention on the contributions given with regards to discrete distributions. What emerges is that results in literature are typically tailored on specific cases because of the hard mathematical tractability. As examples we can cite the paper on the negative binomial distribution (Young, 1970) and the paper on the geometric distribution (Srivastava, 1974; Ciardo et al., 1995). On the other hand, with regards to order statistics from continuous populations, we have over a thousand of contributions characterized by a wide range of different applications: for example survival analysis, robustness evaluation, filtering theory (Balakrishnan and Rao, 1998). Moreover if the first papers considered order statistics for independent and identically distributed random variables, the subsequent attention was payed on non identically distributed and dependant variables but always focused on continuous populations.
In this contribution we present the joint probability distribution of function of order statistics of particular interest such as the minimum and the maximum, the median and the first and third quartile. The appropriate combination of such location measures can be very useful in describing the characteristics of

Email address: `paola.cerchiello@unipv.it` (Paola Cerchiello)

a given discrete distribution when the ultimate goal is to offer a concise and intuitive statical summary.

2. Distribution of order statistics

In general the probability mass function of $i$order statistic over n observations from a discrete parent X can be obtained in three different ways: on one hand if we employ the cumulative distribution function $F_{i:n}(x)$ we have the binomial sum and the beta integral form approaches. On the other hand if we refer to the multinomial trials we have the multinomial argument. The choice typically depends on the aim: computational or theoretical. Below we report the probability mass function $f_{i:n}(x)$ of the i-order statistic in the beta integral form, since it will be employed in the subsequent analysis:

$$f_{i:n}(x) = C(i;n) \int_{F(x-)}^{F(x)} u^{i-1}(1-u)^{n-i} du$$

where

$$C(i;n) = \frac{n!}{(i-1)!(n-i)!}$$

With regards to the joint probability mass function of k order statistics $X_{i_1}(x), \ldots, X_{i_k}(x)$, given $1 \leq i_1 \leq i_2 \leq \ldots \leq i_k \leq n$, we have:

$$f_{i_1,i_2,\ldots,i_k:n}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) =$$
$$= C(i_1, i_2, \ldots, i_k; n) \times \int_S \left\{ \prod_{r=1}^{k} (u_{i_r} - u_{i_{r-1}})^{i_r - i_{r-1} - 1} \right\} (1-u_{i_k})^{n-i_k} du_{i_1} \ldots du_{i_k}$$

where

$$C(i_1, i_2, \ldots, i_k; n) = \frac{n!}{(n-i_k)! \prod_{r=1}^{k} (i_r - i_{r-1} - 1)!}$$

and the support S is

$$S=\{(u_{i_1},\ldots,u_{i_k}):u_{i_1}\leq u_{i_2}\leq\ldots\leq u_{i_k},F(x_r-)\leq u_r\leq F(x_r),r\in\{i_1,i_2,\ldots,i_k\}\}$$

Now we are interested in some linear combinations of order statistics from a discrete parent. One of the most important index, based on the minimum and maximum values of the sample, is the range, whose distribution is proposed in Arnold et al., 2003. In this case the algebraic calculations are shown to be straightforward giving to substantial simplification of expression (3). In fact the probability mass function of the range R_n is

$$\begin{aligned}P(R_n=r)=&\textstyle\sum_{x\in S}\{[F(x+r)-F(x-)]^n-[F(x+r)-F(x)]^n \\ &-[F(x+r-)-F(x-)]^n+[F(x+r-)-F(x)]^n\}\end{aligned} \tag{1}$$

As we will see in the following, when the index of interest is based on other location measures the calculations are more complex.

3. Distribution of Quantile Based Summaries

The main goal of a statistical analysis is to describe the characteristics of ordinal variables; appropriate statistical tools are therefore needed to summarize well such data. Thus it is very useful to employ simple summaries based on the appropriate combination of the most common quantiles of a distribution: minimum, first quartile, median, third quartile, maximum. We also recall that, when the available data have been collected by means of questionnaires, most of the variables to be analyzed is of qualitative nature, typically with a nominal or an ordinal scale. Such data requests for the employment of appropriate statistical tools able to exploit the intrinsic non quantitative nature.

In the sequel, first of all we show the formulation of the index IQ_n (aka inferior quartiles) based on the difference between the median and the minimum and later we report the relative probability mass function.

$$IQ_n=Med-Min$$

For sake of simplicity let us suppose to treat the case of an odd number of observations (the even case can be easily derived). Formally we have:

$$IQ_n = X_{\frac{n+1}{2};n} - X_{1;n}$$

Now, if we establish that $X_1 = x$ and $X_{\frac{n+1}{2}:n} = q + x$ we get:

$$P(IQ_n = q) = C(1, \frac{n+1}{2}; n) \sum_{x \in S} \int_{F(x-)}^{F(x)} \int_{F(x+q-)}^{F(x+q)} (u_{\frac{n+1}{2}} - u_1)^{\frac{n-3}{2}} (1 - u_{\frac{n+1}{2}})^{\frac{n-1}{2}} du_{\frac{n+1}{2}} du_1$$

After some calculations we get

$$P(IQ_n = q) = -\tfrac{n}{C} f(x)(1 - F(x+q))^{n-1} + \tfrac{n}{C} f(x)(1 - F(x+q-))^{n-1} + \textstyle\sum_{i=1}^{n} \frac{2}{n-1} \frac{2}{n+1} \times$$
$$\{(1 - F(x+q))^{\frac{n+i}{2}} [(F(x+q) - F(x))^{\frac{n-i}{2}} - (F(x+q) - F(x-))^{\frac{n-i}{2}}] -$$
$$(1 - F(x+q-))^{\frac{n+i}{2}} [(F(x+q-) - F(x))^{\frac{n-i}{2}} - (F(x+q-) - F(x-))^{\frac{n-i}{2}}]\}$$

where $C = C(1, \frac{n+1}{2}; n)$.
It is evident that the result is not as compact as for the Range index mainly because of the impossibility to simplify terms.
In order to facilitate the interpretation of formulae (10) we report the simplest example with $n = 3$.
After some simple simplifications the final result is

$$P(IQ_3 = q) = -\frac{1}{2} f(x)(1 - F(x+q))^2 + \frac{1}{2} f(x)(1 - F(x+q-))^2$$

The analogous of the IQ index would be the SQ (aka superior quartile) that is evidently based on the difference between the maximum value and the median, that is SQ = Max -Med.

$$SQ_n = X_{(n;n)} - X_{(\frac{n+1}{2};n)}$$

Now, if we establish that $X_{\frac{n+1}{2}:n} = x$ and $X_{n:n} = q + x$ we get:

$$P(SQ_n = q) = C(\frac{n+1}{2}, n; n) \sum_{x \in S} \int_{F(x-)}^{F(x)} \int_{F(x+q-)}^{F(x+q)} (u_{\frac{n+1}{2}})^{\frac{n-1}{2}} (u_n - u_{\frac{n+1}{2}})^{\frac{n-3}{2}} du_n du_{\frac{n+1}{2}}$$

After some calculations we get

$$P(SQ_n = q) = -\frac{n}{C} f(x+q) F(x)^{n-1} + \frac{n}{C} f(x+q) F(x-)^{n-1} + \sum_{i=1}^{n} \frac{2}{n-1} \frac{2}{n+1} \times \\ \{(F(x))^{\frac{n+i}{2}} [(F(x+q) - F(x))^{\frac{n-i}{2}} - (F(x+q) - F(x-))^{\frac{n-i}{2}}] - \\ (F(x-))^{\frac{n+i}{2}} [(F(x+q-) - F(x))^{\frac{n-i}{2}} - (F(x+q-) - F(x-))^{\frac{n-i}{2}}]\}$$

where $C = C(\frac{n+1}{2}, n; n)$.
Similarly to the IQ index, if we consider the simplest case with $n = 3$, we get:

$$P(SQ_3 = q) = -\frac{1}{2} f(x+q) F(x)^2 + \frac{1}{2} f(x+q) F(x-)^2$$

A further function of particular interest is the midrange summary that is defined as $Mid = \frac{X_{(1;n)+X_(n;n)}}{2}$.
Now, if we establish that $X_{(1;n)} = x$ and $X_{n:n} = 2Mid - x$ we get:

$$P(Mid_n = k) = 2 \\ \sum_{Min \in S} [F(2Mid - x) - F(x-)]^n - [F(2Mid - x) - F(x)]^n \\ + [F(2Mid - x-) - F(x)]^n - [F(2Mid - x-) - F(x-)]^n$$

The reader can note that the above expression is similar to the distribution of the range (formula 1)

We finally report the analytical distribution of another interesting summary, the interquartile difference ID defined as $ID = X_{([\frac{3n}{4}]+1;n)} - X_{([\frac{n}{4}]+1;n)}$. In this last case the calculations are more complicated since there is not the possibility to simplify any term:

$$
\begin{aligned}
&P(ID_n = k){=}[-A^n + B^n + G^n - D^n] \\
&+\left(\frac{n!}{[3n/4]-[n/4]+1}\right) E^{(n-[3n/4]-1)}[-A^{[3n/4]-[n/4]+1} * F(Q_1)^{[n/4]} + B^{[3n/4]-[n/4]+1} * F(Q_1^-)^{[n/4]}] \\
&+\left(\frac{n!}{[3n/4]-[n/4]+1}\right) H^{(n-[3n/4]-1)}[+G^{[3n/4]-[n/4]+1} * F(Q_1)^{[n/4]} - D^{[3n/4]-[n/4]+1} * F(Q_1^-)^{[n/4]}] \\
&+n\{A^{n-1}[E - F(Q_1)] + B^{n-1}[-E + F(Q_1^-)] + G^{n-1}[-H + F(Q_1)] + D^{n-1}[H - F(Q_1^-)]\} \\
&+\sum_{i=[3n/4]-[n/4]+2}^{l=n-2} A^i \frac{\varphi^j}{\varphi n}[F(Q_1)^{[n/4]} E^{n-[3n/4]-1}] \\
&+\sum_{i=[3n/4]-[n/4]+2}^{l=n-2} B^i \frac{\varphi^j}{\varphi n}[F(Q_1^-)^{[n/4]} H^{n-[3n/4]-1}] \\
&+\sum_{i=[3n/4]-[n/4]+2}^{l=n-2} G^i \frac{\varphi^j}{\varphi n}[F(Q_1)^{[n/4]} E^{n-[3n/4]-1}] \\
&+\sum_{i=[3n/4]-[n/4]+2}^{l=n-2} D^i \frac{\varphi^j}{\varphi n}[F(Q_1^-)^{[n/4]} H^{n-[3n/4]-1}]
\end{aligned}
$$

where

$$
\begin{aligned}
&A{=}F(Q_3) - F(Q_1) \quad B{=}F(Q_3) - F(Q_1^-) \\
&G{=}F(Q_3^-) - F(Q_1) \quad D{=}F(Q_3^-) - F(Q_1^-) \\
&E{=}1 - F(Q_3) \quad H{=}1 - F(Q_3^-)
\end{aligned}
$$

and $i \leq l$, $j = i - [3n/4] + [n/4] - 1$

Once again, for the sake of interpretation, we show the final result for $n = 3$:

$$P(ID_3 = k){=}\textstyle\sum_{Min \in S}[-A^3 + B^3 + C^3 - D^3]$$

4. Conclusions

In this paper we treat the topic of the distribution of functionals of discrete ordinal variables. In particular we focus on the probability mass function of four indexes that can be profitably used when the goal is to perform a descriptive analysis. In the literature the only distribution for which the analytic form is known is the range. In the paper we have obtained the distribution of a collection of alternative indexes based on appropriate linear combinations of order statistics such as: the minimum, the first quartile, the median, the third quartile and the maximum.

Starting from the general definition of the joint probability mass function for order statistics from a discrete parent (Arnold et al. 2002), we have computed the analytical formulae of some linear combinations of order statistics to be used in practical contexts. The obtained formulae are evidently not so simple as in the range case, especially for the index based on the first and third quartiles. However in this paper we have shown that it is possible in each case to obtain a recursive formula that can be profitably employed and implemented with the available software.

References

[1] Arnold, B.C, Balakrishnan, N., Nagaraja, H.N., 2003. A first course in order statistics. John Wiley & Sons.

[7] Balakrishnan, N., Rao, C.R., 1998. *Order Statistics:Applications.* Elsevier, Amsterdam, The Netherlands.

[7] Ciardo, G., Leemis, L., Nicol, D., 1995. On the minimum of independent geometrically distributed random variables. Statist. Probab. Lett. 23, 313-326.

[4] David, H.A., Nagaraja, H.N., 2003. *Order statistics*, third ed. John Wiley & Sons.

[5] Khatri, C.G., 1962. Distributions of order statistics for discrete case, Annals of the Ins. of Stat. Math. 14, 167-171.

[7] Srivastava, R.C., 1974. Two characterizations of the geometric distribution. J. Amer. Statist. Assoc. 69, 267-269.

[7] Young, D.H., 1970. The order statistics of the negative binomila distribution, Biometrika 57, 181-186.

QUADERNI DEL DIPARTIMENTO DI ECONOMIA, STATISTICA E DIRITTO

n. 1/2011 Paolo Giudici, Emanuela Raffinetti, *A Gini concentration quality measure for ordinal variables,* Serie Statistica.
n. 2/2011 Silvia Figini, Lijun Gao, Paolo Giudici, *Bayesian efficient capital at risk estimation,* Serie Statistica.
n. 3/2011 Luigi Bernardi, *Tendenze dei prelievi tributari ed effetti fiscali della crisi finanziaria nell'Unione europea e in Svizzera,* Serie Economia.
n. 4/2011 Silvio Beretta, Renata Targetti Lenti, *India in the Outsourcing/Offshoring Process: A Western Perspective,* Serie Economia.
n. 5/2011 Paolo Giudici, Eleonora Lorenzini, *SMEs, e-commerce and territorial development: the experience of a "web district",* Serie Statistica.
n. 6/2011 Paola Cerchiello, *On the distribution of functionals of discrete ordinal variables,* Serie Statistica.

COLLANE PRECEDENTI

QUADERNI DEL DIPARTIMENTO DI ECONOMIA PUBBLICA E TERRITORIALE

n. 1/2010 Silvio Beretta, *Variabili finanziarie ed economia globale in tempo di crisi*
n. 2/2010 Silvio Beretta, Renata Targetti Lenti, *L'India nel processo di integrazione internazionale. Dal primo al secondo* unbundling *e la posizione dell'Italia*
n. 3/2010 Margarita Olivera, *Challenges to Regional Integration in Latin America*
n. 4/2010 Italo Magnani, *Un economista liberale guarda alla economia dell'ambiente: impressioni e riflessioni*
n. 5/2010 Italo Magnani, *A cinquant'anni dalla scomparsa di Benvenuto Griziotti: Riflessioni*
n. 6/2010 Luca Mantovan**,** *Class-bias in Technology Adoption: Stagnation and Transformation of Subsistence Agriculture in the Ethiopian Northern Highlands*
n. 7/2010 Marco Missaglia, Giovanni Valensisi, *A trade-focused, post-Keynesian CGE model for Palestine*
n. 8/2010 Giovanni Valensisi, Marco Missaglia, *Reappraising the World Bank CGE model on Palestine: macroeconomic and financial issues*
n. 1/2009 Giorgio Panella, Andrea Zatti, Fiorenza Carraro, *Market Based Instruments for Energy Sustainability*
n. 1/2008 Italo Magnani, *Il pubblico e il privato nella economia della città*
n. 2/2008 Italo Magnani, *Note a margine di una recente opera sull'indirizzo sociologico della scienza delle finanze italiana*
n. 3/2008 Italo Magnani, *La riforma sociale nella formazione di Nitti economista*
n. 4/2008 Marisa Bottiroli Civardi, Renata Targetti Lenti and Rosaria Vega Pansini, *Multiplier Decomposition, Poverty and Inequality in Income Distribution in a SAM Framework: The Vietnamese Case*
n. 5/2008 Luca Mantovan, *A Study on Rural Subsistence in the Ethiopian Northern Highlands*

[per i Quaderni precedenti si rinvia a http://www-5.unipv.it/webdesed/ept/quaderni.php]

QUADERNI DI RICERCA DEL DIPARTIMENTO DI STATISTICA ED ECONOMIA APPLICATE "L. LENTI"

Carla Ge Rondi, *L'après mariage en Italie au début du XXIe siècle* (2005, n. 27)
Carla Ge Rondi, *Casalinga: popolazione attiva senza retribuzione* (2005, n. 25)
Bruno Scarpa, David Dunson, *Bayesian Methods for Searching for Optimal Rules for Timing Intercourse to Achieve Pregnancy* (2005, n. 24)

Bruno Scarpa, *Lo stress in azienda. Modelli di analisi di un'indagine per l'identificazione delle cause di stress* (2004, n. 23)
Bruno Scarpa, *La Customer Satisfaction per un'azienda di servizi informatici. Impostazione e analisi di un'indagine via web* (2004, n. 22)

[per i Quaderni precedenti si rinvia a http://www-5.unipv.it/webdesed/lenti/quaderni.php]

OSSERVATORIO DEI CONTRATTI DELLA P.A.

[Si rinvia a http://www.contratti-appalti.it/]