IFS

Institute for Fiscal Studies

Gordon Anderson

23/41

Working paper

A coefficient of variation for ordered categorical data: Analyzing relative health inequality and ageing in the UK and relative human resource inequality and gender in Canada

Economic and Social Research Council

A Coefficient of Variation for Ordered Categorical Data: Analyzing Relative Health Inequality and ageing in the UK and Relative Human Resource Inequality and Gender in Canada.

Gordon Anderson*

Economics Department, University of Toronto, Canada.

November 2023

Abstract.

The burgeoning use of ordinal data throughout the Empirical Sciences calls for location and variation measurement instruments suitable for such data environments. Neither Pearson's Coefficient of Variation nor the Sharpe Ratio, relative variation comparison workhorses in cardinal worlds, are applicable in ordinal paradigms without artificial data scaling, a practice recently much criticized for its inherent ambiguity. Here, employing the concept of probabilistic distance, unequivocal, scale independent, Coefficient of Variation analogues for use in Multivariate Ordered Categorical environments are introduced and exemplified in analyses of Self-Reported Health outcomes in the UK and Human Resource determinants in Canada.

Keywords: Ordinal Data, Coefficient of Variation, Sharpe Ratio, Probabilistic Distance

JEL classification C18, I32, G10

*The author is grateful for the most helpful comments and suggestions on elements of this paper from participants at the Inequality in the 21st Century Symposium at Queen Mary University of London, the ECINEQ general conference in Aix en Provence and the 5th World Ageing and Rejuvenation Conference in Frankfurt in 2023. Many thanks to the IFS for providing the invaluable research assistance of Ignasi Merediz-Solà. Data are available from the author on request.

Authors Contact Information: Department of Economics, University of Toronto, Max Gluskin House, 150 St. George Street, Toronto, Ontario, M5S 3G7 Canada. Email: gordonj.anderson@utoronto.ca

Introduction.

The number of active pollsters in the United States that collect and analyze ordered categorical data has more than doubled since the turn of the century (Pew Research Centre 2023). Within the Health and Social Sciences, randomized controlled trials that generate ordinal outcomes have seen increasing use (Selman et. al. 2023) to the point where the World Health Organization developed an ordinal scale describing the disease severity of COVID-19 that has been adapted in various treatment trials (Marshall et. Al. 2020). All such data has been generated with the generic intent of comparing relative outcome levels of, and variations within, diverse subgroups which requires a means of measuring and comparing the location and within group variation in ordinal data environments.

In cardinally measured paradigms, means and standard deviations, and their combination in the form of the Coefficient of Variation (COV) (Pearson 1896) and its inverse, the Sharpe Ratio (COV^{-1}) (Sharpe 1964,1994), have been used extensively to answer generic level and variation questions, but cardinality is of the essence in that pursuit. In ordered categorical worlds, some form of artificial scaling has typically been applied to categories to facilitate analysis (see for example, Likert 1932, Rankin 1957 and Cantril 1965) but that presents problems of scale dependency and concomitant equivocation since different, though equally valid, scales frequently yield substantively different conclusions (Bond and Lang 2019, Schroder and Yitzhaki 2017) raising questions concerning the viability of variation measurement about a location measure in ordered categorical data situations.

Without artificial scaling, COV and COV^{-1} have no natural analogues in ordered categorical data environments since that paradigm is bereft of cardinal measure. In its absence, inequality and polarization researchers have used notions of probabilistic distance (Mendelson 1987) and the construct of a median preserving spread in order to quantify variation for the purpose of measuring inequality and polarization (Blair and Lacy 2000, Allison and Foster 2004, Kobus and Kurek 2019). The probabilistic distance of a given category from the median focus category is measured in terms of the likelihood of an outcome occurring in the given or any other category between it and the median category, the higher that probability is, the further apart are the categories deemed to be. Inequality is then quantified as the average probabilistic distance from the median focus category of all non-median categories.

Here, taking a slightly different view of inequality measurement in ordered categorical environments, the notion of probabilistic distance is employed to develop analogs of the Coefficient of Variation (and by implication the Sharp Ratio) for use in multivariate ordered categorical data environments. In the following, details of the conventional Coefficient of Variation and its multivariate versions are outlined in Section 1, Section 2 proposes an analogue for multivariate ordered categorical data environments and Section 3 provides two exemplifying applications. The first, a univariate analysis of the progress of self-reported Health Status in the United Kingdom from 2010 to 2018 (just prior to the Covid outbreak) explores health inequities and the aging process and reveals increasing relative inequalities. The second, a multivariate analysis of the progress of experience and embodied human capital factors underlying of Canadian Human Resource Stocks across gender and time (the 2006-2016 decade) reveals some gender divergence in Human Resource acquisition. Conclusions are drawn in Section 4.

1.The Coefficient of Variation.

First introduced by Pearson (1896) as the ratio of the standard deviation to the mean, COV is a unit free relative variation measure. It, and its inverse, the well-known Sharpe (1964,1994) Ratio used for examining risk adjusted Excess Returns[1], have been used extensively in economics and finance as a measure of economic inequality and relative risk. Despite its disadvantages (Kvalseth 2017), it has also seen extensive use in the physical and biological sciences (Weber et. al 2004), engineering (Jalalibal et.al. 2021) and Industrial Organization fields (Bedeian & Mossholder 2000) where cardinally measurable data abounds. Pearson proposed COV in response to Galtons' practice of using the 13 to 12 male-female size ratio[2] in his work on Natural Inheritance (Galton 1894) and used the mean focused variation measure standardized by the mean to address the comparative variation of organ sizes (usually skull and bone dimensions) across race and gender[3]. His rationale for standardizing the dispersion measure by the mean was a concern for reliability and consistency across disparate distributions, that variation measurement should not be too variable or at least consistently variable i.e. sufficiently stable about the mean value, so as to be comparably useful across races and genders. In more recent times COV has seen a variety of extensions to multivariate

[1] Following concerns that the standard deviation was not an adequate reflection of downside risk, the Sortino Ratio (Sortino and Price 1994) modified the Sharpe Ratio by considering only non-positive deviations from the mean in the standard deviation calculation. Similar modifications are possible in the ordered categorical paradigm.
[2] Galton would rescale a female organ size by 13/12 to obtain a comparable male equivalent.
[3] Pearson's view of his new statistic (Pearson 1896 pp. 276-9) was circumspect but enthusiastic, he wrote:
"*Of course, it does not follow because we have defined in this manner our "coefficient of variation", that this is really a significant quantity in the comparison of various races; it may be only a convenient mathematical expression, but I believe there is evidence to show that it is a more reliable test of "efficiency" in a race than absolute variation*."

environments (see Albert and Zhang 2010 for a survey) based on alternative approaches to dealing with the multiplicity of measurement units.

For expository convenience consider a continuous cardinally measurable variable x with $0 < x < Y < \infty$ and denote group t's PDF: $f_t(x)$ with a corresponding CDF: $F_t(x) = P_t(X < x) = \int_0^x f_t(z)dz$, Survival Function SF: $S_t(x) = P_t(X \geq x) = 1 - F_t(x)$, group t mean: $\mu_t = E_{f_t(x)}(x) = \int_0^Y x f_t(x)dx$ and group t variance of x: $\sigma_t^2 = E_{f_t(x)}((x-\mu_t)^2) = \int_0^Y (x - \mu_t)^2 f_t(x)dx$. Note that letting $u(x) = x$ and $v'(x) = f(x)$, the integration by parts rule[4] reveals the mean to be the integral of the survival function since $\mu = \int_0^Y x f(x)dx = [xF(x)]_0^Y - \int_0^Y F(x)dx = Y - \int_0^Y F(x)dx = \int_0^Y (1 - F(x))dx = \int_0^Y S(x)dx$ which yields an alternative interpretation of the mean as the cumulation of chances of higher outcomes than x over its range. Then $COV_{f_t(x)}(x)$, the group t Coefficient of Variation may be written as:

$$COV_{f_t(x)}(x) = \frac{\sqrt{\sigma_t^2}}{\mu_t} = \frac{\sqrt{\sigma_t^2}}{\int_0^Y S_t(x)dx} \qquad [1]$$

Practically, for a collection of N randomly sampled cardinally measurable values $x_i, i = 1,..N$, where for convenience $x_i \geq 0$, the basic COV is given by:

$$COV = \frac{\sqrt{\sum_{i=1}^N (x_i - \underline{x})^2 / (N-1)}}{\underline{x}} = \frac{\hat{\sigma}}{\hat{\mu}}; where\ \hat{\mu} = \underline{x} = \sum_{i=1}^N x_i / N \qquad [2]$$

When data are sampled from a set of K discrete cardinally measurable values x_k $k = 1,..,K$ where p_k is the proportion of the sample that took on the value x_k, [1] can be computed as:

---

[4] $\int_0^Y u(x)v'(x)dx = [u(x)v(x)]_0^Y - \int_0^Y u'(x)v(x)dx$

$$COV = \frac{\sqrt{\sum_{k=1}^{K}(x_k - \underline{x})^2 p_k}}{\underline{x}} = \frac{\hat{\sigma}}{\hat{\mu}}; where\ \hat{\mu} = \underline{x} = \sum_{k=1}^{K} x_k\, p_k \qquad [2a]$$

As a measure of relative variation, it can be seen to be the square root of the variance estimate, which is the average of the squared distances of the x_i's from the mean, divided by the mean. The mean is very much the focus of the statistic, it is the value that minimises the magnitude of the variance estimate (a similar variation measure around any other value would always be at least as large) and division by it dilutes the standard deviation value and renders the statistic a unit free measure[5]. Its inverse, the Sharpe Ratio (Sharpe 1994), is employed in the finance field in risk vs. return scenarios, where the x_i's are rates of excess return and $\hat{\sigma}$ is a measure of their riskiness level, COV^{-1} can be seen to dilute the average excess return level by the level of riskiness. In the context of income inequality and wellbeing measurement[6], the Sharpe measure dilutes the average income level by a measure of the inequality with which incomes are distributed thus providing an inequality modulated income measure.

Multivariate Extensions of the Coefficient of Variation.

Extending the coefficient of variation to the multivariate paradigm has seen several alternative formulations of a Multivariate COV proposed in the literature. Albert and Zhang (2010) reviewed some of these and proposed a novel formulation themselves,

---

[5] Similar statistics can be contrived if other foci are of interest by making $\underline{x}$ the median or modal value of the collection, indeed dividing the standard deviation by any quantile value would render it a unit free measure relative to the designated quantile.

[6] COV can be shown to satisfy the inequality measurement axioms of anonymity, scale invariance and population independence (Champernowne and Cowell 1999).

all amount to standardizing a function of the variance covariance matrix with the inner product of the dimension means and taking the square root thereof. The object being to reconcile the diverse units of measurement in the various dimensions to obtain a unit free measure. In a $Q > 1$ dimension setting, letting $\underline{\mu}$ be the Q x 1 vector of dimension means and C be the Q x Q covariance matrix, the alternatives (see inter alia Reyment 1960, Van Valen 1974, Voinov and Nikulin 1966 and Albert and Zhang 2010) considered by Albert and Zhang were:

$$\sqrt{\frac{det(C)^{1/K}}{\underline{\mu}'\,\underline{\mu}}};\quad \sqrt{\frac{trC}{\underline{\mu}'\,\underline{\mu}}};\quad \sqrt{\frac{1}{\underline{\mu}'C^{-1}\,\underline{\mu}}};\quad \sqrt{\frac{\underline{\mu}'\,C\underline{\mu}}{\left(\underline{\mu}'\,\underline{\mu}\right)^2}}.$$

It is worthy of note that when $Q = 1$ all of these formulae reduce to the conventional coefficient of variation yet in multivariate empirical settings they can yield very different values for a given sample (Aerts, Haesbroeck and Ruwet 2015).

2. A Coefficient of Variation Analogue for Multivariate Ordered Categorical Data.

To develop Coefficient of Variation or Sharpe-Ratio analogues for Ordered Categorical data, a means of measuring location and variation in the absence of cardinal measure is required[7]. The notion of probabilistic distance, the sense that two ordered outcomes are further apart the greater is the probability of an outcome between them occurring, is useful in this case.

---

[7] Development of the asymptotic distributions of the various constructs used in this analysis is confined to the appendix.

To fix ideas, suppose $K \geq 3$ ordered categories indexed $k = 1,..,K$ with higher k implying higher category. Endow the categories with a Probability Density Function f described by the probabilities $p_{fk}, k = 1,..,K$ of being in the k'th category under distribution f, where $p_{fk} \geq 0$ and $\sum_{k=1}^{K} p_{fk} = 1$. For $k = 1,..,K$, the Cumulative Distribution Function F is given by $F_k = \sum_{i=1}^{k} p_{fi}$ and the Survival Function S is given by $S_k = 1 - F_k$. Analogous to the continuous paradigm formulation of the mean as the integral of the survival function over the range of x, the sum of the Survival Function values over all categories could be considered as a "Mean Ordered Categorical" or MOC location measure where:

$$MOC = \sum_{k=1}^{K} S_k$$

Note that, with a potential minimum value of 0 (when all probability mass is in the lowest category) and a maximum potential value of $K - 1$ (when all probability mass is in the highest category), MOC is not independent of K, the number of categories. While this is of no consequence when group outcomes are being compared across a common number of categories, it does matter when different groups have different numbers of categories. This can be resolved by dividing MOC by $K - 1$ rendering it a number on the unit interval.

Note that in multidimensional ordered categorical data problems, since probabilistic distance is the unit of measurement across all dimensions, a "Mean Ordered Categorical" or MOC location measure is easily developed. For example, consider a two-dimension situation where the second dimension has J categories and the joint probability density function f described by the probabilities $p_{fk,j}, k = 1,..,K, j = 1,..,J$

where $p_{fk,j} \geq 0$ and $\sum_{j=1}^{J}\sum_{k=1}^{K} p_{fk,j} = 1$ yields a Cumulative Distribution Function $F_{fk,j} = \sum_{h=1}^{j}\sum_{i=1}^{k} p_{fi,h}$ and Survival Function $S_{k,j} = 1 - F_{fk,j}$ for $k = 1,..,K\ j = 1,..,J$, then:

$$MOC = \sum_{j=1}^{J}\sum_{k=1}^{K} S_{k,j}$$

In two dimensions MOC will have a minimum potential value of 0 and a maximum potential value of $KJ - 1$ and dividing MOC by this value will yield a location index that is comparable across two dimensioned outcomes with differing numbers of categories. An alternative Median Location measure is problematic in many dimensions since the median is a contour of points.

Developing a measure of variation.

For a given outcome $k^* \in 1,..,K$ and outcomes $k = k^* + 1,..,K$, define the Upper Cumulants of f with respect to k^* as $F_k^{U,k^*} = \sum_{i=k^*+1}^{k} p_{fi}$ (note for $k \leq k^*$, $F_k^{U,k^*} = 0$) and, for outcomes $k = 1,..,k^* - 1$, define its Lower Cumulants as $F_k^{L,k^*} = \sum_{i=k}^{k^*-1} p_{fi}$ (note for $k \geq k^*$, $F_k^{L,k^*} = 0$). It may be seen that $\frac{F_k^{L,k^*}}{F_1^{L,k^*}}$ $k = 1,..,k^* - 1$ is in effect the SF of the below k^* conditional PDF, whereas $\frac{F_k^{U,k^*}}{F_K^{U,k^*}}$ $k = k^* + 1,..,K$ is the CDF of the above k^* conditional PDF. When $k > k^*$, F_k^{U,k^*} is the probability of an outcome between k^* and $k + 1$ occurring which is monotonically non decreasing in k, when $k < k^*$, F_k^{L,k^*} is the probability of an outcome between k^* and $k - 1$ occurring which is monotonically non-decreasing in $k^* - k$. Each record a sense of probabilistic distance of k from k^* in terms of the chance that an outcome will emerge between k and k^* which increases with $|k^* - k|$. Similarly defining G_k^{U,k^*}, G_k^{L,k^*}, the Upper and

Lower Cumulants of g about k^*, then g constitutes an increasing spread of f with respect to outcome k^* when:

$$G_k^{L,k^*} \geq F_k^{L,k^*} \forall\, k = 1,.,k^* - 1 \; and \; G_k^{U,k^*} \geq F_k^{U,k^*} \forall\, k = k^* + 1,.,K \; with > somewhere. \quad [3]$$

The Mendelson (1987) condition [3] amounts to a first order stochastic dominance condition on the "downward looking" below k^* conditional distributions (i.e. imagine the category orderings below k^* were reversed) and the "upward looking" above k^* conditional distributions where f dominates g in each context. Intuitively, with respect to k^* inequality in g distribution is greater than inequality in f distribution with respect to k^* when the chance of below k^* outcomes and the chance of above k^* outcomes are both at least as great in g as they are in f with strictly greater than in at least one case[8].

Given the absence of cardinal measure, setting k^* as the "median" category and using this notion of probabilistic distance has been the basis of inequality and bi-polarization measurement in univariate ordered categorical paradigms (Blair and Lacy 2000, Allison and Foster 2004, Kobus 2015). However, if inequality is conceptualized as the antithesis of equality or complete commonality in the population, the aggregate distance of subjects from a potential focus point of complete commonality would characterize it. In this context, when compared to the mode which is the most likely point of complete commonality, the Median or the Mean may not be very good focus points. They often have lower probabilistic density than the mode (for example in heavily skewed or strongly segmented bimodal distributions), rendering them less likely candidates as point of complete

[8] This construct is similar to notions of left and right distributional separation developed in Anderson (2004) and if a Sortino and Price (1994) type analysis was desired only left separation (i.e. $G_k^{L,k^*} \geq F_k^{L,k^*} \forall\, k = 1,.,k^* - 1$) should be considered.

equality. Furthermore, in multivariate settings mean and median focal points are difficult to determine whereas the modal point is usually uniquely determined (even in a multiplicity of nodes there is usually one node with a density greater than the rest).

Noting that, in a likelihood sense, the mode is the most likely point of complete equality, Anderson and Yalonetzki (2023) provide an alternative to the Median Preserving Spread formulation in the Modally Preserving Spread with the mode as a focal point. As an inequality measure it has a natural likelihood-based interpretation (average probabilistic distance from the most likely point of commonality), is well defined in multidimensional situations, and has a probabilistic unit of measurement which is common to all dimensions.

The Modal Preserving Spread.

Define the Modal outcome of distribution f as outcome k^* $such\ that\ p_{fk^*} = \max_k p_{fk}$. Determining k^* by seeking that category for which $\hat{p}_{fk^*} = \max_k \hat{p}_{fk}$ where $\hat{p}_{fk}, k = 1,.., K$ are the maximum likelihood estimates of category densities, renders k^* as the maximum likelihood estimate of the category most likely to command unanimity of membership. Since the smallest possible value of p_{fk^*} is $\frac{1}{K} + \varepsilon$ where ε has arbitrarily small positive value, $\frac{1}{K} < p_{fk^*} \leq 1$ and, when p_{fk^*} is viewed as the chance that the whole population resides in outcome k^*, $LC(f) = (Kp_{fk^*} - 1)/(K - 1)$ is a very natural likelihood based measure or index on the unit interval of the extent of commonality or equality of outcome in the distribution at the modal outcome. When $LC(f) \to 0$ there is little chance of equality of outcome, when $LC(f) \to 1$ there is every chance of equality of outcome. It follows that its complement, $II(f) = 1 - LC(f) = K\left(1 - p_{fk^*}\right)/(K - 1)$ is an intuitive likelihood-based measure of the extent of inequality of outcome[9]. Unfortunately, it is not responsive to variation in spread in the rest of the distribution in the sense that a marginal shift in mass from k' to k'' where k' , $k'' \neq k^*$ would leave it unaltered unless the shift rendered k'' the new modal outcome. To capture this, the concept of a modal preserving spread needs to be considered. Basically g constitutes a Modal Preserving Spread of f if [3] holds and k^*remains the modal outcome of g i.e. $p_{gk^*} = \max_k p_{gk}$.

This can be readily checked by considering $UAMBI(f,g) = \frac{\sum_{k=1}^{K}((G_k^{U,k^*} - F_k^{U,k^*}) + (G_k^{L,k^*} - F_k^{L,k^*}))}{\sum_{k=1}^{K}\left(\left|(G_k^{U,k^*} - F_k^{U,k^*})\right| + \left|(G_k^{L,k^*} - F_k^{L,k^*})\right|\right)}$, when $UAMBI(f,g)$ = 1, distribution g constitutes an unambiguous Modal Preserving Spread of distribution f. Furthermore, given dispersion from the focus point k^* is maximized when k^*/K mass is allocated to the lowest outcome and $\frac{(K-k^*)}{K}$ is allocated to the highest outcome:

$$0 \leq IMPS(g,f) = \frac{\sum_{k=1}^{K}((G_k^{U,k^*} - F_k^{U,k^*}) + (G_k^{L,k^*} - F_k^{L,k^*}))}{\left(\frac{k^*\sum_{i=1}^{k^*-1} i}{K} + \frac{(K-k^*)\sum_{i=k^*+1}^{K}(i-k^*)}{K} - \sum_{k=1}^{K}\left(F_k^{U,k^*} + F_k^{L,k^*}\right)\right)} \leq 1$$

provides an index measure on the unit interval of the extent of increased Modally Focused relative spread or inequality associated with a move from f to g.

A Modally Focused Inequality Index for Ordered Categorical Data.

Suppose f^e was the distribution of a completely equal group with all elements experiencing outcome k^*, then $p_{fk^*} = 1$ and $p_{fk} = 0 \; \forall \; k \neq k^*$ so that $F_k^{U,k^*} = 0$ and $F_k^{L,k^*} = 0 \; \forall \; k$, then $IMPS(g,f^e)$ becomes:

$$IMPS(g,f^e) = \frac{\sum_{k=1}^{K}(G_k^{U,k^*} + G_k^{L,k^*})}{\left(\frac{(k^*-1)\sum_{i=1}^{k^*-1} i}{K} + \frac{(K-k^*)\sum_{i=k^*+1}^{K}(i-k^*)}{K}\right)} = MFI(g) \qquad [4]$$

[9] Indeed, in unordered categorical worlds IC and II provide equally useful indices of commonality and inequality.

[4] corresponds to a measure of the extent of inequality inherent in the ordered categorical distribution g relative to a state of complete equality at the category most likely to command unanimous membership and thus provides a measure, $MFI(g)$, of the Modally Focused Inequality inherent in distribution g. Let the k^* Focussed Probabilistic Distance vector $\underline{G}^{PD,k^*}$, recording the chance of being in the collection of categories successively further distanced from k^*, be given by:

$$\underline{G}^{PD,k^*} = \begin{bmatrix} G_1^L \\ . \\ G_{k^*-1}^L \\ 0 \\ G_{k^*+1}^U \\ . \\ G_{K-k^*}^U \end{bmatrix}$$

Note that the Probabilistic Distance function is an increasing function of the categorical distance from the i^* category which does not depend upon arbitrary attribution of value to a category in the form of a scale. Letting $\varphi(K,k^*) = \left(\frac{k^* \sum_{i=1}^{k^*-1} i}{K} + \frac{(K-k^*) \sum_{i=,k^*+1}^{K} (i-k^*)}{K}\right)$ then, given a K dimensioned unit vector d, $MFI(g)$ may be written as[10]:

$$0 \le MFI(g,k^*) = \frac{1}{\varphi(K,k^*)} d' \underline{G}^{PD,k^*} \le 1 \qquad [5]$$

A multivariate version.

In a multidimensional ordered categorical context, one of the attractions of the probabilistic distance approach is that, unlike the corresponding cardinal environment, the unit of measure is a probability number that is common to all dimensions. For simplicity, consider the bivariate categorical case where both dimensions are ordered with $p_{f,i,j} \ge 0$: $i = 1,..,I, j =$

---

[10] Note that when comparing groups with common K, k^*, the term $1/\varphi(K,k^*)$ can be omitted.

$1,..,J\ \sum_{i=1}^{I}\sum_{j=1}^{J} p_{f,i,j} = 1$ with the ordering again following the dimension indexing, cumulative and counter cumulative density functions are well defined with $F_{i,j} = \sum_{k=1}^{i}\sum_{l=1}^{j} p_{f,k,l}\ for\ i = 1,..,I, j = 1,..,J.$

In the modal case where k^* coordinates are $\{i^*, j^*\}$ so that $\max_{i,j} p_{f,i,j} = p_{f,i^*,j^*}$:

$$\text{Let } p_{f,i^*,j}^{**} = p_{f,i^*,j}\ j = 1,..,J\ and\ \ p_{f,i,j^*}^{**} = p_{f,i,j^*}\ i = 1,..,I$$

$$F_{i,j}^{**} = F_{i+1,j}^{**} + p_{f,i,j}\ \forall\ i < i^*\ and\ F_{i,j}^{**} = F_{i,j}^{**} + p_{f,i,j}\ \forall\ i > i^*, \forall j = 1,..,J$$

$$F_{i,j}^{\mathrm{L}k^*} = F_{i,j+1}^{\mathrm{L}k^*} + F_{i,j}^{**}\ \forall\ j < j^*\ and\ F_{i,j+1}^{\mathrm{U}k^*} = F_{i,j}^{\mathrm{U}k^*} + F_{i,j}^{**}\ \forall\ j > j^*, i = 1,..,I$$

Again, when p_{f,i^*,j^*} is viewed as the likelihood that the whole population resides in outcome $\{i^*, j^*\}$, $IC(f) = \left(IJp_{f,i^*,j^*} - 1\right)/(IJ - 1)$ is a very natural measure or index on the unit interval of the commonality or equality of outcome in the distribution, so that its complement, $II(f) = IJ\left(1 - p_{f,i^*,j^*}\right)/(IJ - 1)$ provides an intuitive likelihood based measure of inequality of outcome and it is an equally useful index of such in unordered categorical paradigms.

The corresponding 2-dimensional version of [4] is given by:

$$MFI(g) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(G_{i,j}^{U,k^*} + G_{i,j}^{L,k^*}\right)}{\left(\frac{j^*i^*\sum_{i=1}^{i^*-1}\sum_{j=1}^{j^*} ij}{IJ} + \frac{(IJ - j^*i^*)\sum_{i=i^*+1}^{I}\sum_{j=j^*+1}^{J}(ij - i^*j^*)}{IJ}\right)}$$

Appropriately vectorized versions of the I x J matrices $G_{.,.}^{U,k^*}$ and $G_{.,.}^{L,k^*}$ and their corresponding IJ square cumulation matrix $C_{i^*j^*}$ can be constructed to form the i^*, j^* Focused Probabilistic Distance vector $\underline{G}^{PD,i^*,j^*}$, recording the chance of being in the collection of categories successively further distanced from i^*, j^*. Then, given an IJ dimensioned unit vector d, $MFI(g)$ may be written as:

$$MFI(g) = \frac{1}{\varphi(IJ, i^*, j^*)} d'\underline{G}^{PD,i^*,j^*} \qquad [6]$$

Standardization.

Pearsons' concern in ordered categorical environments would have been that such measures of spread would not be "stable" for reliable comparison across groups without suitable locational standardisation. Sharpes' concern would have been that the location measure would have not been diluted by an appropriate measure of uncertainty, in essence it is a variation standardised location measure much like a standard normal statistic. All that remains for an Ordered Categorical Coefficient of Variation or its inverse ($OCCOV$ or $OCCOV^{-1}$) is to standardize $MFI(g)$ with an appropriate probability-based distance measure factor to render it unit free. Analogous to the continuous paradigm formulation of the mean as the integral of the survival function over the range of x (see [1] above) the sum of the SF values over all categories $\sum_{k=1}^{K}(1-G(k))$ could be considered so that:

$$OCCOV = (\mathrm{K}-1)MFI(g,k^*)/(\sum_{k=1}^{K}(1-G(k))) \qquad [7]$$

would provide an ordered categorical Coefficient of Variation analogue appropriate to for the situation at hand[11].

The inverse of $OCCOV$ is the ordered categorical paradigm equivalent of the Sharpe Ratio which is a risk or uncertainty adjusted average returns measure. Thus, it can be viewed as an outcome level measure diluted by a measure of uncertainty surrounding outcome levels.

[11] [7] is clearly dependent upon K, this is of no consequence when variates with a common K are being compared, but when variates with different K's are being compared their respective values of [7] should be rescaled by $\gamma(K) = {}^{(K-1)}/_{K}$

Axiomatics.

Suppose that $\underline{x}$ is the n x 1 dimensioned list of the category locations of n sampled individuals upon which the estimates of $\hat{p}_{fk}, k = 1,..,K$ are based, then $OCCOV$ is readily shown to satisfy the axioms of Anonymity (i.e. it is independent of the ordering of the list $\underline{x}$); Scale Invariance (it is independent of any arbitrary scale accorded the categories) and Population Independence (it will not change when the population is replicated and added to itself) it can also be shown to satisfy a weak version of the Pigou-Dalton Transfer Principal (when the presence in a higher category is reduced by 1 and the presence in a lower category increased by 1 without altering the mode, it will not increase). However, $OCCOV$ is not independent of K, the number of categories. While this is of no consequence when groups are being compared over the same number of categories, it does need attention when groups are being compared over different numbers of categories so that when comparing groups based upon different numbers of categories, it may be prudent to multiply $OCCOV$ by $1/(K + 1)$ for comparison purposes.

3. Two Examples.

To illustrate application of the Ordered Categorical Coefficient of Variation and its components, two examples are reported. To reflect concerns about the health and aging connection in the context of an aging population, a univariate analysis of Self- Reported Health outcomes in the United Kingdom over the pre-covid period 2010-2018 is pursued. To exemplify a multivariate application, an analysis of the progress of the multivariate distribution of experience and embodied human capital factors that underlay Human Resource stocks in Canada over the 2006-2016 decade.

3.1 Health Outcomes in the United Kingdom.

Given the propensity of health to decline with age and the health-longevity gender paradox[12], nations with ageing populations tend to experience declining overall health outcomes which may well differ by gender, raising concerns about inequalities in health outcomes as they relate to the aging process. A challenge with studying these phenomena is that health outcomes are ordered categorical in nature. Here the progress of health outcomes in the UK over the period 2010-2018 are examined using individual seven category self-reported ordered categorical health responses over six life cycle Age Group categories for males and females drawn from the Understanding Society Data Set[13]. Table A1 in the appendix reports details of the age groups, categories and the Probability Density Functions for male and female samples for the two years upon which the results are based.

Table 1 reports the average MOC, MFI and COV over the population for the two years. As may be observed, the Ordered Categorical Coefficient of Variation indicates an increase in relative health inequality over the period, the result of a significant decline in self-reported health levels

---

[12] It has long been understood that health outcomes deteriorate with age (Deaton and Paxson 1998, Kerkhofs and Lindeboom 1997) and even if age groups maintain their health status over time overall health will appear to deteriorate because of the composition effects of increasing proportions of older cohorts in the population. Recent work (Case and Paxon 2005, Nusselder et.al. 2010, Oksuzyan et.al. 2009, Van Oyen 2013) has highlighted a female-male health-longevity paradox - that women typically experience worse health outcomes than men throughout their lives, yet they tend to live longer.

[13] Understanding Society: the UK Household Longitudinal Study (UKHLS) is a University of Essex Institute for Social and Economic Research survey collecting data from participants by surveying the members of approximately 40,000 households a year. After incomplete records were excluded, a 2-period sample of 75487 subjects remained with recorded levels of self-reported Health, Age group and gender categories.

combined with a somewhat more marginal increase in health inequality levels. Table 2 reports more detailed level, inequality and resultant Ordered Categorical Coefficient of Variations for the respective groups and years.

Table 1. Overall Health Levels Inequalities and Coefficients of Variation 2010 - 2018

	Average Health Level Index (MOC) 2010	2018	Average Health Inequality Index (MFI) 2010	2018	Average Coefficient of Variation (COV) 2010	2018
Mean (Standard Error)	3.9006 (0.0167)	3.8226 (0.0194)	3.3170 (0.0090)	3.3373 (0.0107)	0.8516	0.8744
2010-2018 Difference (s.e.) "t"	0.0780 (0.0256) "3.0457"		-0.0203 (0.0140) "-1.4509"			

Note that health levels deteriorate significantly over the life cycle for both males and females in both years, furthermore they tend to be lower for females than for males. Typically, outcomes

Table 2. Health Level, Inequality and Coefficient of Variation by Gender over the Life Cycle.

	Females						Males					
	< 26	26-35	36-45	46-55	56-65	>65	< 26	26-35	36-45	46-55	56-65	>65
Levels MOC												
2010	4.1164	4.0930	3.8942	3.7343	3.7546	3.6737	4.3062	4.0883	3.9591	3.8107	3.7648	3.7570
Std. Error	0.0578	0.0552	0.0504	0.0538	0.0578	0.0540	0.0637	0.0647	0.0576	0.0614	0.0638	0.0582
2018	4.0500	3.9507	3.8324	3.6536	3.5555	3.7364	4.2555	3.9368	3.8978	3.7676	3.7237	3.8404
Std. Error	0.0730	0.0716	0.0645	0.0608	0.0639	0.0560	0.0781	0.0852	0.0735	0.0670	0.0705	0.0598
10-18 Difference	0.0664	0.1423	0.0618	0.0807	0.1991	-0.0627	0.0507	0.1515	0.0613	0.0431	0.0411	-0.0834
Standard Errors	0.0931	0.0904	0.0819	0.0812	0.0862	0.0778	0.1008	0.1070	0.0934	0.0909	0.0951	0.0834
t tests	0.7131	1.5740	0.7550	0.9940	2.3107	-0.8060	0.5031	1.4161	0.6565	0.4743	0.4323	-0.9994
Inequality MFS												
2010	3.3389	3.3890	3.3207	3.2103	3.2656	3.1762	3.3677	3.3968	3.4031	3.3857	3.3668	3.2605
Std. Error	0.0325	0.0299	0.0271	0.0279	0.0300	0.0286	0.0360	0.0361	0.0318	0.0329	0.0340	0.0310
2018	3.2735	3.3822	3.3499	3.2256	3.1685	3.3124	3.2730	3.3588	3.4913	3.4246	3.3732	3.4524
Std. Error	0.0411	0.0404	0.0358	0.0328	0.0337	0.0304	0.0435	0.0486	0.0416	0.0366	0.0381	0.0332
10-18 Difference	0.0654	0.0068	-0.0292	-0.0153	0.0971	-0.1362	0.0947	0.0380	-0.0882	-0.0389	-0.0064	-0.1919
Standard Errors	0.0524	0.0503	0.0449	0.0431	0.0452	0.0417	0.0564	0.0605	0.0524	0.0492	0.0511	0.0455
t tests	1.2483	0.1352	-0.6499	-0.3552	2.1496	-3.2637	1.6780	0.6278	-1.6842	-0.7906	-0.1253	-4.2196
COV												
2010	0.8111	0.8280	0.8527	0.8597	0.8698	0.8646	0.7821	0.8309	0.8596	0.8885	0.8943	0.8678
2018	0.8083	0.8561	0.8741	0.8829	0.8912	0.8865	0.7691	0.8532	0.8957	0.9090	0.9059	0.8990

at each age group are worse in 2018 than they were in 2010 signaling an almost universal intertemporal decline in health over the period, the one notable exception being the over 65's where both genders see an improvement over the period. With regard to inequality, the intertemporal changes are somewhat less universal though the senior group in both genders records a significant increase in inequality. However there does appear to be a decline in inequality over the life cycle in 2010 which is much less apparent in 2018. Overall, the coefficient of variation appears to be growing steadily over the life cycle for both genders with females exhibiting greater relative variation in 2010 than 2018 with miles exhibiting greater variation in 2018 rather than 2010.

3.2 Relative Variation in the distribution of Human Resource factors across the gender divide.

A nations Human Resource Stock (HRS), the aggregation of its constituent agents HRS's, is an amorphous amalgam of their Embodied Human Capital and Cumulated Experience. The fact that males and females face different labour market and life cycle circumstances and have different knowledge acquisition traits suggests that a nations HRS has been acquired and employed differently across the gender divide (Goldin 2014) and differences between the genders of the within gender level and variability of its possession is of interest. To this end an analysis of the corresponding Ordered Categorical Coefficient of Variation can be informative regarding the relative variation and the Sharpe ratio can reveal something about the uncertainty diluted level of human resources.

Assessing the levels is difficult since both components are fundamentally latent and unobservable. Experience - the agents productivity enhancing skills acquired by practice and learning by doing - can be proxied for by the passage of time or the recorded age group of the individual. Embodied Human Capital – the agents' education and training augmented innate abilities– can be proxied for by the Education and Training level they have received. Both proxies are ordered categorical variates and, beyond the afore-mentioned issues associated with using and combining artificially attributed cardinal scales to ordinal variates, their combination in some simple algebraic form is problematic. To examine the progress of Human Resource Stocks in Canada, data on the age, education and training status of individuals have been drawn from the Census of Canada Individual Files for the years 2006 and 2016[14].

The joint probability distributions (PDF) and survival functions (SF) over experience and education and training level groups for Canadian Males and Females in 2006 and 2016 are reported in Table A2 in the appendix.

Table 1 reports $OCCOV$ and Sharpe Ratios together with their components for the joint density for Males and Females in 2006 and 2016. What may be gleaned from Table 1 is that relative variation of human resource stocks increased for both Females and Males over the decade, but much more so for Males, a result of the substantial shift downwards in the Male modal location engendering a substantial increase in relative variation and a sharp reduction in the uncertainty moderated level of human resources.

Table 1. Bivariate Ordered Categorical Coefficient of Variation

	Modal Experience, Training Location	Variation	Location Value	$OCCOV$	Sharpe
Females 2006	3,3	3.0662	18.2842	0.1677	5.9630
Females 2016	4,2	3.6889	19.4335	0.1865	5.3619
Males 2006	3,3	3.0592	17.5026	0.1748	5.7208
Males 2016	1,2	10.5715	18.1212	0.5834	1.7140

[14] All agents over the age of 19 who received an income and reported age and educational status were included in the study resulting in 608538 observations in 2006 (312405 of which were female) and 610346 in 2016 (326676 of which were female). An individual's experience is proxied for by their age group category with 20-29, 30-39, 40-49, 50-59, 60-69 and ≥ 70 being the designated experience categories. The education and training embodied human capital levels are based on 5 ordered categories:- EDU1: Did not finish high school, EDU2: Completed High school, EDU3: Trade or Apprentice certification or University certification or diploma below bachelor degree level and EDU5: University certificate or diploma bachelor level and above, including masters and doctorates. EDU3 and 40-49 age group were deemed the Sufficient Human Resource level.

Table 2. Education Ordered Categorical Coefficient of Variation

		EDU1	EDU2	EDU3	EDU4	EDU5	$OCCOV$	Sharpe
Females 2006	PDF	0.2104	0.2792	**0.3109**	0.0570	0.1425	0.6346	1.5758
	Survival Function	0.7896	0.5104	0.1995	0.1425	0.0000		
Females 2016	PDF	0.1681	**0.3035**	0.2751	0.0399	0.2134	0.6367	1.5706
	Survival Function	0.8319	0.5284	0.2533	0.2134	0.0000		
Males 2006	PDF	0.2139	0.2591	**0.3502**	0.0446	0.1322	0.6139	1.6289
	Survival Function	0.7861	0.5270	0.1768	0.1321	0.0000		
Males 2016	PDF	0.2010	**0.3374**	0.2311	0.0314	0.1992	0.6462	1.5475
	Survival Function	0.7990	0.4616	0.2306	0.1992	0.0000		

Table 3. Experience Ordered Categorical Coefficient of Variation

		20-29	30-39	40-49	50-59	60-69	> 69	$OCCOV$	Sharpe
2006	PDF	0.1689	0.1749	**0.2169**	0.1751	0.1203	0.1439	0.5825	1.7167
Female	Survival Func	0.8311	0.6562	0.4393	0.2642	0.1439	0.0000		
2016	PDF	0.1673	0.1625	0.1647	**0.1921**	0.1609	0.1525	0.5891	1.6975
Female	Survival Func	0.8327	0.6702	0.5054	0.3133	0.1525	0.0000		
2006	PDF	0.1807	0.1782	**0.2232**	0.1872	0.1195	0.1113	0.5853	1.7085
Male	Survival Func	0.8193	0.6412	0.4180	0.2308	0.1113	0.0000		
2016	PDF	0.1904	0.1689	0.1685	**0.1959**	0.1537	0.1225	0.6360	1.5723
Male	Survival Func	0.8096	0.6407	0.4722	0.2763	0.1225	0.0000		

When the marginal Education and Experience distributions are considered in Tables 2 and 3 respectively, a greater increase in relative variation in both Education and Experience for males relative to females can be observed in both dimensions. When viewed separately, the individual dimension mode changes over time are similar across genders (mode levels lowering in the case of education and increasing in the case of experience), but is not reflected in the bivariate distribution (increasing experience, decreasing education for women, both decreasing for men).

Table 4. Education and Experience comparison.

	$OCCOV$	K adjusted $OCCOV$	K adjusted Sharpe
Education 2006 Female	0.6346	0.5077	1.9697
Education 2016 Female	0.6367	0.5094	1.9632
Education 2006 Male	0.6139	0.4911	2.0362
Education 2016 Male	0.6462	0.5170	1.9344
Experience 2006 Female	0.5825	0.4854	2.0601
Experience 2016 Female	0.5891	0.4909	2.0370
Experience 2006 Male	0.5853	0.4878	2.0502
Experience 2016 Male	0.6360	0.5300	1.8868

The relative variation in education as opposed to experience outcomes is examined in Table 4. To make this comparison the ordered categorical coefficient of variation needs to be adjusted by the number of categories in the respective variates. When this is done it can be observed that with the exception of males in 2016 education has greater relative variation than does experience so the position is reversed for 2016 males.

4. Conclusions.

By invoking the notion of probabilistic distance and developing measures of levels and spreads analogous to means and standard deviations in cardinal paradigms, the construction of a measure of relative ordinal variation which is unit free and comparable across populations is possible, despite the lack of cardinality. Furthermore, the measures are easily implemented in multidimensional environments. Exemplifying applications of the measure to examine the relative variability in univariate health outcomes in the UK and multivariate human resource factors across the gender divide in Canada revealed substantial and meaningful differences.

References.

Aerts, S., Haesbroeck, G. and C. Ruwet (2015) "Multivariate coefficients of variation Comparison and influence functions" Journal of Multivariate Analysis. 142 183-198.

Albert, A., and L. Zhang (2010) "A novel definition of the multivariate coefficient of variation" Biometrical Journal, Volume 52, 5, 667-675

Allison, R. A. and J. E. Foster (2004). Measuring health inequality using qualitative data. Journal of Health Economics 23, 505- 24.

Anderson G.J. (2004) "Toward an Empirical Analysis of Polarization" Journal of Econometrics 122 1-26

Anderson G.J. and G. Yalonetzky (2023) "The mode and the measurement of inequality with single and multiple ordinal variables" Mimeo Economics Department University of Toronto.

Bedeian, A. G., & Mossholder, K. W. (2000). On the Use of the Coefficient of Variation as a Measure of Diversity. Organizational Research Methods, 3(3), 285–297

Blair, J. and M.G. Lacy (2000) "Statistics of Ordinal Variation" Sociological Methods and Research 28 251-280.

Bond T.N. and K. Lang (2019) "The Sad Truth about Happiness Scales," Journal of Political Economy 127, no. 4: 1629-1640.

Case A. and Paxson C (2005) Sex differences in morbidity and mortality. Demography 42:189–214

Cantril, H. (1965). The pattern of human concerns. New Brunswick, NJ: Rutgers University Press

Champernowne, D. G. and Cowell, F. A. (1999). Economic Inequality and Income Distribution. Cambridge University Press.

Deaton, Angus S., and Christina H. Paxson, "Aging and Inequality in Income and Health," The American Economic Review 88(2), (1998), 248–253

Galton, F. (1894), Natural Inheritance (5th ed.), New York: Macmillan and Company.

Goldin, C. (2014) "A Grand Gender Convergence: Its Last Chapter" American Economic Review 2014, 104(4): 1–30.

Zahra Jalilibal, Amirhossein Amiri, Philippe Castagliola, Michael B.C. Khoo. Monitoring the coefficient of variation: A literature review. Computers & Industrial Engineering, 2021, 161, pp.107600. ff10.1016/j.cie.2021.107600ff. ffhal-03336179f

Kerkhofs M. and M. Lindeboom (1997) "Age Related Health Dynamics and Changes in Labour Market Status" Health Economics 6: 407-423.

Kobus, M. (2015). Polarization measurement for ordinal data. Journal of Economic Inequality 13, 275-297.

Kobus M. and Kurek R. (2019) "Multidimensiona1 polarization for ordinal data" The Journal of Economic Inequality 17: 301-317.

Kvalseth, T.O. (2017) "Coefficient of Variation: the Second-Order Alternative" Journal of Applied Statistics 44 402-415.

Likert R. (1932) "A Technique for the measurement of Attitudes" Archives of Psychology 1-55.

Marshall JC, Murthy S, Diaz J, Adhikari N, Angus DC, Arabi YM, et al. A minimal common outcome measure set for COVID-19 clinical research. Lancet Infect Dis. 2020;20(8):e192–7.

Mendelson, H. (1987). Quantile- preserving spread. Journal of Economic Theory 42, 334-51.

Nusselder W, Looman C, Van Oyen H, Robine JM, Jagger C (2010) Gender differences in health of EU10 and EU15 populations: the double burden of EU10 men. European Journal of Ageing 7:219–227

Oksuzyan A, Petersen I, Stovring H, Bingley P, Vaupel JW, Christensen K (2009) The male-female health–survival paradox: a survey and register study of the impact of sex-specific selection and information bias. Annals of Epidemiology 19:504–511

Herman Van Oyen, Wilma Nusselder, Carol Jagger, Petra Kolip, Emmanuelle Cambois, Jean-Marie Robine (2013) “Gender differences in healthy life years within the EU: an exploration of the ‘‘health–survival’’ paradox” International Journal of Public Health (2013) 58:143–155

Pearson K. (1896) “Mathematical contributions to the theory of evolution: Regression, heredity and panmixia” Philos. Transactions of the Royal Society London. A, Math. Phys. Eng. Sci 187, 253– 318.

Pew Research Center, (2023), “How Public Polling Has Changed in the 21st Century”.

Rankin L. Cerebral vascular accidents in patients over the age of 60. II. Prognosis. Scott Med J. 1957; 2: 200–215.

Reyment, R.A. (1960) Studies on Nigerian upper Cretaceous and lower tertiary ostracoda: part 1 Senonian and maastrichtian ostracoda. Stockolm Contributions in Geology 7 1-238.

Schroder, C. and Yitzhaki, S. (2017). Revisiting the evidence for cardinal treatment of ordinal variables. European Economic Review, 92:337-358.

Chris J. Selman, Katherine J. Lee, Clare L. Whitehead, Brett J. Manley, and Robert K. Mahar (2023) “Statistical analyses of ordinal outcomes in randomised controlled trials: protocol for a scoping review” Trials. 2023; 24: 286.

Sharpe, W.F. (1964) “Capital Asset Prices: A theory of Market Equilibrium Under Conditions of Risk” Journal of Finance 428-442.

Sharpe, W. (1994) “The Sharpe ratio”, Journal of Portfolio Management, Vol. 21, No. 1, Fall, pp. 49-58.

Sortino F.A. and L.N. Price (1994) “Performance Measurement in a Downside Risk Framework” The Journal of Investing 3 59-64.

Van Valen L. (1974) “multivariate structural statistics in Natural History” Journal of Theoretical Biology 45 235-247.

Voinov V.G. and M.S. Nikulin (1966) “Unbiased Estimators and their Applications” Multivariate Case Vol. 2 Kluwer.

Weber, E. U., Shafir, S., & Blais, A.-R. (2004). Predicting Risk Sensitivity in Humans and Lower Animals: Risk as Variance or Coefficient of Variation. Psychological Review, 111(2), 430–445.

Apppendix.

Inference.

Following Rao (2009), given an independent random sample of size n, $\widehat{\underline{p_g}}$, the estimator of the vector of outcome probabilities $\underline{p_g}$ is multivariate normal:

$$\sqrt{n}\left(\widehat{\underline{p_g}} - \underline{p_g}\right) \sim N\left(\underline{0}, V_g\right)$$

where:

$$V_g = \begin{bmatrix} p_{1,g} & 0 & 0 & . & 0 \\ 0 & p_{2,g} & 0 & . & 0 \\ 0 & 0 & p_{3,g} & . & 0 \\ . & . & . & . & 0 \\ 0 & 0 & 0 & . & p_{\mathrm{K},g} \end{bmatrix} - \begin{bmatrix} p_{1,g} \\ p_{2,g} \\ . \\ . \\ p_{\mathrm{K},g} \end{bmatrix} [p_{1,g} \quad p_{2,g} \quad . \quad . \quad p_{K,g}]$$

Given a K dimensioned square cumulation matrix C_{k^*} with typical element $c_{i,j}$ $i, j = 1, .., I$ where for $i, j < k^*$, $c_{i,j} = 1$ $when\ j \geq i\ and\ 0$ otherwise, and for $i, j > k^*$, $c_{i,j} = 1$ $when\ j \leq i\ and\ 0$ otherwise, all other elements of the matrix are 0 [15], and a summation vector $\underline{d}$ which is a K x 1 column of ones, then $\underline{G}^{PD,k^*} = C_{k^*}\underline{p_g}$ and $\underline{\hat{G}_g}^{PD,k^*} = C_{k^*}\widehat{\underline{p_g}}$ so that:

$$\sqrt{n}\left(\underline{\hat{G}_g}^{PD,k^*} - \underline{G_g}^{PD,k^*}\right) \sim N\left(\underline{0}, C_{k^*} V_g C_{k^*}'\right)$$

So that $\widehat{MFI}(g)$, estimates of $MFI(g)$ will be such that:

$$\sqrt{n}\left(\widehat{MFI}(g) - MFI(g)\right) \sim N\left(0, \frac{1}{\varphi(K, k^*)^2} \underline{d}' C_{k^*} V_g C_{k^*}' \underline{d}\right)$$

[15] As an example, for I = 6 and $k^* = 3$, C_{k^*} is of the form:

$$C_{k^*} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Table A1. PDF's For the UK Health Outcomes.

Category*	Females						Males					
	< 26	26-35	36-45	46-55	56-65	>65	< 26	26-35	36-45	46-55	56-65	>65
2010												
1	0.0259	0.0307	0.0445	0.0589	0.0564	0.0558	0.0227	0.0236	0.0334	0.0502	0.0491	0.0498
2	0.0497	0.0579	0.0728	0.0796	0.0852	0.0842	0.0357	0.0538	0.0582	0.0701	0.0806	0.0810
3	0.1170	0.1249	0.1355	0.1586	0.1481	0.1553	0.0957	0.1320	0.1425	0.1476	0.1491	0.1458
4	0.0939	0.0722	0.0765	0.0772	0.0748	0.1011	0.0857	0.0770	0.0839	0.0806	0.0816	0.0960
5	0.1720	0.1436	0.1473	0.1274	0.1265	0.1419	0.1667	0.1668	0.1570	0.1389	0.1364	0.1399
6	0.3860	0.4299	0.4087	0.3935	0.4112	0.3622	0.4058	0.4085	0.4138	0.4276	0.4236	0.3882
7	0.1555	0.1408	0.1147	0.1048	0.0978	0.0995	0.1877	0.1383	0.1112	0.0850	0.0796	0.0993
2018												
1	0.0254	0.0362	0.0383	0.0430	0.0641	0.0406	0.0267	0.0271	0.0214	0.0369	0.0452	0.0416
2	0.0616	0.0621	0.0787	0.0998	0.0991	0.0871	0.0418	0.0596	0.0643	0.0771	0.0901	0.0658
3	0.1129	0.1226	0.1345	0.1564	0.1598	0.1425	0.0988	0.1306	0.1438	0.1512	0.1394	0.1389
4	0.1049	0.0916	0.0894	0.0938	0.0907	0.1095	0.0920	0.1078	0.1058	0.1048	0.0994	0.0941
5	0.1834	0.1826	0.1755	0.1610	0.1442	0.1505	0.1599	0.1975	0.1763	0.1449	0.1430	0.1611
6	0.3565	0.3912	0.3871	0.3604	0.3647	0.3850	0.3843	0.3618	0.4071	0.4165	0.4128	0.4209
7	0.1553	0.1137	0.0965	0.0856	0.0774	0.0848	0.1965	0.1156	0.0813	0.0686	0.0701	0.0776

*The seven categories were: 1 Completely dissatisfied, 2 Mostly dissatisfied, 3 Somewhat dissatisfied, 4 neither dissatisfied nor satisfied, 5 Somewhat satisfied, 6 Mostly satisfied and 7 Completely satisfied. Note that category 6, the mostly satisfied category is always the modal category.

Table A2*.

		Joint PDF					Survival Function				
		EDU1	EDU2	EDU3	EDU4	EDU5	EDU1	EDU2	EDU3	EDU4	EDU5
Females 2006	20-29	0.0181	0.0552	0.0543	0.0077	0.0337	0.9819	0.9268	0.8725	0.8648	0.8311
	30-39	0.0182	0.0390	0.0666	0.0105	0.0405	0.9637	0.8696	0.7487	0.7305	0.6562
	40-49	0.0294	0.0628	0.0795	0.0129	0.0323	0.9343	0.7773	0.5769	0.5459	0.4393
	50-59	0.0320	0.0540	0.0557	0.0114	0.0219	0.9023	0.6913	0.4351	0.3928	0.2642
	60-69	0.0408	0.0321	0.0304	0.0081	0.0089	0.8615	0.6184	0.3318	0.2814	0.1439
	> 69	0.0719	0.0361	0.0243	0.0066	0.0050	0.7896	0.5104	0.1995	0.1425	0.0000
Females 2016	20-29	0.0137	0.0588	0.0436	0.0052	0.0461	0.9863	0.9275	0.8839	0.8787	0.8327
	30-39	0.0135	0.0364	0.0511	0.0067	0.0548	0.9728	0.8775	0.7828	0.7710	0.6702
	40-49	0.0166	0.0404	0.0544	0.0075	0.0459	0.9562	0.8206	0.6714	0.6522	0.5054
	50-59	0.0272	0.0638	0.0597	0.0080	0.0335	0.9290	0.7296	0.5209	0.4936	0.3133
	60-69	0.0332	0.0580	0.0408	0.0066	0.0223	0.8958	0.6385	0.3890	0.3551	0.1525
	> 69	0.0639	0.0461	0.0256	0.0060	0.0108	0.8319	0.5284	0.2533	0.2134	0.0000
Males 2006	20-29	0.0265	0.0673	0.0565	0.0067	0.0238	0.9736	0.9063	0.8498	0.8431	0.8193
	30-39	0.0244	0.0444	0.0675	0.0083	0.0335	0.9491	0.8374	0.7135	0.6985	0.6412
	40-49	0.0390	0.0557	0.0873	0.0100	0.0311	0.9101	0.7427	0.5313	0.5063	0.4180
	50-59	0.0372	0.0473	0.0683	0.0093	0.0251	0.8728	0.6582	0.3785	0.3442	0.2308
	60-69	0.0374	0.0245	0.0400	0.0059	0.0117	0.8354	0.5962	0.2766	0.2364	0.1113
	> 69	0.0493	0.0199	0.0306	0.0044	0.0071	0.7861	0.5270	0.1768	0.1321	0.0000
Males 2016	20-29	0.0236	0.0855	0.0397	0.0041	0.0375	0.9764	0.8909	0.8512	0.8471	0.8096
	30-39	0.0223	0.0525	0.0454	0.0051	0.0436	0.9540	0.8160	0.7309	0.7217	0.6407
	40-49	0.0246	0.0524	0.0446	0.0061	0.0408	0.9295	0.7390	0.6093	0.5940	0.4722
	50-59	0.0412	0.0634	0.0494	0.0062	0.0356	0.8882	0.6344	0.4552	0.4337	0.2763
	60-69	0.0373	0.0507	0.0333	0.0055	0.0269	0.8509	0.5464	0.3339	0.3068	0.1225
	> 69	0.0520	0.0328	0.0186	0.0043	0.0149	0.7990	0.4616	0.2306	0.1992	0.0000

* Note that the distributions have multiple nodes with point densities greater than all contiguous points but there is always a unique universal modal with a density value greater than any other point in the range.