Gli Indici di Concentrazione: *una misura del rischio*

Maurizio Rosina

INTRODUZIONE

Sono una azienda che produce tanti diversi tipi di prodotti ed ha un nutrito portafoglio Clienti. Mi accorgo, però, che le vendite si 'concentrano' su pochi tipi di prodotti. Corro un grande 'rischio' se qualcuno di tali prodotti smettesse di essere venduto. Poi, magari, mi accorgo che le vendite si 'concentrano' su solo pochi Clienti. Corro un grande 'rischio' se qualcuno di tali importanti Clienti smettesse di acquistare. Qualcuno potrebbe dubitare che tale ipotetica azienda stia correndo dei grandi rischi, e che il valore dell'indice che misura la 'concentrazione' possa essere visto come una misura del 'rischio'? Se, viceversa, tale ipotetica azienda proponesse una 'equidistribuzione' nella vendita dei tipi di prodotti ed una 'equidistribuzione' delle vendite tra i sui tanti Clienti, qualcuno potrebbe dubitare che anche se qualcuno dei prodotti non venisse più venduto o qualcuno dei Clienti non acquistasse più, in ogni caso l'azienda sarebbe in grado di sopportare tali difficoltà?
Nel seguito si tratterà proprio degli Indici che misurano la concentrazione/equidistribuzione, e se ne proporrà un utilizzo come modelli e strumenti per la misura del rischio.

Date n entità (ovvero n «unità statistiche») oggetto di studio di un particolare aspetto (aspetto detto «carattere») di un fenomeno su di esse osservabile, in generale si ha che il carattere è tanto più «concentrato» quanto maggiore è la frazione dell'ammontare complessivo del carattere che spetta alla frazione di unità statistiche che ne possiedono di più.

Quanto sopra detto non è un gioco di parole, anche se, forse, non è proprio di immediata appercezione. Lo sono, probabilmente, un po' di esempi.

L'analisi della concentrazione è, ad esempio, importante sia per i soggetti «Azienda» che per un ipotetico soggetto «Stato» che a regole aziendali voglia/debba ricondursi. Per tali soggetti è importante determinare il livello di rischio nelle rispettive attività.

Un tipo di rischio da tenere sotto controllo è, ad esempio, quello legato ad una eccessiva «concentrazione» del fatturato per prodotto o per cliente. Ovviamente per il soggetto «Stato» il *fatturato per prodotto* potrebbe essere l'incoming fiscale per tipologia di

tassa o di tariffa, oppure l'incoming derivante da tariffe correlate a specifici codici ATECO; ed il *fatturato per cliente* potrebbe essere l'entità dell'incoming fiscale per le varie classi e/o stratificazioni dei cittadini-contribuenti o per le varie localizzazioni geografiche dei cittadini-contribuenti e relativa stratificazione delle loro attività ricadenti in codici ATECO.

Per entrambi i soggetti «Azienda» e «Stato», rendersi conto che il fatturato è fortemente legato a soli pochi «prodotti» e/o che solo pochi «clienti» contribuiscono al fatturato, le porrebbe in una potenziale condizione di "elevato rischio". Basterebbe infatti che tali «prodotti» non fossero più di interesse e/o che qualcuno dei pochi «clienti» non contribuisse più per creare seri problemi finanziari.

Esempio classico è quello di una Azienda che operi nel settore manifatturiero. Il superamento tecnologico (obsolescenza) del prodotto di «punta» e/o una diminuzione della «qualità percepita» del brand e/o in quanto prodotto, porterebbe ad una seria crisi, come pure l'avere una situazione in cui il fatturato fosse legato a pochi grandi clienti. Per evitare e/o tenere sotto controllo situazioni a rischio, l'Azienda manifatturiera dovrebbe verificare se il fatturato è dovuto in egual misura a tutti i prodotti (cioè se si ha una equilibrata equidistribuzione del fatturato tra i vari prodotti) oppure, viceversa, se il fatturato deriva in gran misura dalla vendita di soli pochi «prodotti» (caso di concentrazione). Ed analogamente dovrebbe analizzare se il fatturato è dovuto in egual misura a molti/tutti i clienti (cioè se si ha equidistribuzione del fatturato tra i vari clienti) oppure, viceversa, se il fatturato deriva in larga misura da pochi grandi clienti (caso di elevata concentrazione).

Nelle situazioni in cui si rileva una elevata concentrazione generalmente si annida un potenziale «rischio». Occorre quindi prevedere di attuare **misure della concentrazione, atte ad individuare potenziali situazioni di «rischio».**

GLI INDICI DI CONCENTRAZIONE

La misura della concentrazione ha dato luogo allo sviluppo di tutta una serie di Indici di Concentrazione, basati su analisi di variabili diverse.

Nel seguito, senza cercare di omettere alcun passaggio matematico, verranno presentati 5 diversi tipi di Indici di Concentrazione, accumunati dal fatto che tutti operano solo sulla conoscenza di due nozioni, date dal *numero n delle unità statistiche oggetto di esame*, e dalla *frazione dell'ammontare complessivo del «carattere» oggetto di esame* (ad es. il fatturato, il numero addetti, i clienti, ecc) *che ciascuna i-esima unità statistica possiede*.

Una delle caratteristiche più richieste per gli Indici di Concentrazione è che *tutti* gli Indici operino fornendo valori entro uno stesso range di valori, auspicabilmente [0..1]. Ciò nei desiderata pratici; poi talvolta però gli stessi autorevoli testi che citano tale importate caratteristica richiesta nella «pratica» omettono di perseguirla, e presentano indici che operano nei range [1/n…1], [0..log(n)], ecc. Nel seguito tutti gli Indici di Concentrazione verranno ricavati in modo che forniscano risultati nel range [0..1], come «pratica» d'uso richiede.

Prima di cominciare a ricavare i vari indici, osserviamo che detta A una variabile rappresentativa di un carattere in esame, e dette a_1, a_2,.. a_i… a_n n determinazioni (valori) di A, il valore totale del carattere rilevato sarà $T = \sum_{i=1}^{n} a_i$. La **quota** del carattere detenuta dalla i-esima determinazione sarà quindi $s_i = a_i/T$, per cui, ovviamente, la somma delle quote detenute da tutte le determinazioni sarà sempre pari ad uno, ovvero $\sum_{i=1}^{n} s_i = 1$.

Ad esempio, se A fosse la variabile rappresentativa del fatturato di una industria composta di n aziende, e a_1, a_2, ... a_i … a_n fossero i fatturati delle varie aziende, allora s_i, con $i \in [1..n]$, rappresenterebbe la quota del fatturato totale detenuta dalla i-esima Azienda, valendo sempre che $s_i \in [0..1]$ e $\sum_{i=1}^{n} s_i = 1$.

Tutti gli indici di concentrazione verranno nel seguito ricavati solo sulla base della conoscenza del numero e dei valori delle n determinazioni, ovvero dalla conoscenza dei valori delle n quote.

Nel seguito verranno presentati i seguenti 5 Indici di Concentrazione, ben noti in letteratura:

- Herfindal
- Hanna e Kay ($\alpha = 1{,}5$ e $\alpha = 2{,}5$)
- Hall e Tideman
- Horvat
- Theil

Tutti gli indici verranno 'normalizzati' al fine che forniscano valori nel range [0..1], con valore 0 rappresentativo di equidistribuzione delle quote tra tutte le determinazioni/unità statistiche, e valore 1 rappresentativo di massima concentrazione. La massima concentrazione si ha quando una sola determinazione/unità statistica, tra tutte quelle in esame, assomma in sé l'intero ammontare del carattere. Ragionando in termini di quote ciò significa che una sola determinazione/unità statistica presenta quota del carattere pari ad 1, o, in termini percentuali, il 100%.

INDICE di HERFINDAL

È l'indice di concentrazione forse più noto ed utilizzato. La sua formulazione di base è semplicissima

$$H = \sum_{i=1}^{n} s_i^2$$

in cui n sono il numero delle determinazioni, ed s_i, con i = 1..n, sono le relative quote. Si noti come l'elevamento a potenza esalti l'importanza delle quote di valore maggiore, ovvero, per dirla in altro modo, come le determinazioni di minori dimensioni contribuiscano in misura meno che proporzionale alla determinazione del valore dell'indice. L'indice di Herfindal (in realtà di Herfindal e Hirschman) si caratterizza proprio per tale proprietà.

Se vi fosse una equidistribuzione dei valori delle determinazioni, ovvero delle quote, (ovvero se ciascuna determinazione a_i avesse la stessa porzione del valore Totale, pari a T/n, ciò che comporterebbe, in termini di quote, $s_i = 1/n$) Herfindal varrebbe

$$H = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 = n\,\frac{1}{n^2} = 1/n$$

Viceversa, se vi fosse massima concentrazione, ovvero una sola determinazione possedesse l'intero ammontare, quindi possedesse una quota pari a 1/1 = 1, con tutte le altre quote uguali a zero, Herfindal varrebbe $H = \sum_{i=1}^{n} s_i^2 = (1)^2 = 1$.

H fornisce, quindi, valori nel range [1/n..1], range dipendente da n. Normalizziamo quindi H mappandolo nel range [0..1], indipendente da n, ed a tal fine semplicemente poniamo[1]:

$$\mathbf{H^* = (H - 1/n)/(1 - 1/n)}$$

con H* che ora fornisce valori nel range [0..1], valendo 0 nel caso di equidistribuzione e 1 nel caso di massima concentrazione.

INDICE di HANNA e KAY

Una quasi ovvia estensione dell'indice di Herfindal è ipotizzare valori diversi dell'esponente nell'elevamento a potenza. Da qui nasce l'idea dell'indice di Hanna e

[1] Mappare un [min .. x .. max] in [0 .. t .. 1] è il ricavare, per qualsiasi valore x nell'intervallo xmin, xmax il corrispondente valore di t nell'intervallo 0, 1, ed è ottenibile da x = xmin + t (xmax – xmin)

Kay, che in realtà, individua una «famiglia» di indici, dipendentemente dal valore dell'esponente. La formulazione di 'base' dell'indice Hanna-Kay è piuttosto semplice

$$HK = \sum_{i=1}^{n} s_i^{\alpha} \quad \text{con } \alpha > 0 \text{ e } \alpha \neq 1$$

Se vi fosse una equidistribuzione di valori delle determinazioni, ovvero per ogni quota valesse $s_i = 1/n$, Hanna-Kay varrebbe

$$HK = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^{\alpha} = n(1/n^{\alpha}) = n*n^{-\alpha} = n^{1-\alpha}$$

Viceversa, se vi fosse massima concentrazione, ovvero una sola determinazione possedesse l'intero ammontare, quindi esistesse una quota $s_i^{\alpha} = 1$, con tutte le altre quote uguali a zero, Hanna-Kay varrebbe $HK = 1^{\alpha} = 1$.

HK di 'base' fornisce, quindi, valori nel range $[n^{1-\alpha}..1]$, dipendente da n e da α, mentre a noi interessa operare nel classico range [0..1]. Tramite una sequenza di trasformazioni/mapping (vedi APPENDICE) è possibile giungere ad Hanna-Kay che fornisce valori nel range [1/n..1].

$$HK^{+} = {1}\Big/{\left(\sum_{i=1}^{n} s_i^{\alpha}\right)^{(1/1-\alpha)}} = \left(\sum_{i=1}^{n} s_i^{\alpha}\right)^{-(1/1-\alpha)} = \left(\sum_{i=1}^{n} s_i^{\alpha}\right)^{1/(\alpha-1)}$$

La formulazione HK+ è la più nota ed utilizzata per Hanna-Kay. Però ancora presta il fianco ad operare in un range dipendente da n.

Operiamo, quindi, la classica normalizzazione

$$\mathbf{HK^{*} = (HK^{+} - 1/n)/(1 - 1/n)}$$

con HK* che ora fornisce valori nel range [0..1].

Permane il problema di che valori assegnare al parametro α. Si noti come per α = 2 si ricade nella formulazione di Herfindal, infatti

$$HK^{+} = \left(\sum_{i=1}^{n} s_i^{2}\right)^{1/(2-1)} = \sum_{i=1}^{n} s_i^{2} = H$$

Generalmente, quindi, si preferisce utilizzare in Hanna-Kay valori di α rispettivamente minori o maggiori di 2.

Nel seguito verranno proposti alcuni esempi di utilizzo nei quali il calcolo dell'indice di Hanna-Kay verrà eseguito per due valori di alfa a cavallo di α = 2, e precisamente per i valori α = 1,5 ed α = 2,5.

INDICE di HALL e TIDEMAN

L'idea di base di questo indice è di ordinare le determinazioni (ovvero, di riflesso, le quote) in senso non crescente (ovvero dalla più grande alla più piccola), quindi «pesare» ciascuna quota in modo crescente, utilizzando quale peso la sua posizione nell'ordinamento. Ne consegue che le quote maggiori saranno le prime nel ranking, e quindi saranno pesate con valori minori delle successive. La formulazione di 'base' di Hall-Tideman risulta

$$HT = \left(\sum_{i=1}^{n} i s_i\right) \quad \text{in cui } s_1 \geq s_2 \geq \ldots \geq s_n$$

In questa formulazione di 'base' HT fornisce una misura *inversa* rispetto alla concentrazione, ed opera nel range [1..(n+1)/2], nel quale il valore che rappresenta l'equidistribuzione qui risulta essere *l'estremo superiore* del range.

Per verificarlo osserviamo che in caso di equidistribuzione, ovvero $s_i = 1/n$, $i \in [1..n]$, HT vale

$$(1/n \sum_{i=1}^{n} i) = 1/n * n(n+1)/2 = (n+1)/2$$

e nel caso di massima concentrazione vale $\sum_{i=1}^{1} 1 * 1 = 1$.

Il range[1.. (n+1)/2] è, ovviamente, assai poco fruibile, ed HT fornisce una misura *inversa*. Tramite una sequenza di trasformazioni/mapping (vedi APPENDICE) è possibile giungere ad Hanna-Kay che fornisce una misura diretta e propone valori nel range [1/n..1].

$$HT^{+} = 1/[2(\sum_{i=1}^{n} i s_i) - 1]$$

Questa è la formulazione più nota ed usata di Hall-Tideman, ma ha ancora lo svantaggio di dipendere da n.

Operiamo, quindi, la classica normalizzazione finale che permetterà di operare nel range [0..1], ovvero

$$\mathbf{HT^{*} = (HT^{+} - 1/n)/(1 - 1/n)}$$

con HT* che ora fornisce valori nel range [0..1].

INDICE di HORVAT

Tale indice è talvolta anche noto come Comprehensive Concentration Index (CCI). L'idea di base di questo indice è inizialmente ordinare le determinazioni (ovvero, di riflesso, le quote) in senso non crescente (dalla più grande alla più piccola), esattamente come fatto per Hall-Tideman, quindi operare il calcolo trattando e pesando in modo differente la quota più grande rispetto alle rimanenti. L'indice di Horvat può, infatti, essere espresso come

$$HV = s_1 p_1 + \sum_{i=2}^{n} s_i^2 \, p_i$$

in cui $s_1 \geq s_2 \geq \ldots \geq s_n$ e pesi $p_1 = 1$ e $p_i = (1+(1- s_i)) = (2 - s_i)$ per $i \in [2..n]$, da cui

$$HV = s_1 + \sum_{i=2}^{n} s_i^2 (2 - s_i)$$

Horvat può essere visto composto da due termini, il primo relativo alla quota più grande (la s_1) pesata con peso unitario, alla quale poi vengono sommati i valori delle restanti quote, con ciascuna restante quota elevata al quadrato e pesata per un fattore moltiplicatore inversamente proporzionale alla sua grandezza. Si noti infatti che ciascuna quota corrente s_i (con $i > 1$) è elevata al quadrato e moltiplicata per $(1 + (1 - s_i))$, ovvero viene moltiplicata per il valore totale delle quote (valore pari ad 1) a cui si somma il valore totale a cui viene sottratta la quota corrente. Il fattore moltiplicatore è quindi sempre ≥ 1 e inversamente proporzionale alla grandezza di ciascuna quota, ovvero, in definitiva, si pesano di più le quote più piccole (si noti la diversa logica di conduzione del calcolo rispetto all'indice di Herfindal, che tende ad esaltare le quote di valore maggiore).

I pesi $p_i = (2 - s_i)$ propongono quindi valori $1 < p_i \leq 2$, con $i \in [2..n]$.

In particolare se $s_i \approx 0 \rightarrow p_i \approx 2$ e se $s_i \approx 1 \rightarrow p_i \approx 1$ pur senza poter mai raggiungere il valore 1, in quanto tale valore lo potrebbe possedere solo s_1, qualora s_1 assommasse in sé la totalità delle quote (condizione di massima concentrazione).

L'indice di Horvat opera quindi in un range [min .. 1] in cui il valore min rappresenta quello di equidistribuzione. Si può facilmente dimostrare (vedi APPENDICE) che il valore min = $[(3n^2 - 3n + 1)/n^3]$.

L'indice di Horvat, ed è questa la sua formulazione più nota ed utilizzata, è quindi dato, come già detto, da

$$HV = s_1 + \sum_{i=2}^{n} s_i^2 (2 - s_i)$$

e fornisce valori nel range $[(3n^2 – 3n + 1)/n^3 .. 1]$, range dipendente da n.

Operiamo, quindi, la consueta classica normalizzazione finale che permetterà di operare nel range [0..1], ovvero

$$\mathbf{HV^* = (HV - [(3n^2 – 3n + 1)/n^3])/(1 - [(3n^2 – 3n + 1)/n^3])}$$

con HV* che ora fornisce valori nel range [0..1].

INDICE di THEIL

Questo indice è correlato al concetto di misura del «disordine» (ovvero alla misura della entropia), inquadrabile come una misura della non uniformità di una distribuzione. L'indice deriva, come noto, dai lavori di Shannon nel campo della teoria dell'informazione. La 'Shannon entropy' è il valore atteso della informazione contenuta in un messaggio. La formulazione di 'base' di tale misura è

$$T = - \sum_{i=1}^{n} s_i \log(s_i) = \sum_{i=1}^{n} s_i \log(1/s_i)$$

In tale formulazione si assume che eventuali valori $s_i = 0$ forniscano un contributo nullo, ovvero pari a zero, alla sommatoria. Con tale formulazione se si ha uniformità di distribuzione (ovvero equidistribuzione, cioè quote tutte uguali a 1/n) si perviene a

$$\sum_{i=1}^{n} \frac{1}{n} \log(1/(1/n)) = n * \frac{1}{n} * [\log(1) - \log(1/n)] =$$
$$- \log(1/n) = - \log(1) + \log(n) = \log(n)$$

e se vi è massima concentrazione $1*\log(1) = 0$. Ovvero tale formulazione fornisce valori nel range [log(n)..0], assai poco piacevole per i nostri scopi. Tramite una sequenza di trasformazioni/mapping (vedi APPENDICE) è possibile giungere a Theil che fornisce una misura diretta e propone valori nel range [1/n..1].

$$T^+ = \sum_{i=1}^{n} s_i * \log(n * s_i) = \sum_{i=1}^{n} [s_i * \log\left(s_i / 1/n\right)]$$

La T^+ è la formulazione più nota ed usata dell'indice di Theil, la quale opera nel range [0..log(n)]. Basterà ora normalizzare per ottenere un indice di Theil che fornisca valori nel range [0..1]. Per normalizzare basterà trasformare la funzione dividendola per log(n), ovvero

$$\mathbf{T^* = T^+/\log(n) = \sum_{i=1}^{n} [s_i * log\left(s_i / 1/n\right)]/log(n)}$$

con T* che ora fornisce valori nel range [0..1].

L'INTERPRETAZIONE DEI VALORI DEGLI INDICI

Esposti i 5 Indici di Concentrazione, tutti normalizzati per operare nel range [0..1], occorre ora entrare in un argomento poco frequentato.

Come interpretare i valori degli indici?

Si noti che sebbene tutti gli indici operino sugli stessi parametri e forniscano valori nello stesso range [0..1], ciascuno di essi fornisce risultati secondo un diverso modello di calcolo.

Gli unici due risultati con uguale interpretazione per tutti gli indici sono che al valore 0 corrisponde 'equidistribuzione', ed al valore 1 'massima concentrazione'. Qualsiasi altro valore intermedio ricadrà nel range [0..1], ma tramite un modello di calcolo diverso. Per dirla in altri termini non è così semplice, anche sullo stesso set di dati, paragonare (o meglio mettere in corrispondenza) i valori dei 5 indici. Talché, ad esempio, a volte nella interpretazione di tali valori si fa riferimento a valori di 'soglia' definiti da una qualche Autorità (vedi ad es. in wikipedia, nel caso dell'indice di Herfindal, l'affermazione «… Secondo le "US Merger Guidelines", un valore di … compreso tra … indica un mercato moderatamente concentrato, mentre un valore superiore ne indica uno fortemente concentrato ...».

Ora qui non si pretende di dare soluzione ad un problema a cui valenti matematici/statistici ed economisti non hanno dato soluzione. Si fornirà, quindi, solamente un metodo che personalmente trovo di qualche utilità nell'interpretare i valori dei vari indici.

Dato un set di n determinazioni di cui calcolare gli indici di concentrazione, ipotizziamo di voler conoscere il valore che i vari indici proporrebbero qualora la somma dei valori delle n determinazioni fosse equidistribuita nelle prime n/2 determinazioni. Detta T la somma dei valori delle n determinazioni, ci poniamo quindi nell'ipotesi che le prime n/2 ne possiedano ciascuna T/(n/2) e le restanti 0 (che in termini di quote vale ad assumere che le prime n/2 quote presentino tutte eguale valore 1/(n/2) e le restanti zero), e ricaviamo in tale condizione i valori che assumerebbero i vari indici.

È come se stessimo ricavando dei valori intermedi di «riferimento» per i vari indici, a cui confrontare i valori reali forniti dagli indici. Per un qualsiasi valore reale x di un indice, con $x \in [0..1]$, ora potremo dire se x è minore o maggiore del valore di «riferimento», ovvero se il valore reale indica che la concentrazione reale è minore o maggiore di quanto si avrebbe se tutto l'ammontare delle quote derivate delle n determinazioni del caso in esame fosse «concentrato», in modo «equidistribuito», nelle prime n/2 determinazioni.

Tramite l'empirico approccio delineato, si potrà almeno verificare se un indice propone un valore reale di concentrazione maggiore di quanto l'indice varrebbe se la somma dei valori di tutte le sue n determinazioni fosse equidistribuita nelle prime n/2 determinazioni - che è lo stesso che dire che la somma delle n quote (somma sempre pari ad uno) viene equidistribuita nelle sole prime n/2 determinazioni. Quindi, come detto, dei valori di «riferimento» contro cui confrontare i valori reali degli indici. Osserviamo, inoltre, che il valore di riferimento di ciascun indice dipenderà esclusivamente dal numero delle determinazioni e dal modello di calcolo dell'indice. Che è come dire che in più problematiche totalmente diverse, ma nelle quali sono coinvolte sempre lo stesso numero n di determinazioni/unità statistiche, il valore di riferimento per ognuno dei 5 indici sarà lo stesso. Ad esempio il valore di riferimento di Herfindal in qualsiasi problematica che implichi n=6 determinazioni/unità statistiche sarà sempre pari a 0,2, in quanto la somma delle quote (che è sempre pari ad uno) viene sempre ad essere equidistribuita nelle prime 6/2 = 3 determinazioni.

Ricaviamo, quindi, i valori di «riferimento» per i vari indici, seguendo l'approccio proposto.

Herfindal

Se si equidistribuisce la somma totale T dei valori delle n determinazioni nelle prime n/2 determinazioni, ciò significa che ciascuna quota vale $s_i = [T/(n/2)]/T = 1/(n/2)$ per $i \in [1..n/2]$ e $s_i = 0$ per $i \in [n/2 + 1..n]$. Ora non è sempre detto che n sia pari, ovvero esattamente divisibile in due parti intere, quindi poniamoci, nel caso generale, nelle condizioni di equidistribuire il totale in $k=int(n/2)$ determinazioni. In tal caso le varie determinazioni valgono $s_i = [T/k]/T = 1/k$ per $i \in [1..k]$ e $s_i = 0$ per $i \in [k + 1..n]$. Ciò posto otteniamo per Herfindal, direttamente omettendo nella sommatoria i valori per $i \in [k+1..n]$ che forniscono un contributo nullo

$$Hrif = \sum_{i=1}^{k} s_i^2 = \sum_{i=1}^{k} (1/k)^2 = k(1/k)^2 = 1/k$$

Ricordiamo, infine, che Herfindal formalmente fornisce valori nel range [1/n..1], ed è rispetto a valori ottenuti entro tale range che vorremo confrontare il valore di «riferimento» che stiamo costruendo, quindi normalizziamo rispetto al range [1/n..1] per operare nel range [0..1], ottenendo il definitivo

$$\mathbf{H^*rif = (1/k - 1/n)/(1 - 1/n)}$$

con H*rif rappresentativo (valore intermedio di «riferimento») della concentrazione dell'ammontare totale equidistribuito in k = int(n/2) determinazioni delle n totali.

Hanna-Kay

Per Hanna-Kay procediamo come per il precedente indice, ponendoci direttamente nel caso generale di k = int(n/2) , ovvero $s_i = 1/k$ per $i \in [1..k]$ e $s_i = 0$ per $i \in [k + 1..n]$ ed utilizzando la formulazione di Hanna-Kay che produce risultati nel range[1/n..1]

$$HK^{+}rif = \sum_{i=1}^{k}(1/k^{\alpha})^{1/(\alpha\text{ -}1)} = \left(k/_{k^{\alpha}}\right)^{1/(\alpha\text{ -}1)} = k^{(1-\alpha/\alpha-1)} = 1/k^{-(1-\alpha/\alpha-1)} = 1/k$$

Otteniamo quindi il risultato che $HK^{+}rif$ è esattamente uguale a Hrif di Herfindal.

Quindi non procediamo oltre con la normalizzazione, perché già sappiamo che HK*rif risulterà uguale a H*rif , ovvero il valore di «riferimento», su stessa numerosità di determinazioni, sarà lo stesso sia per Herfindal che per Hanna-Kay, e ciò per qualsiasi valore di α.

Hall-Tideman

Anche per Hall-Tideman si ottiene un risultato simile al precedente. Infatti assumendo come sempre k = int(n/2), ed utilizzando anche in questo caso la formulazione di Hall-Tideman che produce risultati nel range [1/n..1], si ottiene

$$HT^{+}rif = 1/[2*\left(\sum_{i=1}^{k} i * 1/k\right) - 1] = 1/[(2/k)\left(\sum_{i=1}^{k} i\right) - 1] = 1/[2/k * k(k+1)/2 - 1] = 1/k$$

Ancora lo stesso risultato di di Herfindal ed Hanna-Kay. Quindi non procediamo oltre con i calcoli, in quanto sappiamo che basta calcolare, su stessa numerosità di detrminazioni, ad es. il valore di riferimento (normalizzato nel range [0..1]) di Herfindal, in quanto

H*rif = HK*rif = HT*rif

Horvat

Operiamo nel consueto modo anche per Horvat, utilizzando la classica formulazione che fornisce valori nel range $[(3n^2 - 3n + 1)/n^3 .. 1]$, e ponendoci nel consueto caso di k = int(n/2), ovvero $s_i = 1/k$ per $i \in [1..int(n/2)]$ e $s_i = 0$ per $i \in [int(n/2) + 1..n]$.

$$HVrif= \frac{1}{k} + \sum_{i=2}^{k}\left[\left(\left(1/_{k}\right)^{2}\right)\left(2 - 1/_{k}\right)\right] = \frac{1}{k} + \sum_{i=2}^{k}\left(\frac{2k-1}{k^3}\right) = \frac{1}{k} + (k\text{-}1)\left(\frac{2k-1}{k^3}\right) = \frac{1}{k} + \left(\frac{2k^2-k-2k+1}{k^3}\right) = \left(\frac{3k^2-3k+1}{k^3}\right)$$

e normalizzando, per operare nel range [0..1] si ottiene

$$\mathbf{HV^*rif = (HVrif - [(3n^2 - 3n + 1)/n^3])/(1 - [(3n^2 - 3n + 1)/n^3])}$$

Theil

Operiamo allo stesso modo per Theil, sempre ponendoci direttamente nel caso generale di k = int(n/2), con $s_i = 1/k$ per $i \in [1..int(n/2)]$ e $s_i = 0$ per $i \in [int(n/2) + 1..n]$, ed utilizzando quale formulazione di Theil quella che fornisce valori nel range [0..log(n)].

$$T^+rif = \sum_{i=1}^{k} \frac{1}{k} * log\left(\frac{\frac{1}{k}}{1/n}\right) = \sum_{i=1}^{k} \frac{1}{k} \left(log\left(\frac{1}{k}\right) - log\left(\frac{1}{n}\right)\right) =$$

$$k \frac{1}{k} (\log(\frac{1}{k}) - \log(\frac{1}{n})) = \log(\mathbf{\frac{1}{k}}) - \log(\mathbf{\frac{1}{n}})$$

e normalizzando per operare nel range [0..1]

$$\mathbf{T^*rif = T^+rif / log(n)}$$

In conclusione, per qualsiasi indice conosciamo perfettamente il significato dei valori degli estremi dell'intervallo (0 = equidistribuzione, 1= max concentrazione), ma ora, per ogni set di dati, possiamo anche calcolare un valore intermedio di «riferimento», ovvero [0… val «riferimento» …1], valore intermedio che sebbene talvolta numericamente diverso da indice a indice, ha però lo stesso «significato». Quando il valore di un indice è maggiore del valore di riferimento ciò è indicativo di una «accentuata concentrazione». Infatti in tale occorrenza è come se il valore dell'indice indicasse una concentrazione maggiore di quella che si avrebbe se tutto l'ammontare del carattere fosse detenuto, in modo equidistribuito, solamente dalle prime int(n/2) unità statistiche. Il valore dell'indice di «riferimento» può quindi essere assimilato ad una condizione di «oligopolio», cioè ad una situazione in cui, date n unità statistiche oggetto di indagine, int(n/2) unità si «accaparrano/detengono», dividendoselo in modo equidistribuito, tutto il mercato/carattere. *Quindi un valore dell'indice maggiore di quello di «riferimento» indica una concentrazione maggiore di quanto si avrebbe se ci si trovasse in una condizione di «oligopolio»*.

Si osservi, infine, che poiché il valore intermedio di «riferimento» è dipendente solo dalla numerosità n delle determinazioni, sarebbe persino possibile tabellare i valori di riferimento dei vari indici in funzione di n, senza operare il calcolo a run time.

Con quanto sopra abbiamo terminato la prima puntata, dedicata alla presentazione dei 5 indici di concentrazione (Herfindal, Hanna e Kay, Hall e Tidemann, Horvat, Theil), alle loro normalizzazioni nel range [0..1], ed all'aver individuato una tecnica per ricavare un valore intermedio di «riferimento» che nel seguito permetterà una lettura/comparazione ed analisi critica dei valori degli indici di grande semplicità e facile comprensibilità. Rimane da applicare tutto ciò a dei casi d'uso, e ciò verrà fatto nella prossima puntata, nella quale verranno proposti due esempi, per ciascun dei quali

verranno presentati i valori dei vari indici, i relativi valori di «riferimento», e le considerazioni che potranno essere desunte dai risultati che si otterranno.

Alla prossima.

APPENDICE

Kanna-Kay

HK di 'base' fornisce valori nel range $[n^{1-\alpha}..1]$, dipendente da n e da α. Tale range è assai poco fruibile, quindi per prima cosa cerchiamo di «ricondurlo» ad un più comprensibile range [n..1], in modo da eliminare la dipendenza da α. Per ricondurre ad un range del tipo [n..1] si opera una trasformazione tale da ricondurre ad n il limite inferiore nel range, ovvero una trasformazione del tipo

$$\sqrt[1-\alpha]{n^{1-\alpha}} = n$$

ed applicando tale modalità di trasformazione alla formulazione di 'base' di Hanna-Kay si ottiene

$$\sqrt[1-\alpha]{\sum_{i=1}^{n} s_i^{\alpha}} = \left(\sum_{i=1}^{n} s_i^{\alpha}\right)^{1/1-\alpha}$$

Se vi fosse una equidistribuzione di valori delle determinazioni, ovvero delle quote, $s_i = 1/n$, Hanna-Kay, dopo tale trasformazione, varrebbe

$$\left(\sum_{i=1}^{n}\left(\frac{1}{n}\right)^{\alpha}\right)^{\left(\frac{1}{1-\alpha}\right)} = \left\{n\left(\frac{1}{n}\right)^{\alpha}\right\}^{\left(\frac{1}{1-\alpha}\right)} = \{n * n^{-\alpha}\}^{\left(\frac{1}{1-\alpha}\right)} = (n^{1-\alpha})^{\left(\frac{1}{1-\alpha}\right)} = n$$

Viceversa, se vi fosse massima concentrazione, quindi una sola quota possedesse tutto l'ammontare $s_i = 1$, con tutte le altre quote uguali a zero, Hanna-Kay dopo tale trasformazione varrebbe

$$1^{\alpha\left(\frac{1}{1-\alpha}\right)} = 1$$

Tale range [n..1], benché migliore del precedente, non è ancora quello ottimale, e tra l'altro presenta l'estremo inferiore maggiore di quello superiore. Cerchiamo quindi di «ricondurlo» ad un più generale range [1/n..1], in cui l'estremo inferiore è minore del superiore.

Per ricondurre ad un range del tipo [1/n..1] si opera una trasformazione tale da calcolare il reciproco della funzione, ovvero

$$HK^{+} = {1}\Big/{\left(\sum_{i=1}^{n} s_i^{\alpha}\right)^{(1/1-\alpha)}} = \left(\sum_{i=1}^{n} s_i^{\alpha}\right)^{-(1/1-\alpha)} = \left(\sum_{i=1}^{n} s_i^{\alpha}\right)^{1/(\alpha-1)}$$

che fornisce valori nel range [1/n..1].

Per verificarlo osserviamo che in caso di equidistribuzione

$$\left(\sum_{i=1}^{n} 1/n^{\alpha}\right)^{1/(\alpha-1)} = n\left(\frac{1}{n^{\alpha}}\right)^{1/(\alpha-1)} = \left(n^{(1-\alpha)}\right)^{1/(\alpha-1)} = \left(1/n^{\alpha-1}\right)^{1/(\alpha-1)} = 1/n$$

e nel caso di massima concentrazione $(1^{\alpha})^{(1/\alpha-1)} = 1$.

La formulazione HK+ è la più nota ed utilizzata per Hanna-Kay. Però ancora presta il fianco ad operare in un range dipendente da n.

Hall-Tidemann

Il range[1.. (n+1)/2] è, ovviamente, assai poco fruibile, cerchiamo quindi, inizialmente di «ricondurci» ad un più generale range [1..n].

Per ricondurci ad un range [n..1] occorre che l'estremo superiore del range precedente, ovvero (n+1)/2, si trasformi in n. Per fare ciò occorre operare una trasformazione su tale estremo del tipo 2* [(n+1)/2] – 1 = n.

Modifichiamo quindi in tal senso la formulazione di 'base' di Hall-Tideman, ottenendo:

$$2\left(\sum_{i=1}^{n} i s_i\right) - 1$$

che ora fornisce valori in [1..n].

Infatti nel caso di equidistribuzione $2/n^{*}\left(\sum_{i=1}^{n} i\right) - 1 = 2/n^{*}\,[n(n+1)/2] - 1 = n$ e nel caso di massima concentrazione $2^{*}1 - 1 = 1$.

Come detto, sinora però questo indice fornisce una misura 'inversa' di concentrazione (infatti il valore di equidistribuzione si ha in presenza del valore massimo, n, del range), si vuole, viceversa, una misura 'diretta' di concentrazione, ed a tal fine si opera una trasformazione che calcoli il reciproco della funzione, affinché il valore di equidistribuzione si abbia per il valore 1/n e quello di max concentrazione, come di consueto, per 1/1 = 1, ovvero si operi nel range [1/n..1].

Operando la trasformazione descritta, ovvero il reciproco della funzione, si ottiene

$$H^+ = 1/[2(\sum_{i=1}^{n} is_i) - 1]$$

con HT^+ che ora fornisce valori nel range [1/n..1].

Horvat

Si può facilmente dimostrare che il valore min = $[(3n^2 – 3n + 1)/n^3]$. Infatti, nel caso di equidistribuzione, ovvero quando tutte le quote sono pari a 1/n, si ottiene

$$HV = 1/n + \sum_{i=2}^{n} n^2(2 – 1/n) = 1/n + (n-1)*1/n^2*(2 – 1/n) =$$
$$= 1/n + 1/n^2(n(2-1/n) – (2-1/n)) = 1/n + 1/n^2(2n - 1 -2 + 1/n) =$$
$$= 1/n + 2/n – 1/n^2 – 2/n^2 + 1/n^3 = (2n^2 – n – 2n + 1 + n^2)/n^3 =$$
$$= \mathbf{(3n^2 – 3n + 1)/n^3}$$

Theil

Cominciamo con il trasformare la formulazione di base dell'indice inizialmente sottraendo tale formulazione al valore di massima concentrazione, ovvero

$$T^+ = \log(n) - \sum_{i=1}^{n} s_i \log(1/s_i) = \log(n) + \sum_{i=1}^{n} s_i \log(s_i)$$

e tenendo presente che $\sum_{i=1}^{n} s_i = 1$, possiamo riscrivere quanto sopra come

$$T^+ = \sum_{i=1}^{n} s_i \log(n) + \sum_{i=1}^{n} s_i \log(s_i) = \sum_{i=1}^{n} s_i (\log(n) + \log(s_i)) =$$
$$= \sum_{i=1}^{n} s_i * \log(n * s_i) = \sum_{i=1}^{n} [s_i * \log\left(s_i / 1/n\right)]$$

che è la formulazione più nota ed usata dell'indice di Theil, la quale opera nel range [0..log(n)]. Infatti in caso di equidistribuzione $T^+ = n*(1/n)*\log(1) = 0$ e nel caso di massima concentrazione $T^+ = 1*\log(1/(1/n)) = \log(n)$.