Concentrazione di un carattere

Prendiamo in esame un carattere che corrisponda ad oggetti che siano trasferibili, ovvero che possano passare da una persona all'altra, per esempio 40 caramelle. Supponiamo di volerle dividere tra 5 bambini: Andrea, Baldo, Arturo, Luca e Marco. La divisione delle caramelle puo' avvenire in molti modi diversi, quello che ci interessa è studiare un modo per capire quanto il carattere sia concentrato o distribuito tra i diversi bambini. Vedremo tre situazioni, partiamo con la prima in cui le caramelle sono equamente distribuite:

i	X	n	f	F	$p = i/m$
1	Andrea	8	0,2	0,2	$0,2 = 1/5$
2	Baldo	8	0,2	0,4	$0,4 = 2/5$
3	Arturo	8	0,2	0,6	$0,6 = 3/5$
4	Luca	8	0,2	0,8	$0,8 = 4/5$
5	Marco	8	0,2	1	$1 = 5/5$
		40			

Nella tabella di lato, oltre alle classiche colonne, ne abbiamo aggiunta una, la colonna "p" che si calcola per ogni riga effettuando il rapporto tra il numero di riga ("i" da 1 a 5) e il numero totale di righe ($m = 5$).

Quello che si nota è che, se il carattere è equidistribuito, F e p hanno valori coincidenti

Mettiamoci ora in un secondo caso, quello in cui la distribuzione delle caramelle non sia uniforme. Ad esempio daremo a Baldo 3 caramelle, ad Andrea 5, ad Arturo 9, a Luca 11 e a Marco 12.
La prima cosa da capire è che in questo caso la tabella va costruita partendo dal carattere meno numeroso a quello più numeroso:

i	X	n	f	F	p
1	Baldo	3	0,075	0,075	0,2
2	Andrea	5	0,125	0,2	0,4
3	Arturo	9	0,225	0,425	0,6
4	Luca	11	0,275	0,7	0,8
5	Marco	12	0,3	1	1
		40	1		

In questo caso si nota che F e p sono diversi.

Si può dimostrare, ma non lo faremo, che, escluso l'ultimo valore, nel caso in cui non ci sia distribuzione equa, accade sempre che: $F_i < p_i$

In generale per tener presente del caso di distribuzione uniforme, nonché dell'ultimo valore, la relazione generale tra "F" e "p" è: $\boxed{F_i \leq p_i}$

Dunque in assenza di concentrazione $F_i = p_i$, mentre più c'è concentrazione del carattere più i termini F_i e p_i cominciano ad essere discosti.
Il caso limite di concentrazione è il seguente, in cui, per esempio, tutte le caramelle sono a Marco.
Nella tabella, Marco sarà l'ultimo degli

elementi elencati, visto che questi vanno in ordine:

i	X	n	f	F	p
1	Baldo	0	0	0	0.2
2	Andrea	0	0	0	0.4
3	Arturo	0	0	0	0.6
4	Luca	0	0	0	0.8
5	Marco	40	1	1	1

E' facile capire che in questo caso la differenza tra i valori di F e p assume il massimo valore possibile. Infatti considerando che l'ultimo valore deve essere uguale ad 1 sia per "p" che per "F", e che i valori di "p" sono sempre gli stessi, la massima differenza si ottiene nel caso in cui gli F (frequenze cumulate) siano tutte nulle.

Quanto visto finora ci convince che un buon indice per misurare la concentrazione deve essere proporzionale alla somma delle differenze tra i valori di "F" e di "p" (somma da cui si può escludere l'ultima riga, visto che già sappiamo che per essa $F - p = 0$).
L'indice in questione si chiama indice di Gini e si calcola come segue:

$$g = \frac{\sum_{i=1}^{m-1} (p_i - F_i)}{\sum_{i=1}^{m-1} p_i} =$$

la sommatoria al numeratore si può scindere in 2 sommatorie:

$$= \frac{\sum_{i=1}^{m-1} p_i - \sum_{i=1}^{m-1} F_i}{\sum_{i=1}^{m-1} p_i}$$

i due termini al numeratore si possono entrambi dividere per il denominatore, ottenendo:

$$= 1 - \frac{\sum_{i=1}^{m-1} F_i}{\sum_{i=1}^{m-1} p_i}$$

ora si utilizza la definizione di $p_i = \frac{i}{m}$

$$= 1 - \frac{\sum_{i=1}^{m-1} F_i}{\frac{1}{m}\sum_{i=1}^{m-1} i} =$$

al denominatore la "m" che è un numero si è portato fuori dalla sommatoria. Questa sommatoria è diventata ora la somma dei primi "m-1" numeri naturali, la quale si può dimostrare essere:

$$\sum_{i=1}^{m-1} i = \frac{m\cdot(m-1)}{2}$$

(formula di Gauss)

sostituendo:

$$= 1 - \frac{\sum_{i=1}^{m-1} F_i}{\frac{\cancel{m}\cdot(m-1)}{2\cancel{m}}} = \boxed{1 - \frac{2}{m-1}\sum_{i=1}^{m-1} F_i}$$

L'ultima formola ottenuta è quella che si usa nelle applicazioni. Applichiamola alle tre distribuzioni viste finora:

caso 1: carattere equidistribuito

$$g = 1 - \frac{2}{5-1}(0{,}2+0{,}4+0{,}6+0{,}8) =$$
$$= 1 - \frac{2}{4}\cdot 2 = 1 - \frac{4}{4} = 1 - 1 = 0$$

caso 2: carattere distribuito casualmente

$$g = 1 - \frac{2}{5-1}(0{,}075+0{,}2+0{,}425+0{,}7) =$$
$$= 1 - \frac{2}{4}\cdot 1{,}4 = 0{,}3$$

caso 3 : carattere con massima concentrazione

$$g = 1 - \frac{2}{5-1}\left(0+0+0+0\right) =$$

$$= 1 - \frac{2}{4} \cdot 0 = 1 - 0 = 1$$

L'indice di Gini, dunque, assume sempre un valore compreso tra zero ed uno:

$0 \leq g \leq 1$ dove $g = 0$ corrisponde al caso di concentrazione nulla

e $g = 1$ corrisponde al caso di massima concentrazione.