MatematicaMente

ISSN: 2037-6367

Pubblicazione mensile della sezione veronese della MATHESIS – Società Italiana di Scienze Matematiche e Fisiche – Fondata nel 1895 – Autorizzazione del Tribunale di Verona n. 1360 del 15 – 03 – 1999 – I diritti d'autore sono riservati. Direttore: Luciano Corso – Redazione: Alberto Burato, Fabrizio Giugni, Michele Picotti, Sisto Baldo, Andrea Sellaroli, Bruno Stecca – Via IV Novembre, 11/b – 37126 Verona – 338 6416432 – e-mail: – info@mathesisverona.it, lcorso@iol.it – Stampa in proprio – Numero 237 – Pubblicato il 07 - 06 - 2018

Entropia come misura dell'eterogeneità [1]

di Luciano Corso [*]

Abstract: *In this paper, I consider the entropy concept by Claude Shannon as a measure of the heterogeneity. In statistics, this measure is important to verify the degree of dispersion of a population of data in a given space. In this case, we use the entropy as measure of the dispersion of a polluting substance in a certain environment.*

Il presente lavoro nasce da un corso di Calcolo delle Probabilità e di Statistica che ho tenuto presso l'ARPA del Veneto nel 2007. Il corso era indirizzato ai funzionari del settore scientifico e consisteva in un ciclo annuale di lezioni (90 ore) sui più significativi argomenti della Statistica. Un argomento importante della Statistica descrittiva è la variabilità, o dispersione, di un gruppo di dati sperimentali. Ho affrontato, in questo ambito, anche il tema dell'entropia come misura dell'eterogeneità. Ho ritenuto che fosse necessario sostenere la teoria con un'applicazione simulata, un po' fuori dagli schemi classici, per mezzo della quale i tecnici potessero rendersi conto di come applicare il concetto di entropia per misurare la dispersione sul territorio di sostanze inquinanti. I corsisti (personale già assunto dall'ARPAV), una quarantina, erano tutti laureati con lauree a orientamento scientifico tecnologico e provenivano da ogni parte del Veneto. Il corso si tenne presso la sede dell'ARPAV di Verona e venne fortemente sostenuto dall'allora dirigente responsabile Attilio Tacconi (dottore in Fisica).

Il concetto di entropia si studia in termodinamica. Il 2° principio della termodinamica è legato, com'è noto, alle strutture profonde delle trasformazioni termodinamiche ed è il fondamento per comprendere la dispersione caotica del calore (energia) nello spazio. Fu Boltzmann a intuire bene il fenomeno. Poi, Claude Shannon comprese che il concetto poteva essere esteso nel suo significato anche per capire come un messaggio, nel passare da una sorgente a un punto di ricezione, potesse arrivare lacunoso. La relazione di entropia di Shannon corrisponde perfettamente a quella di Boltzmann, com'è stato dimostrato da più autori. Shannon, tuttavia, dette al concetto una valenza molto ampia, tanto da poterlo oggi considerare anche una misura dell'eterogeneità statistica. Peraltro, l'entropia, come misura dell'eterogeneità, non ha avuto un grande successo applicativo, probabilmente perché si preferiscono altre misure molto più semplici da usare, efficaci ed efficienti come questa. Per esempio, l'indice di eterogeneità di Corrado Gini arriva alle stesse conclusioni con calcoli meno complessi. Tuttavia l'entropia, come misura, ha un rilievo semantico particolare per le implicazioni che l'idea assume ogni qual volta si prendono in considerazione trasformazioni fisiche, chimiche, biologiche, ecologiche ed economiche. L'applicazione del concetto di entropia di Shannon che qui presento permette al lettore di capire come si usa la misura e quali processi devono essere attuati per renderla operativa. È possibile considerare il presente lavoro come unità didattica relativa alle misure di dispersione statistica, sotto la voce «eterogeneità».

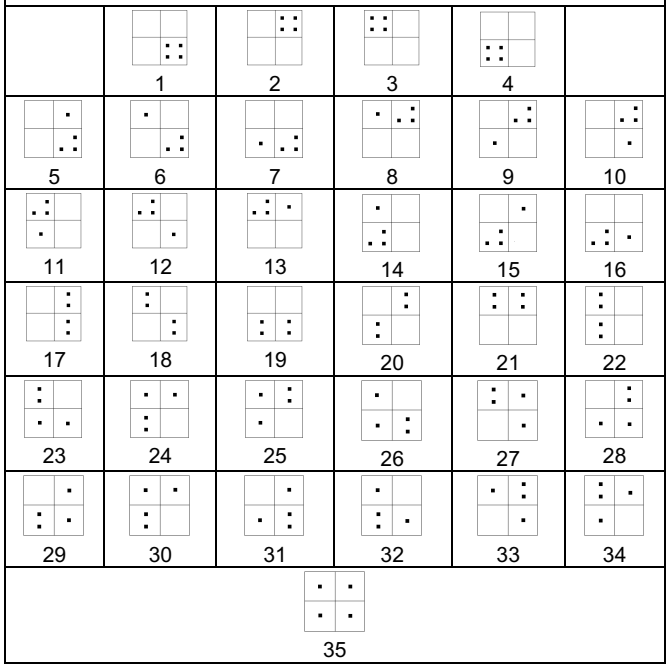
Concludo questa breve introduzione con un'avvertenza: sia i disegni, sia le tabelle sono stati costruiti ad hoc, allo scopo di meglio chiarire il processo applicativo usato e la semantica associata a ogni formula.

Consideriamo $n = 4$ particelle d'inquinante in un reticolo territoriale quadrato suddiviso in $k = 4$ parti (parcelle) di una partizione. Ci si chiede in quanti modi le 4 particelle possono occupare queste 4 parti di territorio. Il calcolo combinatorio ci dà la possibilità di conoscere quante soluzioni ci possono essere. Esse sono:

$$\binom{n+k-1}{k} = \binom{4+4-1}{4} = 35\,. \qquad (1)$$

Nella Tabella A, sono presentate tutte le possibilità.

Tabella A: con riferimento all'esempio, sono 35 gli stati distinti possibili



L'entropia media, come misura dell'eterogeneità di uno stato è data da:

$$H_s = \frac{1}{Ln(2)} \cdot \sum_{i=1}^{k} p_i \cdot Ln\left(\frac{1}{p_i}\right) \qquad \forall\, s \qquad (2)$$

dove s =1, …, 35 sta per stato considerato, k è il numero degli elementi eterogenei presenti in ogni stato (numero di modalità, che corrisponde, nel nostro caso, al numero di celle di ogni stato), p_i è la probabilità di trovarsi in presenza dell'elemento (cella) i. Per esempio, per lo stato 34 l'applicazione della (2) dà:

$$H_{34} = \left[\frac{1}{4} \cdot Ln(4) + \frac{2}{4} \cdot Ln(2) + \frac{1}{4} \cdot Ln(4)\right] \cdot \frac{1}{Ln(2)} = 1{,}5.$$

Per giustificare (2) si consideri un sistema di k eventi A_1 con frequenza n_1, A_2 con frequenza n_2, … A_k con frequenza n_k. In questa popolazione, per ogni s, si ha:

$$\sum_{i=1}^{k} n_i = n\,, \quad \frac{n}{n_1} = 2^{h_1}\,, \quad \frac{n}{n_2} = 2^{h_2}\,, \ldots.\,, \quad \frac{n}{n_k} = 2^{h_k}\,,$$

dove h_i è l'entropia associata all'evento A_i, cioè è il numero mini-

mo di domande *sì-no* che si devono fare per sapere se si è verificato l'evento A_i tra i k eventi distinti possibili. Nel nostro caso, dando per noto che, a priori, si conoscano le dislocazioni delle particelle nelle varie parti, l'interpretazione è la seguente: scelta a caso una particella da una delle parcelle, per gli stati fino a 4, le domande *sì-no* da fare sono zero in quanto, se pesco a caso una particella da una delle parcelle, non ho da fare alcuna domanda per sapere da quale parcella è stata estratta, nel senso che è stata estratta di sicuro dalla parcella che ne contiene 4. Se, invece, sono in uno degli stati da 5 a 16, allora dato che so, sempre a priori, in quali parti sono dislocate le particelle, la particella estratta può provenire da una delle due parti e quindi basta una sola domanda *sì-no* per sapere da quale. In più, in tal caso, poiché una parcella ha più probabilità di essere stata scelta dell'altra, l'entropia è lievemente minore di uno perché si dà più credito al fatto di aver pescato dalla parcella che contiene tre particelle rispetto all'aver pescato in quella che ne contiene una sola. Se, come nei casi dal 17 al 22, vi fossero due palline in ogni parte che le contengono e zero nelle altre due, allora l'entropia è uguale a uno, cioè con una sola domanda *sì-no* si riesce a selezionare la parcella delle due da cui è stata estratta la particella. E così via. Considerando che $n_i / n = p_i$ e passando poi al logaritmo naturale e alla media si ottiene

$$M(h) = \sum_{i=1}^{k} h_i \cdot p_i = \frac{1}{Ln(2)} \cdot \sum_{i=1}^{k} p_i \cdot Ln\left(\frac{n}{n_i}\right) =$$

$$= \frac{1}{Ln(2)} \cdot \sum_{i=1}^{k} p_i \cdot Ln\left(\frac{1}{p_i}\right).$$

Teniamo presente che questa misura non distingue tra stati che hanno una distribuzione simile di mutazioni di stato. Perciò gli stati dal 23 al 34, per esempio, hanno entropia uguale. Per ciascuno di questi 35 stati possibili calcoliamo l'entropia come misura dell'eterogeneità di ogni stato.

Se i gruppi contenenti i particelle hanno numerosità k_i, allora i modi di porre n particelle in k parti sono $k! / (k_0! \ldots k_n!)$, dove k è il numero totale di parti (celle) del sistema, $k_0, k_1, \ldots, k_n$ contano rispettivamente le parti con uguale presenza di particelle i (cioè nelle k_i celle ci sono i particelle). Inoltre, $\sum k_i = k$. Con riferimento allo stato s =1 (4 particelle tutte in un'unica cella, la quarta) si ha n=4, k=4, k_0=3, k_1=0, k_2=0, k_3=0, k_4=1 e quindi 4! / (3! 0! 0! 0! 1!) = 4.

Ecco i 5 casi:

1) 4! / (3! 0! 0! 0! 1!) = 4
2) 4! / (2! 1! 0! 1! 0!) =12
3) 4! / (2! 0! 2! 0! 0!) = 6
4) 4! / (1! 2! 1! 0! 0!) =12
5) 4! / (0! 4! 0! 0! 0!) = 1.

L'entropia H_s ha un massimo dato da: $Max(H_s)=Log_2 k$. Dividendo H_s per il suo massimo si ottiene una misura normalizzata dell'entropia. Assegniamo a questa misura il simbolo H'_s. Si ha:

$$H'_s = \frac{H_s}{Max(H_s)}. \qquad (3)$$

La (2) e la (3) portano a:

$$H'_s = \frac{1}{Ln(2)} \cdot \frac{\sum_{j=1}^{k} p_j \cdot Ln\left(\frac{1}{p_j}\right)}{Log_2(k)}. \qquad (4)$$

Presentiamo gli stati sotto forma di tabelle in cui per ogni casella A_{ij} è presentata la corrispondente frequenza osservata di particelle inquinanti:

Stato 1 (da 1 a 4)		
Parcelle	n_i	p_i
A_{11}	0	0
A_{12}	0	0
A_{21}	0	0
A_{22}	4	1

$$H_1 = \frac{1}{Ln(2)} \cdot \sum_{i=1}^{4} p_i \cdot Ln\left(\frac{1}{p_i}\right) = 0,$$

$$H'_1 = \frac{H_1}{Max[H_1]} = 0.$$

Stato 2 (da 5 a 16)		
Parcelle	n_i	p_i
A_{11}	0	0
A_{12}	0	0
A_{21}	1	1/4
A_{22}	3	3/4

$$H_2 = \frac{1}{Ln(2)} \cdot \sum_{i=1}^{4} p_i \cdot Ln\left(\frac{1}{p_i}\right)$$

$$= \frac{1}{Ln(2)} \cdot \left[\frac{1}{4} \cdot Ln(4) + \frac{3}{4} \cdot Ln\left(\frac{4}{3}\right)\right] \cong 0{,}8113,$$

$$H'_1 = \frac{H_1}{Max[H_1]} \cong \frac{0{,}8113}{2} \cong 0{,}4057.$$

Stato 3 (da 17 a 22)		
Parcelle	n_i	p_i
A_{11}	0	0
A_{12}	0	0
A_{21}	2	1/2
A_{22}	2	1/2

$$H_3 = \frac{1}{Ln(2)} \cdot \sum_{i=1}^{4} p_i \cdot Ln\left(\frac{1}{p_i}\right)$$

$$= \frac{1}{Ln(2)} \cdot \left[\frac{2}{4} \cdot Ln(2) + \frac{2}{4} \cdot Ln(2)\right] = 1,$$

$$H'_3 = \frac{1}{2} = 0.5.$$

Stato 4 (da 23 a 34)		
Parcelle	n_i	p_i
A_{11}	0	0
A_{12}	1	1/4
A_{21}	1	1/4
A_{22}	2	2/4

$$H_4 = \frac{1}{Ln(2)} \cdot \sum_{i=1}^{4} p_i \cdot Ln\left(\frac{1}{p_i}\right)$$

$$= \frac{1}{Ln(2)} \cdot \left[\frac{1}{4} \cdot Ln(4) + \frac{1}{4} \cdot Ln(4) + \frac{2}{4} \cdot Ln(2)\right] = 1.5,$$

$$H'_4 = \frac{1.5}{2} = 0.75.$$

Stato 5 (situazione 35)		
Parcelle	n_i	p_i
A_{11}	1	1/4
A_{12}	1	1/4
A_{21}	1	1/4
A_{22}	1	1/4

$$H_5 = \frac{1}{Ln(2)} \cdot \sum_{i=1}^{4} p_i \cdot Ln\left(\frac{1}{p_i}\right)$$

$$= \frac{1}{Ln(2)} \cdot \left[\frac{1}{4} \cdot Ln(4) + \frac{1}{4} \cdot Ln(4) + \frac{1}{4} \cdot Ln(4) + \frac{1}{4} \cdot Ln(4)\right] = 2,$$

$$H'_5 = \frac{2}{2} = 1.$$

[*Segue al numero* 238]

[1] Il presente lavoro è già stato pubblicato dall'autore sul Periodico di Matematiche, ISSN-1582-8832, ed. Mathesis, n. 2 Mag-Ago 2011 Volume 3 Serie XI Anno CXXI, pagg. 97-106.

[*] Presidente della sezione di Verona della Mathesis, fondatore e direttore della rivista MatematicaMente.