CALCOLO DELLA DIFFERENZA MEDIA.

(In collaborazione con U. Paciello)

In: « *Metron* », 1930, vol. VIII, n. 3, pp. 1-6.

B. DE FINETTI E U. PACIELLO

Calcolo della differenza media.

Uno dei metodi introdotti nella metodologia statistica dal GINI (*) per misurare la variabilità di una seriazione consiste nel calcolo della *differenza media* (con o senza ripetizione, Δ_R e Δ).

Vari metodi sono stati proposti per eseguirne il calcolo.

1) Il GINI stesso, nel lavoro citato, diede un metodo che si basa sulla considerazione delle *distanze graduali* di due elementi simmetrici (che occupano cioè lo stesso posto nella graduatoria rispettivamente in ordine crescente e decrescente). La sua formula è

$$\Delta_R = \frac{n-1}{n}\Delta = \frac{1}{n^2}\sum_1^n{}_i \, d_{i,n-i+1}\,|a_i - a_{n-i+1}|$$

ove $d_{i,n-i+1} = n + 1 - 2i$ è la distanza graduale.

2) Lo CZUBER ha proposto un metodo che utilizza le successive somme parziali dei termini disposti in ordine crescente e decrescente. È da questo metodo che si deduce la disposizione di calcolo data dal PIETRA (**), per il caso cosiddetto della « differenza media ponderata », che qui in particolar modo c'interessa. Un'abbreviazione al calcolo è stata portata recentemente dal DE GLERIA (***), grazie

---

(*) CORRADO GINI, *Variabilità e Mutabilità*, « Studi economico-giuridici pubblicati per cura della facoltà di Giurisprudenza della R. Università di Cagliari », Anno III, Parte 2ª, 1912.

(**) GAETANO PIETRA, *The theory of statistical relations with special reference to cyclical series*, « Metron », Vol. IV, N. 3-4, 1925.

(***) AMADIO DE GLERIA, *Una abbreviazione nel calcolo della differenza media*, « Rivista Italiana di Statistica », Anno I, N. 4, 1929.

all'osservazione, per se stessa ovvia, che le due serie di somme sono complementari.

La formula di CZUBER è

$$\Delta_R = \frac{n - 1}{n} \Delta = \frac{2}{n^2} (S' - S)$$

dove

$$S = s_1 + s_2 + \ldots + s_n \qquad s_i = a_1 + a_2 + \ldots + a_i$$
$$S' = s'_1 + s'_2 + \ldots + s'_n \qquad s'_i = a_n + a_{n-1} + \ldots + a_{n-i+1}$$

3) Il GINI ha introdotto poi il *rapporto di concentrazione* R (*), e ha dimostrato che è collegato alla differenza media dalla relazione

$$\Delta_R = \frac{n - 1}{n} \Delta = 2 A R$$

dove A è la media aritmetica delle a_i. Un metodo per il calcolo della differenza media consiste dunque nel dedurla dal rapporto di concentrazione, che si può calcolare con metodi aritmetici od anche grafici, illustrati nel lavoro citato di GINI.

Dovendo però eseguire spesso tali calcoli, specialmente nel caso della « differenza media ponderata », per cui il procedimento preferibile era quello (nominato) del PIETRA, abbiamo veduto che essi riescono molto lunghi e faticosi, e abbiamo cercato quindi di procedere in modo più conveniente e che si presti meglio all'impostazione meccanica. Ne è risultato il metodo che qui esponiamo, e che, impiegato molte volte, ha mostrato di consentire un risparmio enorme di tempo e di lavoro.

1. — Detta

$$a_1 \quad a_2 \quad \ldots \quad a_n \qquad (a_1 \leq a_2 \leq \ldots \leq a_n)$$

una seriazione statistica, si dicono differenza media con ripetizione, Δ_R, e differenza media senza ripetizione, Δ, le espressioni

$$\Delta_R = \frac{\sum_{1}^{n} {}_{ij} |a_i - a_j|}{n^2} \qquad \Delta = \frac{\sum_{1}^{n} {}_{ij} |a_i - a_j|}{n (n - 1)}$$

(*) CORRADO GINI, *Sulla misura della concentrazione e della variabilità dei caratteri*, « Atti del R. Ist. Veneto di S. L. A. », Anno 1913-14, Tomo LXXV, Parte II.

per il cui calcolo, ovviamente, non interessa in sostanza che valutare la sommatoria

$$\sum_{1}^{n}{}_{ij} |a_i - a_j|.$$

Possiamo scrivere intanto

$$\sum_{1}^{n}{}_{ij} |a_i - a_j| = 2 \sum_{1}^{n}{}'_{ij} (a_i - a_j)$$

ove Σ' è esteso alle sole coppie con $i > j$. Ogni termine di tale somma si può scomporre:

$$a_i - a_j = (a_i - a_{i-1}) + (a_{i-1} - a_{i-2}) + \dots + (a_{j+2} - a_{j+1}) + (a_{j+1} - a_j)$$

ed è ovvio quindi che si può scrivere

$$\sum_{1}^{n}{}'_{ij} (a_i - a_j) = \sum_{1}^{n-1}{}_{h} C_h (a_{h+1} - a_h)$$

ove C_h è il numero dei termini del tipo $(a_i - a_j)$ che contengono l'addendo $(a_{h+1} - a_h)$, ossia per cui $i \geq h + 1$, $j \leq h$. Essendo $(n - h)$ i valori di i maggiori o uguali ad $(h + 1)$, e h i valori di j minori o uguali ad h, risulta

$$C_h = h\ (n - h),$$

$$\sum_{1}^{n}{}'_{ij} (a_i - a_j) = \sum_{1}^{n-1}{}_{h} h\ (n - h)\ (a_{h+1} - a_h).$$

2. — Il metodo risulta particolarmente vantaggioso quando molte delle quantità a_i hanno lo stesso valore (o si attribuisce loro lo stesso valore raggruppandole in classi più o meno estese), perchè allora la più parte dei termini è nulla.

Consideriamo infatti il caso in cui la seriazione è costituita degli n valori

$$a_1 \quad a_2 \quad \dots \quad a_n \qquad\qquad (a_1 < a_2 < \dots < a_n)$$

assunti rispettivamente in

$$p_1 \quad p_2 \quad \dots \quad p_n$$

casi. Oppure, ciò che non cambia nulla, supponiamo che p_1 casi siano raggruppati in una classe cui compete il valore a_1, p_2 in una classe cui compete il valore $a_2, \dots$ e si voglia calcolare la differenza media

come se nei $p_1, p_2, \ldots, p_n$ casi il carattere a assumesse effettivamente il valore $a_1, a_2, \ldots, a_n$.

Posto allora

$$q_h = p_1 + p_2 + \ldots + p_h \qquad\qquad q = q_n$$

la formula precedente diviene

$$\sum_{1}^{n}{}'_{ij} (a_i - a_j) = \sum_{1}^{n-1}{}_h \; q_h \, (q - q_h) \, (a_{h+1} - a_h)$$

e si presta al calcolo in modo facilissimo, come mostra l'esempio che segue.

3. — Vogliamo calcolare la differenza media della natalità nel 1921 per i 16 Compartimenti, tenendo conto dell'ammontare delle rispettive popolazioni.

Poniamo allora :

h = numero d'ordine dei Compartimenti disposti per natalità crescente,

a_h = natalità nell'h-esimo Compartimento (anno 1921),

p_h = popolazione dell'h-esimo Compartimento (cens. 1° dic. 1921),

e, al solito

$q_h = p_1 + p_2 + \ldots + p_h =$ popolazione dei Compartimenti con natalità non superiore ad a_h,

$q = q_{16} =$ popolazione censita totale,

e quindi

$q - q_h =$ popolazione dei Compartimenti con natalità non inferiore ad a_{h+1}.

Nella tabella seguente le somme S_h sono riportate in migliaia ; il calcolo è stato eseguito tenendo conto di tutte le cifre, e ciò spiega delle piccole differenze che si incontrerebbero ripetendolo. Va notato che, pur tenendo conto di tutte le cifre, il calcolo si è potuto eseguire sulla Nova Brunsviga a (10, 10, 15) cifre, mentre col metodo precedente sarebbe stato necessario ricorrere a macchine con un numero di cifre maggiore.

Ecco la tabella dei dati e dei calcoli che danno la sommatoria

$$\sum_{1}^{n}{}'_{ij} (a_i - a_j).$$

h	Compartimenti	a_h	p_h	q_h	$q - q_h$	$a_{h+1} - a_h$	$q_h (q - q_h) \times (a_{h+1} - a_h)$
1	Piemonte	20,22	3.383.646	3.384	33.801	0,82	93.783.718
2	Liguria	21,04	1.335.466	4.719	32.465	6,31	966.743.552
3	Sicilia	27,35	4.061.452	8.781	28.404	0,81	202.016.640
4	Lombardia . . .	28,16	5.086.338	13.867	23.318	0,40	129.337.566
5	Toscana	28,56	2.759.767	16.627	20.558	1,14	389.662.881
6	Lazio	29,70	1.956.908	18.584	18.601	0,62	214.317.350
7	Sardegna	30,32	864.174	19.447	17.737	0,28	96.583.586
8	Emilia	30,60	3.027.009	22.474	14.710	1,97	651.285.855
9	Marche	32,57	1.145.685	23.620	13.564	1,05	336.410.397
10	Umbria	33,62	645.515	24.266	12.919	0,43	134.797.550
11	Campania . . .	34,05	3.243.739	27.510	9.675	0,36	95.815.077
12	Veneto	34,41	3.999.027	31.509	5.676	0,91	162.743.370
13	Abruzzi e Molise	35,32	1.387.215	32.896	4.289	0,57	80.414.886
14	Puglie	35,89	2.307.762	35.204	1.981	0,34	23.709.606
15	Calabrie	36,23	1.512.318	36.716	469	1,65	28.385.855
16	Basilicata. . . .	37,88	468.557	37.184	—	—	—
			37.184.578				3.606.007.889

Si ricava

$$\sum_{1}^{n}{}'_{i\,j} (a_i - a_j) = 3.606.007.889.000.000 :$$

vanno aggiunti 6 zeri perchè la popolazione si è valutata in migliaia (mancano 3 cifre a q_h e $q - q_h$), La popolazione totale è 37.184.578, e quindi

$$\Delta_R = 2 \cdot \frac{3.606.007.889.000.000}{(37.184.578)^2} = 5{,}216$$

$$\Delta = 2 \cdot \frac{3.606.007.889.000.000}{37.184.578 \times 37.184.577} = 5{,}216.$$

È ovvio che su un numero di termini così grande la differenza di un'unità non influisce sulle cifre decimali che possono interessare, e si ha quindi praticamente $\Delta = \Delta_R$.

4. — Una semplificazione notevole si ha quando tutte le differenze $(a_{h+1} - a_h)$ sono uguali, come avviene in molti casi. Sono infatti molto usate delle classificazioni in gradi di uguale ampiezza: le età, le stature, i pesi di un gruppo d'individui si ripartiranno ad es. per classi ciascuna d'un anno, d'un centimetro, d'un chilogrammo.

Allora si ha

$$\sum_{1}^{n-1}{}_h \; q_h \, (q - q_h) \, (a_{h+1} - a_h) = d \cdot \sum_{1}^{n-1}{}_h \; q_h \, (q - q_h)$$

ove

$$d = a_n - a_{n-1} = a_{n-1} - a_{n-2} = \ldots = a_2 - a_1$$

è l'ampiezza d'ogni singola classe.