

Introduzione di Statistica descrittiva

Silvia Bacci¹

Dipartimento di Economia - Università di Perugia

¹silvia.bacci@unipg.it

Outline

- 1 Introduzione
- 2 Distribuzioni di frequenza
 - Le frequenze assolute
 - Le frequenze relative
 - Le frequenze cumulate
- 3 Le misure di centralità
 - La media aritmetica
 - La mediana e altri quantili
 - La moda
- 4 Le misure di variabilità
 - Il caso dei caratteri quantitativi
 - Il caso dei caratteri qualitativi

Che cosa è la Statistica?

- La **Statistica** è la disciplina che studia i **fenomeni collettivi**
 - I **fenomeni collettivi** sono tutti quei fenomeni che presentano una **pluralità** di manifestazioni **diverse**
- La **Statistica** è l'insieme delle metodologie per raccogliere, organizzare, sintetizzare, analizzare e interpretare le manifestazioni dei fenomeni e trasformarle in **informazioni** utili per i **processi decisionali**

Alcune domande sulle prove Invalsi alle quali può rispondere la Statistica

- Qual è il punteggio medio conseguito nel test di matematica da parte della totalità degli studenti? E distinguendo per sesso e per scuola di provenienza?
- La variabilità dei punteggi dei test Invalsi è maggiore nel test di matematica o nel test di italiano? Ed è maggiore tra gli studenti del Sud, del Centro o del Nord Italia?
- I risultati conseguiti nel test di matematica sono tendenzialmente migliori o peggiori rispetto ai risultati conseguiti nel test di italiano?
- È vero che il fenomeno del “cheating” è più diffuso nelle scuole del Sud Italia rispetto a quelle del Centro-Nord?
- Qual è la probabilità che uno studente maschio ottenga un punteggio nel test di italiano maggiore di una studentessa?
- Esiste una relazione tra punteggio conseguito in un test Invalsi e scuola di provenienza?

Alcuni concetti di base

- **Carattere, variabile**: sono sinonimi del termine fenomeno
- **Unità statistica**: è il singolo soggetto (individuo, studente) sul quale si osserva il fenomeno studiato
- **Popolazione**: è l'insieme di tutte le unità statistiche oggetto dello studio; di norma non è possibile né tantomeno economicamente e temporalmente conveniente osservare la popolazione nella sua interezza
- **Campione**: è un sottoinsieme rappresentativo della popolazione

Alcuni concetti di base

- **Studio osservativo**: i dati vengono raccolti tramite l'osservazione delle variabili d'interesse sulle unità statistiche, senza nessun intervento atto a modificarne i valori
- **Esperimento pianificato**: le unità statistiche vengono prima sottoposte ad un certo trattamento e poi vengono osservate
- **Parametro**: è una caratteristica, usualmente incognita, della popolazione (es., livello medio delle conoscenze in matematica degli studenti italiani di terza media)
- **Statistica**: è una caratteristica, usualmente nota, del campione (es., punteggio medio conseguito nel test di Matematica dagli studenti del campione)
- **Dati**: sono le manifestazioni del fenomeno oggetto di studio; usualmente sono riportati in una matrice con le unità statistiche sulle righe e le variabili sulle colonne

Dati utilizzati

- Dati Invalsi rilevati a giugno 2009 sugli studenti di terza media
- Sono considerate solo le scuole/classi con almeno 20 studenti partecipanti alla rilevazione
- 3774 studenti provenienti da 171 scuole
- Test di italiano: 30 domande di comprensione del testo e 10 domande di grammatica
- Test di matematica: 27 domande
- Tutte le domande sono a scelta multipla con 1 risposta corretta e 3 sbagliate; la codifica usata è 1 per la risposta corretta e 0 per quella sbagliata
- Informazioni aggiuntive disponibili: sesso dello studente, scuola di provenienza, area geografica sede della scuola

Statistica descrittiva e Inferenza statistica

- La **statistica descrittiva** è quell'insieme di metodologie statistiche che si occupano della descrizione, cioè dell'organizzazione e della sintesi, dei dati raccolti (provenienti dalla popolazione o dal campione)
 - Le informazioni tratte dal processo di organizzazione e sintesi consentono di pervenire ad una conoscenza del fenomeno studiato che è **limitata all'insieme di dati raccolti**
- L'**inferenza statistica** è quell'insieme di metodologie statistiche che consentono di **generalizzare all'intera popolazione** la conoscenza acquisita sui dati osservati su un campione
 - **Processo induttivo** di acquisizione della conoscenza $\neq$ Processo deduttivo
 - Le conclusioni inferenziali sono sempre accompagnate da una certa **probabilità di errore** dovuta alle fluttuazioni casuali nel campione (errore campionario vs errore non campionario)
 - Premessa fondamentale dell'inferenza: **Teoria delle Probabilità**

La natura delle variabili

- **Variabili qualitative o categoriali:** sono variabili le cui manifestazioni sono di natura non numerica
 - **Variabili nominali o sconnesse:** si manifestano con due o più modalità (o categorie) qualitative non ordinabili (es., scuola frequentata, sesso, area geografica di provenienza)
 - **Variabili ordinali:** le modalità (o categorie) sono intrinsecamente ordinabili (es., livelli di soddisfazione definiti come “per niente”, “abbastanza”, “molto”, livelli di abilità definiti come “insufficiente”, “sufficiente”, “buono”, “ottimo”)
- **Variabili quantitative:** sono variabili che si manifestano tramite conteggi o misure
 - **Variabili discrete:** possono assumere un numero finito o un'infinità numerabile di valori (es., punteggio conseguito al test di matematica: 0, 1, 2, ..., 27)
 - **Variabili continue:** assumono un'infinità non numerabile di valori (es., età, peso, altezza, temperatura, distanze)

Le frequenze assolute

- Punto di partenza: dati disaggregati

modalità	x_1	x_1	x_2	x_2	x_2	x_3	x_4	...	x_i	...	x_k
frequenza	1	1	1	1	1	1	1	...	1	...	1

- La **frequenza assoluta** n_i indica il **numero di volte** in cui una certa modalità x_i di X è stata osservata

Table : Distribuzione delle frequenze assolute

modalità	x_1	x_2	x_3	x_4	...	x_i	...	x_k
freq. ass.	2	3	1	1	...	n_i	...	n_k

Table : Distribuzione degli studenti in base all'area geografica di provenienza (frequenze assolute)

Area x_i	n_i
Nord-Est	566
Nord-Ovest	448
Centro	626
Sud	1218
Isole	916
Totale	3774

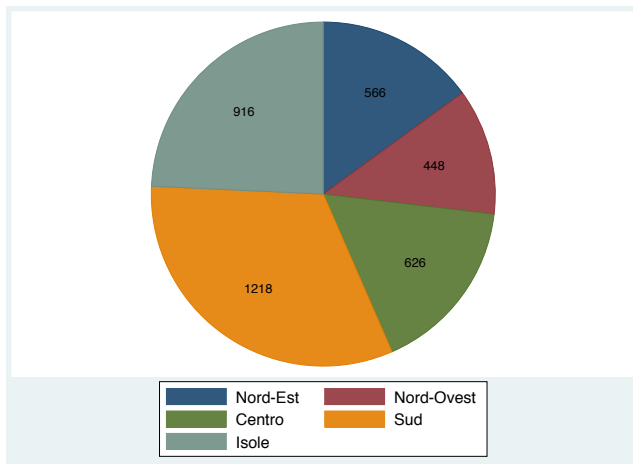
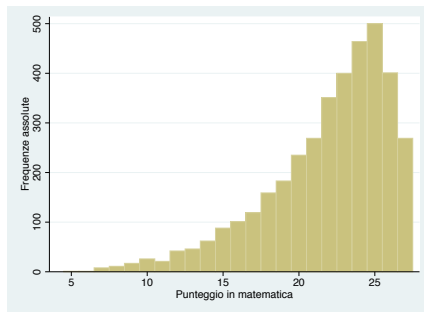
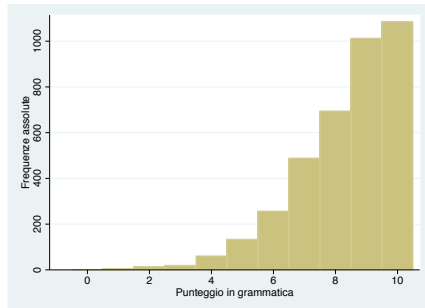
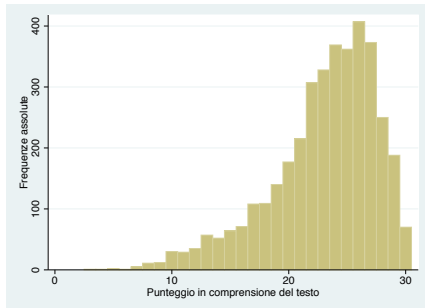


Figure : Distribuzione degli studenti in base all'area geografica di provenienza (frequenze assolute)

Table : Test Invalsi di italiano: distribuzione del punteggio in comprensione del testo (frequenze assolute)

Punteggio x_i	n_i	Punteggio x_i	n_i
3	1	17	108
4	1	18	109
5	2	19	140
6	1	20	177
7	5	21	215
8	11	22	307
9	12	23	328
10	30	24	369
11	29	25	362
12	35	26	407
13	57	27	373
14	52	28	250
15	64	29	188
16	71	30	70
		Totale	3774



Dati in classi

- Se un carattere è quantitativo è spesso necessario o utile **aggregare i dati in classi**

Table : Test Invalsi di italiano: distribuzione del punteggio in comprensione del testo, dati in classi (frequenze assolute)

i	Classi di punteggio	n_i
1	0-4	2
2	5-9	31
3	10-14	203
4	15-19	492
5	20-24	1396
6	25-30	1650
Totale		3774

Le frequenze relative

- La **frequenza relativa** f_i indica il numero di volte in cui una certa modalità x_i di X è stata osservata rapportata al numero totale di osservazioni

$$f_i = \frac{n_i}{n},$$

$$\text{con } n = \sum_{i=1}^k n_i$$

Table : Test Invalsi di italiano: distribuzione del punteggio in comprensione del testo, dati in classi (frequenze assolute e relative)

i	Classi di punteggio	n_i	f_i
1	0-4	2	0,0005
2	5-9	31	0,0082
3	10-14	203	0,0538
4	15-19	492	0,1304
5	20-24	1396	0,3699
6	25-30	1650	0,4372
Totale		3774	1,000

Le frequenze cumulate

- Se i dati sono **almeno ordinabili**, un'altra distribuzione utile è quella delle **frequenze cumulate**
- La **frequenza assoluta cumulata** N_i indica il numero di unità che presentano la modalità x_i o una modalità inferiore ad x_i (o che si trovano nella classe i -esima o in una classe inferiore)

$$N_i = \sum_{j=1}^i n_j$$

- La **frequenza relativa cumulata** F_i indica il numero relativo di unità che presentano la modalità x_i o una modalità inferiore ad x_i (o che si trovano nella classe i -esima o in una classe inferiore)

$$F_i = \sum_{j=1}^i f_j = \frac{N_i}{n}$$

Table : Test Invalsi di italiano: distribuzione del punteggio in comprensione del testo, dati in classi (frequenze assolute e relative, semplici e cumulate)

i	Classi di punteggio	n_i	f_i	N_i	F_i
1	0-4	2	0,0005	2	0,0005
2	5-9	31	0,0082	33	0,0087
3	10-14	203	0,0538	236	0,0625
4	15-19	492	0,1304	728	0,1929
5	20-24	1396	0,3699	2124	0,5628
6	25-30	1650	0,4372	3774	1,0000
Totale		3774	1,000		

- Ben il 43,72% degli studenti ($1 - 0,5628 = 0,4372$) ha conseguito un punteggio sulla sezione di comprensione del testo del test di italiano pari ad almeno 25/30

Le misure di centralità

- Le misure di centralità sono degli indicatori sintetici che descrivono ciò che di caratteristico c'è in una distribuzione di frequenza
- Le più note misure di centralità sono:
 - media aritmetica
 - mediana e altri quantili
 - moda

La media aritmetica

- Media aritmetica nel caso di dati disaggregati:

$$\mu = \frac{\sum_{i=1}^n x_i}{n}$$

- Media aritmetica nel caso di distribuzioni di frequenza:

$$\mu = \frac{\sum_{i=1}^k x_i n_i}{n} = \sum_{i=1}^k x_i f_i$$

- Media aritmetica nel caso di dati in classi:

$$\mu = \frac{\sum_{i=1}^k x_i^* n_i}{n} = \sum_{i=1}^k x_i^* f_i,$$

con $x_i^* = \frac{c_i + c_{i+1}}{2}$ valore centrale della classe i -esima

Caratteristiche della media aritmetica:

- Può essere calcolata soltanto per **caratteri quantitativi** (es., punteggio medio osservato nel test Invalsi di matematica)
- Il suo valore è sempre compreso tra il minimo e il massimo dei valori osservati
- $\sum_{i=1}^n (x_i - \mu) = 0$, cioè scostamenti positivi e scostamenti negativi dei singoli valori osservati rispetto alla media si compensano tra loro
- $\sum_{i=1}^n (x_i - H)^2$ è minima se e solo se $H = \mu$: tale proprietà risulta utile per definire una idonea misura di variabilità (vedi oltre)
- **Risente di valori anomali (outliers)** presenti nella distribuzione, cioè valori eccezionalmente grandi o piccoli

Table : Test Invalsi di italiano e matematica: punteggi medi per sesso e area geografica

	Compr. testo	Grammatica	Matematica
Maschi	22,624	8,237	21,998
Femmine	23,359	8,454	21,810
Nord-Est	23,067	8,150	21,564
Nord-Ovest	22,136	7,855	20,739
Centro	22,831	8,051	21,244
Sud	23,454	8,732	22,481
Isole	22,869	8,396	22,365
Totale	22,994	8,346	21,903

Nell'interpretare la tabella:

- è necessario tenere conto che il massimo punteggio conseguibile varia in base al test (30 per comprensione del testo, 10 per grammatica e 27 per matematica);
- ha senso confrontare i punteggi medi degli studenti di sesso diverso o provenienti da aree geografiche diverse per un dato test; non ha senso il confronto tra test diversi

La mediana

- La mediana è il **valore centrale** di una distribuzione ordinata, cioè è il valore che divide a metà una distribuzione
- Può essere calcolata solo su dati ordinabili, quindi ha un'applicabilità più ampia della media
- Essa **non risente di valori anomali**, quindi in certi casi è da preferire alla media aritmetica
- Come si calcola:
 - nel caso di **distribuzione disaggregata**:
 - se n è dispari, la mediana è il valore che occupa la posizione $(n + 1)/2$
 - se n è pari, la mediana è il valore dato dalla semisomma dei valori che occupano le posizioni $n/2$ e $n/2 + 1$
 - nel caso di **distribuzione aggregata**, la mediana è il valore al quale corrisponde la frequenza relativa cumulata F_i immediatamente maggiore di 0,50

- Si considerino i seguenti punteggi ordinati riportati nel test di matematica dai 20 studenti provenienti da una determinata scuola del Centro Italia:

5 12 13 16 16 17 17 18 19 19 19 20 20 20 20 22 23 25 25 25

- Calcolare media e mediana dei punteggi
 - Punteggio medio: $\mu_1 = 18,550$
 - Punteggio mediano: $\tilde{\mu}_1 = 19,000$
- Calcolare nuovamente media e mediana, eliminando il punteggio più basso
 - Punteggio medio: $\mu_2 = 19,263$
 - Punteggio mediano: $\tilde{\mu}_2 = 19,000$
- Eliminando il punteggio "5", che è particolarmente basso rispetto a tutti gli altri punteggi osservati, la mediana rimane invariata, mentre la media aritmetica aumenta di 0,713 punti

I quantili

- I **quantili** sono valori che dividono in 4 parti uguali una distribuzione ordinata
- Il **primo quartile** Q_1 è tale che almeno il 25% dei dati è uguale o inferiore a Q_1 e almeno il 75% dei dati è maggiore o uguale a Q_1
- Il **secondo quartile** Q_2 è tale che almeno il 50% dei dati è uguale o inferiore a Q_2 e almeno il 50% dei dati è maggiore o uguale a Q_2
- Il secondo quartile coincide con la mediana
- Il **terzo quartile** Q_3 è tale che almeno il 75% dei dati è uguale o inferiore a Q_3 e almeno il 25% dei dati è maggiore o uguale a Q_3

- I **percentili** sono valori che dividono in 100 parti uguali una distribuzione ordinata
- Il **k -esimo percentile P_k** è tale che almeno il $k\%$ dei dati è uguale o inferiore a P_k e almeno il $(100 - k)\%$ dei dati è maggiore o uguale a P_k
- Nota che Q_1 coincide con P_{25} , Q_2 coincide con P_{50} e Q_3 coincide con P_{75}
- Come calcolare il k -esimo percentile:
 - calcoliamo $L = k/100 \cdot n$
 - se L non è un numero intero $\Rightarrow P_k$ è il valore in posizione $[L + 1]$ (cioè arrotondo L per eccesso)
 - se L è un numero intero $\Rightarrow P_k$ è la semisomma tra i valori in posizione L e $L + 1$

- Si considerino i seguenti punteggi ordinati riportati nel test di matematica dai 20 studenti provenienti da una determinata scuola del Centro Italia:

5 12 13 16 16 17 17 18 19 19 19 20 20 20 20 22 23 25 25 25

- Calcolare Q_1 , Q_3 e P_{10}
 - Primo quartile: $L = 25/100 \cdot 20 = 5 \Rightarrow Q_1$ è la semisomma tra i valori in posizione 5 e 6, cioè: $Q_1 = 16,5$
 - Terzo quartile: $L = 75/100 \cdot 20 = 15 \Rightarrow Q_3$ è la semisomma tra i valori in posizione 15 e 16, cioè: $Q_3 = 21,0$
 - Decimo centile: $L = 10/100 \cdot 20 = 2 \Rightarrow P_{10}$ è la semisomma tra i valori in posizione 2 e 3, cioè: $P_{10} = 12,5$

Table : Test Invalsi di italiano, sezione di comprensione del testo: punteggio medio e quantili (decimo centile, primo quartile, mediana, terzo quartile) per sesso

Sesso	μ	P_{10}	Q_1	$\tilde{\mu}$	Q_3
Maschi	22,624	16,000	20,000	24,000	26,000
Femmine	23,359	17,000	21,000	24,000	27,000
Totale	22,994	17,000	21,000	24,000	26,000

La moda

- La moda è la **modalità con frequenza più elevata**
- Può essere calcolata su dati di qualsiasi natura, quindi ha **un'applicabilità più ampia** della media e della mediana
- È **meno informativa** di media e mediana:
 - problema di distribuzioni multimodali
 - problema di distribuzioni con modalità aventi frequenze simili

x_i	5	12	13	16	17	18	19	20	22	23	25
n_i	1	1	1	2	2	1	3	4	1	1	3

Le misure di variabilità

- Le misure di variabilità misurano **il grado di dispersione** di un carattere intorno ad una certa misura di centralità
- Le più note misure di variabilità sono:
 - **range**: differenza tra il valore massimo e il valore minimo
 - **range interquartile**: $Q_3 - Q_1$
 - **varianza** e **deviazione standard** (o scostamento quadratico medio)
 - **coefficiente di variazione**

La varianza

- La **varianza** è una misura della variazione dei valori intorno alla propria media aritmetica
- È la media degli scarti al quadrato rispetto alla media aritmetica; il quadrato è necessario per evitare che scarti positivi e negativi si compensino

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n} = \frac{\sum_{i=1}^k (x_i - \mu)^2 n_i}{n}$$

- La varianza assume sempre valori **maggiori o uguali a 0**
- La varianza è pari a 0 quando le unità assumono tutte la stessa modalità del carattere
- L'unità di misura della varianza è il quadrato dell'unità di misura del carattere osservato

La deviazione standard

- La **deviazione standard** è la radice quadrata della varianza e si indica con σ
- Essa ha il vantaggio di essere espressa nella **stessa unità di misura** del carattere osservato

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}} = \sqrt{\frac{\sum_{i=1}^k (x_i - \mu)^2 n_i}{n}}$$

Varianza e deviazione standard

Table : Test Invalsi di italiano e matematica: punteggi medi e deviazioni standard per sesso e area geografica

	Compr. testo		Grammatica		Matematica	
	μ	σ	μ	σ	μ	σ
Maschi	22,624	4,695	8,237	1,680	21,998	4,064
Femmine	23,359	4,384	8,454	1,553	21,810	4,025
Nord-Est	23,067	4,188	8,150	1,642	21,564	4,172
Nord-Ovest	22,136	4,756	7,855	1,587	20,739	4,602
Centro	22,831	4,466	8,051	1,723	21,244	4,176
Sud	23,454	4,426	8,732	1,437	22,481	3,735
Isole	22,869	4,828	8,396	1,661	22,365	3,770
Totale	22,994	4,555	8,346	1,621	21,903	4,045

Nell'interpretare la tabella:

- è necessario tenere conto che il massimo punteggio conseguibile varia in base al test (30 per comprensione del testo, 10 per grammatica e 27 per matematica);
- ha senso confrontare - tramite la deviazione standard (o la varianza) - i punteggi medi degli studenti di sesso diverso o provenienti da aree geografiche diverse per un dato test; non ha senso il confronto tra test diversi

Il coefficiente di variazione

- L'assenza di un limite superiore e la dipendenza dall'unità di misura fa sì che varianza e deviazione standard non siano utilizzabili per il **confronto di variabilità tra popolazioni diverse**
- Ad esempio, si pensi al confronto tra la variabilità del punteggio del test Invalsi di italiano sezione di comprensione del testo, la variabilità del punteggio del test Invalsi di italiano sezione di grammatica e la variabilità del punteggio del test Invalsi di matematica
- Una misura relativa di variabilità rappresentata da un numero puro è data dal **coefficiente di variazione CV**

$$CV = \frac{\sigma}{|\mu|} \cdot 100$$

sotto la condizione che $\mu \neq 0$

Table : Test Invalsi di italiano e matematica: deviazioni standard e coefficienti di variazione per sesso e area geografica

	Compr. testo		Grammatica		Matematica	
	σ	CV	σ	CV	σ	CV
Maschi	4,695	20,750	1,680	20,397	4,064	18,477
Femmine	4,384	18,767	1,553	18,372	4,025	18,454
Nord-Est	4,188	18,157	1,642	20,153	4,172	19,348
Nord-Ovest	4,756	21,485	1,587	20,202	4,602	22,192
Centro	4,466	19,563	1,723	21,406	4,176	19,659
Sud	4,426	18,871	1,437	16,462	3,735	16,615
Isole	4,828	21,113	1,661	19,778	3,770	16,858
Totale	4,555	19,809	1,621	19,421	4,045	18,468

- **Interpretazione errata, basata sui valori di σ :** il test di italiano sezione di grammatica presenta una variabilità nei punteggi inferiore (sia globalmente sia distinguendo per sesso e per area geografica) rispetto ai test di italiano sezione di comprensione del testo e di matematica; il test di italiano sezione di comprensione del testo presenta una variabilità nei punteggi leggermente superiore al test di matematica
- **Interpretazione corretta, basata sui valori di CV :** la variabilità dei punteggi è molto simile nei tre test posti a confronto; nel Nord-Est e nel Centro Italia i punteggi del test di italiano sezione di grammatica presentano una variabilità leggermente superiore rispetto ai punteggi degli altri due test

Eterogeneità

- Nel caso di caratteri qualitativi, si parla di **eterogeneità** (anziché di variabilità) del carattere
- **Minima eterogeneità** (o massima omogeneità): quando tutte le unità del collettivo hanno la stessa modalità del carattere (es., tutti gli studenti sono maschi)

modalità	x_1	x_2	x_3	x_4	...	x_i	...	x_k
freq. rel.	0	0	0	0	...	1	...	0

- **Massima eterogeneità** (o minima omogeneità): quando tutte le modalità del carattere presentano la medesima frequenza relativa nel collettivo esaminato (es., il 50% degli studenti è maschio e il 50% degli studenti è femmina)

modalità	x_1	x_2	x_3	x_4	...	x_i	...	x_k
freq. rel.	$\frac{1}{k}$	$\frac{1}{k}$	$\frac{1}{k}$	$\frac{1}{k}$	...	$\frac{1}{k}$	...	$\frac{1}{k}$

Indici di eterogeneità

- Indice relativo di eterogeneità di Gini

$$e_1 = \frac{k}{k-1} \left(1 - \sum_{i=1}^k f_i^2 \right)$$

- Indice relativo di entropia

$$e_2 = \frac{-\sum_{i=1}^k f_i \log(f_i)}{\log(k)}$$

- Entrambi gli indici assumono **valore tra 0 (minima eterogeneità) ed 1 (massima eterogeneità)**

Table : Tabella per il calcolo degli indici di eterogeneità e_1 ed e_2 relativi alla distribuzione degli studenti per area geografica

Area	n_i	f_i	f_i^2	$f_i \log(f_i)$
Nord-Est	566	0,150	0,022	-0,285
Nord-Ovest	448	0,119	0,014	-0,253
Centro	626	0,166	0,028	-0,298
Sud	1218	0,323	0,104	-0,365
Isole	916	0,243	0,059	-0,344
	3774	1,000	1,000	0,000
Somma			0,227	-1,544

- Indice relativo di eterogeneità di Gini

$$e_1 = \frac{5}{4}(1 - 0,227) = 0,966$$

- Indice relativo di entropia

$$e_2 = \frac{1,544}{\log(5)} = 0,959$$

- Entrambi gli indici indicano un'elevata eterogeneità nella distribuzione degli studenti per area geografica

Table : Frequenze relative f_i e indici di eterogeneità e_1 ed e_2 relativi alla distribuzione delle risposte corrette per i 10 item della sezione di Grammatica del test Invalsi

Item	f_i	e_1	e_2
c1	0,936	0,241	0,345
c2	0,784	0,677	0,752
c3	0,857	0,489	0,591
c4	0,653	0,906	0,931
c5	0,921	0,292	0,399
c6	0,728	0,793	0,845
c7	0,789	0,666	0,744
c8	0,949	0,195	0,292
c9	0,964	0,139	0,224
c10	0,766	0,717	0,785