Statistica I

Esercitazione 4: dati qualitativi, eterogeneità, concentrazione

Tommaso Rigon

Università Milano-Bicocca

Descrizione del problema

- La **litrotripsia extracorporea** è un trattamento relativamente poco invasivo per il paziente per la calcolosi.
- Si avvale dell'utilizzo del litotritore, un'apposita apparecchiatura che genera onde d'urto capaci di frammentare i calcoli.
- Per valutarne l'efficiacia, nel caso della calcolosi uretrale, sono stati considerati in quest'indagine un totale di $n = 80$ **pazienti**.
- L'**efficacia** è stata misurata utilizzando la seguente **scala di modalità** che si riferisce al grado di frammentazione dei calcoli dopo la prima seduta di trattamento:
  - `Buono`: tutti i frammenti sono più piccoli di 3mm.
  - `Medio`: nessun frammento sopra i 5mm, almeno uno maggiore di 3mm.
  - `Scarso`: frammenti maggiori di 5mm.
  - `Assente`: nessun segno di frammentazione deicalcoli originari.
- Per ogni paziente è inoltre nota la posizione dell'uretere (`lombare`, `presacrale` o `pelvico`), ovvero la posizione in cui si erano formati i calcoli.

Dati grezzi e domande

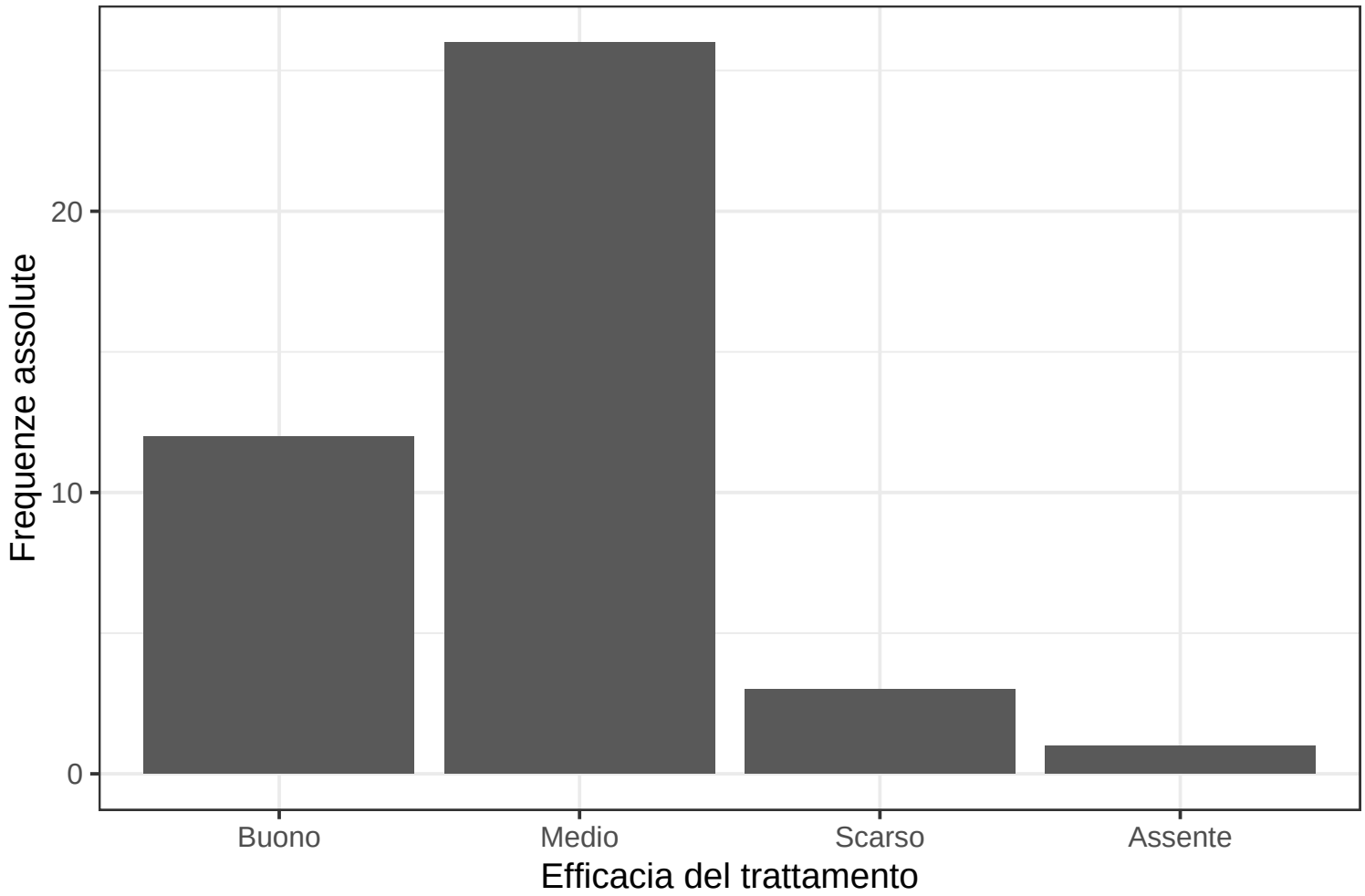
- I dati sono riassunti nella seguente **tabella**:

	buono	medio	scarso	assente
lombare	12	26	3	1
pre-sacrale	2	8	0	0
pelvico	12	13	2	1

Domande

- Qual è la **moda** della variabile "efficacia del trattamento" per i pazienti con calcolosi lombare? Si produca un **grafico** a supporto della risposta.
- Si ottengano gli **indici di eterogeneità** (normalizzati) di Gini e l'entropia, nei tre casi.
- Quale uretere è **meno variabile** in termini di efficacia del trattamento?

Diagramma a barre



Schema della soluzione

- Gli indici richiesti sono indicati nella tabella seguente. I **calcoli** sono stati **omessi**.

	`lombare`	`pre-sacrale`	`pelvico`
G_{norm}	0.71	0.43	0.79
H_{norm}	0.67	0.36	0.74

- Pertanto, gli indici normalizzati indicano una maggiore variabilità della risposta per le sedi `lombare` e `pelvica` rispetto a `pre-sacrale`.

Commenti ai risultati

- Il diagramma a barre mostra chiaramente una **buona/media efficacia** della **terapia**.
- Gli indici di **eterogeneità complementano la descrizione** del problema, misurando, zona per zona, la variabilità del fenomeno, ovvero l'**affidabilità** del metodo.
- Sebbene abbiano eterogeneità simili, l'**efficacia del trattamento** nella zona `lombare` sembra essere maggiore rispetto al caso `pelvico`.
- Inoltre, il **numero di osservazioni** è davvero **molto limitato**, specialmente nel caso `pre-sacrale`. Questo potrebbe portare a delle **oscillazioni casuali** di questi indici.
- Immaginate di osservare una nuova coppia di osservazioni: (`scarso`, `pre-sacrale`). Questo nuovo dato porterebbe a $G_{\text{norm}} = 0.57$, ben superiore a 0.43!
- **Aggregare le classi**. Cosa succederebbe se aggregassimo i valori delle categorie buona e media? L'eterogeneità diminuirebbe moltissimo!

Interpretazione degli indici di eterogeneità

- Più in generale, come si **interpreta** il valore dell'indice di eteregeneità di Gini/Shannon?
- Si potrebbe essere portati a pensare, **erroneamente**, che $G_{\text{norm}} > 0.5$ implica alta eterogeneità, mentre mentre $G_{\text{norm}} < 0.5$ implica bassa eterogeneità.
- Purtroppo, non è così: l'interpretazione, così come giudizi "alto" e "basso", sono molto **difficili** proprio perchè legati al **contesto applicativo**.
- Per esempio, il valore $G_{\text{norm}} = 0.9993$ potrebbe essere cosiderato "basso" in ecologia, se fa riferimento alla biodiversità delle specie dell'intera Amazzonia.
- In mancanza di esperienza nel contesto applicativo, pone meno difficoltà il **confronto** tra **gruppi di dati** relativi allo stesso fenomeno, come in questo esercizio.

Descrizione del problema

- Consideriamo i dati relativi alle **quote di mercato** dei principali operatore di rete mobile in Italia nel 2023. Fonte: relazione annuale AGCOM 2024, Grafico 1.1.13.

Descrizione del problema

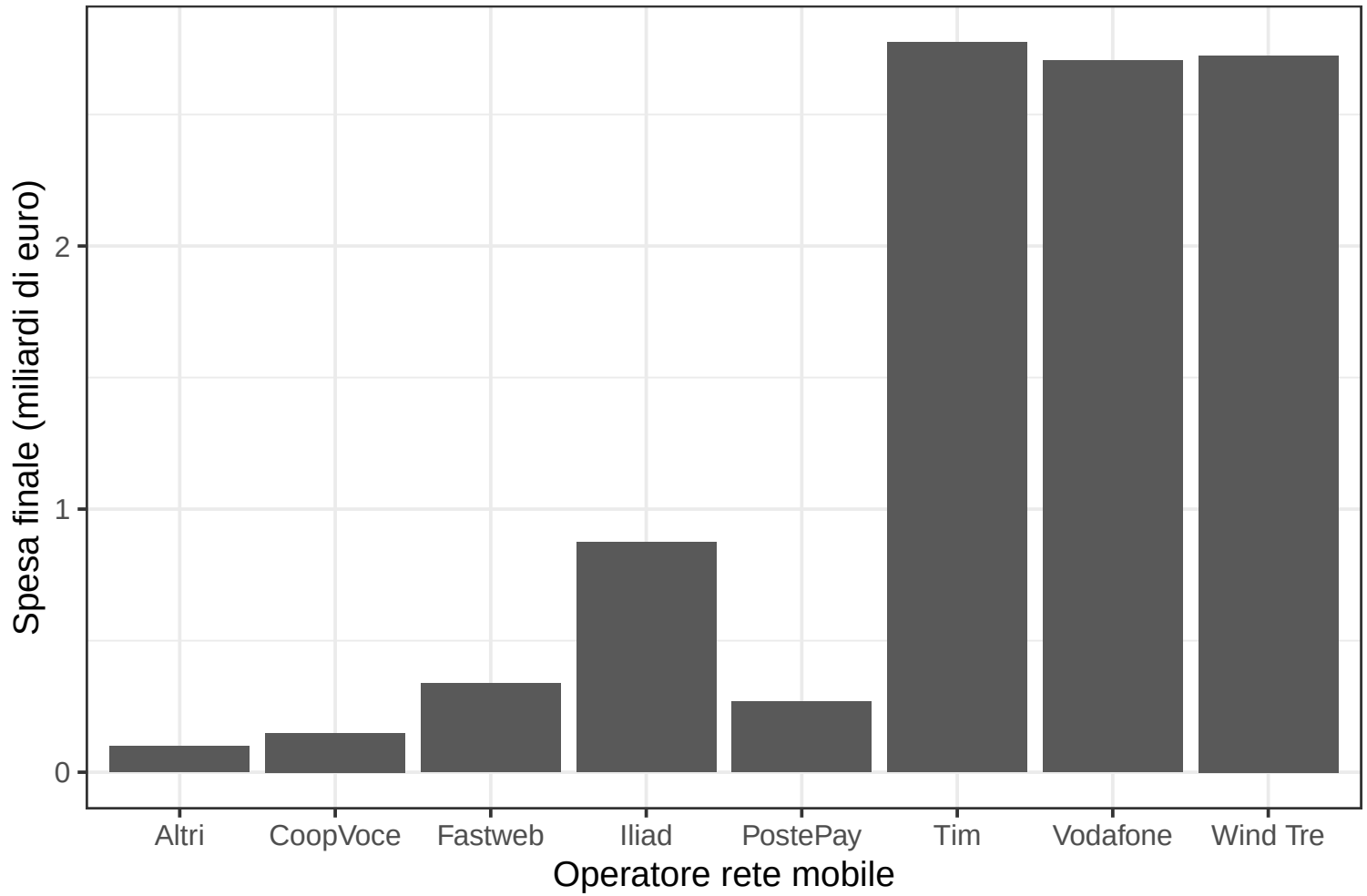
- In relazione a un mercato, gli indici di **eterogeneità**/**concentrazione** possono essere pensati come una misura del **grado di concorrenza** nel mercato stesso.

Operatore	Spesa finale (miliardi di euro)	Quota percentuale (%)
Tim	2.774	27.928
Wind Tre	2.724	27.427
Vodafone	2.704	27.227
Iliad	0.875	8.809
Fastweb	0.338	3.403
PostePay	0.268	2.703
CoopVoce	0.149	1.502
Altri	0.099	1.001

Domande

Si valuti il grado di concorrenza con appositi indici di **eterogeneità** e/o di **concentrazione**.

Diagramma a barre



Eterogeneità

- Le **quote** $f_1, \ldots, f_k$, ovvero rapporti tra la spesa totale di ciascun operatore e la spesa complessiva, possono essere trattate come se fossero delle **frequenze relative**.

Operatore	f_j	f_j^2	$\log f_j$	$f_j \log f_j$
Tim	0.2793	0.0780	-1.2755	-0.3562
Wind Tre	0.2743	0.0752	-1.2936	-0.3548
Vodafone	0.2723	0.0741	-1.3010	-0.3542
Iliad	0.0881	0.0078	-2.4294	-0.2140
Fastweb	0.0340	0.0012	-3.3804	-0.1150
PostePay	0.0270	0.0007	-3.6109	-0.0976
CoopVoce	0.0150	0.0002	-4.1987	-0.0630
Altri	0.0100	0.0001	-4.6042	-0.0461

- L'indice di **eterogeneità Gini** e l'indice di Gini normalizzato sono pari a

$$G = 1 - \sum_{j=1}^{k} f_j^2 = 1 - 0.237 = 0.763, \qquad G_{\text{norm}} = \frac{k}{k-1} G = 8/7 \times 0.763 = 0.87.$$

- L'**entropia** e l'entropia normalizzata sono invece pari a:

$$H = -\sum_{j=1}^{k} f_j \log f_j = 1.601, \qquad H_{\text{norm}} = H / \log k = 1.601 / \log 8 = 0.77.$$

Rapporto di concentrazione di Gini I

- Una possibilità alternativa consiste nel valutare la concorrenza del mercato tramite il **rapporto di concentrazione di Gini**.

- Si ottenga anzitutto la tabella contenente le coppie (p_j, q_j), ordinate per quota. In questo contesto, si ha che

$$p_j = \frac{j}{n}, \qquad q_j = \sum_{i=1}^{j} f_{(i)}, \qquad j = 1, \ldots, n,$$

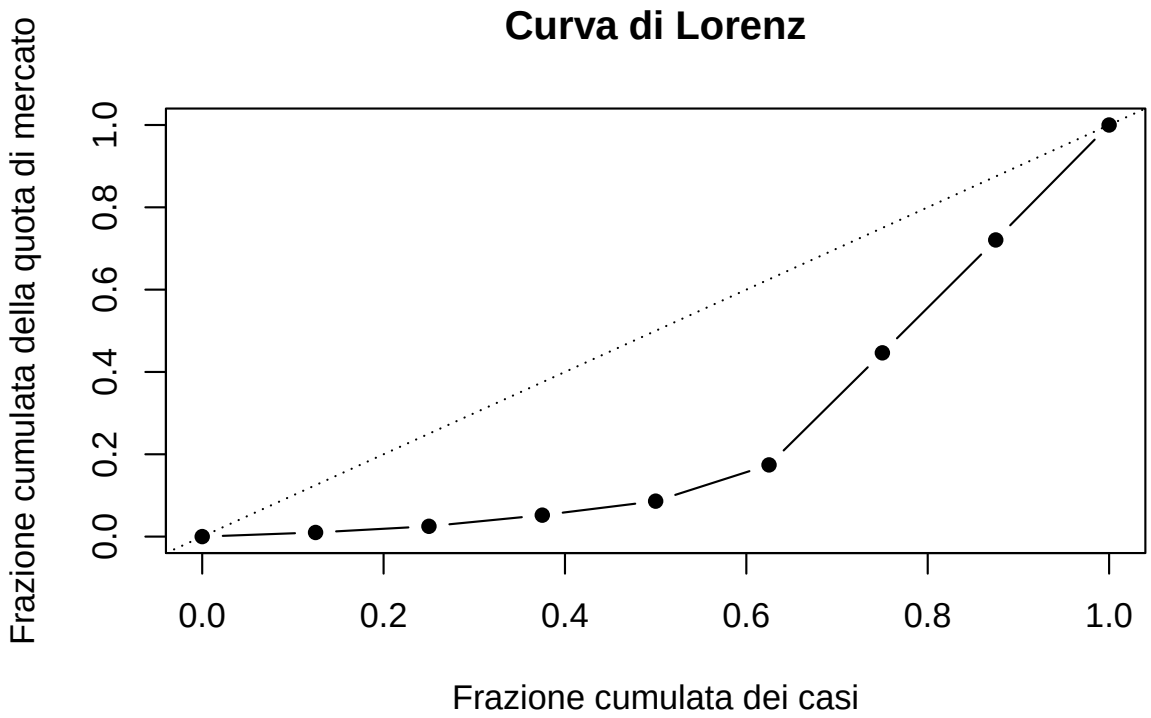
dove $f_{(1)}, \ldots, f_{(n)}$ sono le **quote di mercato ordinate**.

Operatore	$f_{(j)}$	p_j	q_j
Altri	0.0100	0.1250	0.0100
CoopVoce	0.0150	0.2500	0.0250
PostePay	0.0270	0.3750	0.0521
Fastweb	0.0340	0.5000	0.0861
Iliad	0.0881	0.6250	0.1742
Vodafone	0.2723	0.7500	0.4464
Wind Tre	0.2743	0.8750	0.7207
Tim	0.2793	1.0000	1.0000

Rapporto di concentrazione di Gini II

- Dalla precedente tabella il calcolo del **rapporto di concentrazione di Gini** è agevole, così come il grafico della curva di Lorenz:

$$\mathcal{R} = 1 - \frac{2}{n-1}\sum_{j=1}^{n-1} q_j = 1 - \frac{2}{7}(0.01 + 0.025 + 0.0521 + \cdots + 0.7207) = 0.5673.$$



Commento ai risultati

- Sia il report di AGCOM che il diagramma a barre, indicano la presenza di un **oligopolio** nel mercato di telefonia mobile.
- Tim, WindTre e Vodafone controllano circa l'80% del mercato. Questo suggerisce che in questo contesto applicativo, $G_{\text{norm}} = 0.87$ non è un valore "alto", anzi.
- Analogamente, l'indice di concentrazione $\mathcal{R} = 0.5673$, che non va considerato come un valore "basso".
- Entrambi gli indici sono quindi **informativi** solo **se ben contestualizzati**. Sono inoltre utili per descrivere la concorrenza del mercato nel tempo; in pratica, sarebbe interessante **confrontare** $\mathcal{R}_{2023}$ con $\mathcal{R}_{2022}, \mathcal{R}_{2021}$, etc.
- **Effetti distorsivi**. In questi dati alcuni operatori sono stati accorpati (ed anche le loro quote!). Sebbene non sia un grosso problema in questo caso specifico, dato che gli operatori accorpati sono pochi, questo **riduce artificialmente la concentrazione**.