APPROFONDIMENTI – SCHEDA 1

1. Terminologia

Il problema della terminologia in statistica descrittiva è significativo.
In queste schede abbiamo adottato la terminologia statistica più diffusa a livello scientifico internazionale (tradotta, ovviamente, in italiano).

A. Alcuni libri di statistica descrittiva di scuola italiana utilizzano una terminologia diversa. Forniamo quindi alcune "traduzioni"

Variabile qualitativa	Mutabile
Distribuzione di una variabile qualitativa	Serie statistica
Distribuzione di una variabile qualitativa ordinale	Serie statistica ordinata
Distribuzione di una variabile quantitativa	Seriazione statistica
Valori assunti da una variabile quantitativa	Intensità
.....	.....

B. Su altri termini ci sono differenti usi e in alcuni di questi casi abbiamo messo la doppia dicitura.
- In genere si usa "diagramma a barre" per il caso di variabili qualitative e "istogramma" per variabili quantitative, ma talvolta si usa il secondo anche per le qualitative.
- Talvolta per indicare i marginali si usa anche il + al posto del punto: f_{+j} invece di $f_{\cdot j}$
- Per indicare i profili riga si usa anche $f_{/|i}$ e per quelli colonna $f_{i|j}$ ma abbiamo preferito non appesantire le notazioni.

C. Le indicazioni "variabile" o "**variabile casuale** (o aleatoria)" o "variabile casuale empirica (o statistica)" sono usate in modo equivalente. La terminologia – seppur universalmente diffusa – è comunque ambigua.
Precisiamo anzitutto che le variabili casuali non sono variabili, ma *funzioni*:

$$X : \Omega \to E \subseteq IR$$

dove Ω è l'insieme delle unità sperimentali, $\Omega = \{\omega_1, \ldots, \omega_n\}$ e E è l'insieme dei valori assunti da X; nel caso di variabili qualitative, l'insieme E sarà formato dalle codifiche numeriche.
La parola variabile indica piuttosto che si è in presenza di fenomeni che hanno variabilità.

Si indicano con le lettere *maiuscole* le variabili e con le lettere *minuscole* i valori assunti; quindi in genere: $X(\omega_i) = x_i$. Questa differente notazione è estremamente importante per non fare confusioni, soprattutto quando si affrontano argomenti di statistica inferenziale; è utile iniziare già l'uso corretto a livello di statistica descrittiva.

In *questo contesto* di statistica descrittiva anche la parola *casuale* non ha il significato usuale: non c'è nulla di casuale nell'assegnare a una unità sperimentale il valore assunto dalla variabile.
La casualità interviene quando si parla di probabilità; è per questo che taluni autori preferiscono definire le variabili casuali riferite a rilevazioni sperimentali come "variabili casuali empiriche (o statistiche)"; noi non abbiamo fatto questa distinzione, visto che – per ora – non affronteremo argomenti di statistica inferenziale.

D. Nella scheda è scritto che una variabile qualitativa corrisponde a un attributo non misurabile; si può precisare dicendo che – anche se è misurabile – non si utilizzano tali misure nella determinazione del valore. Ad esempio nel caso del sesso o di altri attributi fisici si possono misurare quantità legate al DNA che forniscono informazioni sulla variabile, ma quando si usano le modalità "maschio" o "femmina" non si fa riferimento a tali quantità.

2. Il problema dell'indipendenza

Per lo studio dei legami fra le variabili si è scelto invece di soffermarsi molto sulle rappresentazioni grafiche dei profili riga e colonna e delle loro deviazioni dal profilo medio.

Non si è affrontato invece nei dettagli il problema dell'indipendenza perché assume un significato più interessante a livello probabilistico, in quanto – in presenza di dati reali – non si avrà "mai" uguaglianza su tutte le celle della tabelle di una delle condizioni di indipendenza.

A questo proposito precisiamo che le condizioni equivalenti di indipendenza sono:

a) uguaglianza di tutti i profili riga fra loro e con il profilo marginale della variabile colonna:
$$f_{j|i} = f_{.j} \quad \text{per ogni riga i e per ogni cella j della riga}$$

b) uguaglianza di tutti i profili colonna fra loro e con il profilo marginale della variabile riga:
$$f_{i|j} = f_{i.} \quad \text{per ogni colonna j e per ogni cella i della colonna}$$

c) uguaglianza dei valori osservati con il prodotto dei corrispondenti marginali:
$$f_{ij} = f_{i.}\ f_{.j} \quad \text{per ogni cella, quindi per ogni i e j}$$

Dimostriamo a) implica c): $f_{j|i} = \frac{f_{ij}}{f_{i.}}$; essendo $f_{j|i} = f_{.j}$ si ha: $\frac{f_{ij}}{f_{i.}} = f_{.j}$ e quindi c).

Il viceversa è analogo.

Dimostriamo a) implica b): $\frac{f_{ij}}{f_{i.}} = f_{j.}$ implica $\frac{f_{ij}}{f_{.j}} = f_{i.}$ e quindi b). Il viceversa è analogo.

Da un punto di vista didattico ci sembra che le condizioni a) e b) siano più immediate.

Abbiamo scelto di non trattare indici *descrittivi* di indipendenza quali l'indice chi quadro (o X quadro) e l'indice di Mortera:

$$\chi^2 = n \sum \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} \qquad \text{e} \quad W = n \sum |f_{ij} - f_{i.}f_{.j}|$$

in quanto assumono un significato solo se sono relativizzati rispetto al loro valore massimo, ad esempio: $\frac{\chi^2}{\chi^2_{max}}$ in modo da renderli compresi fra 0 e 1. La condizione più lontana dall'indipendenza comporta l'indice massimo.

Tale valore massimo ovviamente dipende dal numero di righe e colonne della matrice, ma anche da come sono formati i marginali. Fissati i marginali, il valore massimo per gli indici precedenti si ha quando le frequenze relative sono "il più possibile" concentrate sulla diagonale. Però il calcolo esplicito per ogni gruppo di marginali fissati non è affatto semplice se il numero di righe e di colonne è elevato.

Se non si considerano i marginali, si può dimostrare (omettiamo la non difficile dimostrazione) che per una tabella a I righe e J colonne, il massimo valore del χ^2 è:

$$n \times \min \{I - 1, J - 1\}$$

Però relativizzare il valore dell'indice della tabella osservata rispetto a questo valore, senza tener conto dei marginali, può portare ad cattive interpretazioni dei risultati. Esempio.

osservati

5	2	3	10
20	6	4	30
55	2	3	60
80	10	10	100

indipendenza

8	1	1	10
24	3	3	30
48	6	6	60
80	10	10	100

massimo χ^2

10	0	0	10
20	10	0	30
50	0	10	60
80	10	10	100

L'indice χ^2 della tabella è 15.313; il massimo è 31.250. Il massimo indipendentemente dai marginali per una tabella a 3 righe e 3 colonne e 100 unità sperimentali è 400. Quindi l'indice relativo a marginali fissati è 0.490 mentre e quello relativo al massimo senza marginali è 0.154.

L'indice χ^2 assume invece importanza in ambito inferenziale.