Statistica I

Unità O: tabelle di contingenza

Tommaso Rigon

Università Milano-Bicocca

Anno Accademico 2020-2021

Unità O

Argomenti affrontati

- Tabelle di contingenza
- Distribuzione congiunta, marginale e condizionata
- Indipendenza in distribuzione
- Frequenze attese, indice χ^2 di Pearson

Riferimenti al libro di testo

- §7.1 — §7.5

Descrizione del problema

- Dopo il **disastro del Titanic**, una commissione d'inchiesta del **British Board of Trade** ha compilato una lista di tutti i 1316 passeggeri includendo le seguenti aggiuntive:
  - l'esito (salvato, non salvato)
  - la classe (I, II, III) in cui viaggiavano
  - il sesso, l'età, etc.

- In questa unità ci limitiamo a considerare le informazioni sull'esito e la classe. I dati sono quindi costituiti da una lunga lista del tipo:

Passeggero	**Classe**	**Esito**
Nome 1	II	Salvato
Nome 2	III	Non salvato
Nome 3	I	Non salvato
⋮	⋮	⋮
Nome 1316	III	Salvato

Le frequenze marginali

- La variabile **Classe** ha la seguente distribuzione di **frequenza marginale**:

Classe	**Frequenze assolute**	**Frequenze relative**
I	325	0.247
II	285	0.216
III	706	0.537

- La variabile **Esito** ha la seguente distribuzione di **frequenza marginale**:

Esito	**Frequenze assolute**	**Frequenze relative**
Salvato	499	0.379
Non salvato	817	0.621

Le frequenze congiunte

- Una sintesi che possiamo operare consiste nel costruire una tabella, detta **tabella di contingenza** oppure **tabella a doppia entrata**.

Esito	**Classe** I	II	III	Totale
Salvato	203	118	178	499
Non salvato	122	167	528	817
Totale	325	285	706	1316

- In questa tabella sono riportate le **frequenze congiunte**, ad esempio, il valore 203 rappresenta il numero di passeggeri che viaggiavano in I classe e che sono sopravvissuti.

- È quindi evidente che viaggiatori della I classe hanno ricevuto un **trattamento preferenziale**.

- La frazione di individui della I classe che si sono salvati è $203/325 \approx 0.63$.

- Invece, la frazione di viaggiatori della III classe che si sono salvati è $178/706 \approx 0.25$.

Le frequenze congiunte relative

- Possiamo anche considerare le **frequenze congiunte relative**, ottenute dividendo le frequenze congiunte per il numero totale $n = 1316$.

Esito	**Classe** I	II	III	Totale
Salvato	0.154	0.090	0.135	0.38
Non salvato	0.093	0.127	0.401	0.62
Totale	0.247	0.217	0.536	1

- La frazione di individui della I classe che si sono salvati è $0.154/0.247 \approx 0.63$.

- Invece, la frazione di viaggiatori della III classe che si sono salvati è $0.135/0.536 \approx 0.25$.

Tabella di contingenza

- Siano x ed y due variabili aventi modalità $c_1, \dots, c_k$ e $d_1, \dots, d_h$, rispettivamente.

- Una **tabella di contingenza** (a due variabili) per le coppie di dati $(x_1, y_1), \dots, (x_n, y_n)$ si presenta nella seguente forma:

	Variabile y					
Variabile x	d_1	$\dots$	d_j	$\dots$	d_k	**Totale**
c_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1k}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{ik}	n_{i+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_h	n_{h1}	$\dots$	n_{hj}	$\dots$	n_{hk}	n_{h+}
Totale	n_{+1}	$\dots$	n_{+j}	$\dots$	n_{+k}	n

- La frequenza n_{ij} è il numero di unità statistica che presentano contemporaneamente le modalità c_i e d_j.

Tabella di contingenza, frequenze relative

- Dividendo per n ciascun termine della precedente tabella, si ottiene inoltre:

	Variabile y					
Variabile x	d_1	$\dots$	d_j	$\dots$	d_k	**Totale**
c_1	f_{11}	$\dots$	f_{1j}	$\dots$	f_{1k}	f_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_i	f_{i1}	$\dots$	f_{ij}	$\dots$	f_{ik}	f_{i+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_h	f_{h1}	$\dots$	f_{hj}	$\dots$	f_{hk}	f_{h+}
Totale	f_{+1}	$\dots$	f_{+j}	$\dots$	f_{+k}	1

- La frequenza relativa $f_{ij} = n_{ij}/n$ è quindi la frazione di osservazioni che presentano contemporaneamente le modalità c_i e d_j.

Alcune proprietà

- **Proprietà**. Per definizione, vale quindi che

$$n_{i+} = \sum_{j=1}^{k} n_{ij}, \qquad n_{+j} = \sum_{i=1}^{h} n_{ij}, \qquad n = \sum_{i=1}^{h}\sum_{j=1}^{k} n_{ij}.$$

- **Proprietà**. Sempre per definizione, vale quindi che

$$f_{i+} = \frac{n_{i+}}{n} = \sum_{j=1}^{k} f_{ij}, \qquad f_{+j} = \frac{n_{+j}}{n} = \sum_{i=1}^{h} f_{ij}, \qquad \sum_{i=1}^{h}\sum_{j=1}^{k} f_{ij} = 1.$$

- **Esercizio**. Si verifichino le proprietà precedenti nel caso del Titanic.

Terminologia & notazione

Distribuzione congiunta

- Una tabella di contingenza nel suo complesso ci mostra la distribuzione congiunta di x ed y. Le n_{ij} per $i = 1, \dots, h$ e $j = 1, \dots, k$ sono chiamate frequenze congiunte.

- I valori f_{ij} per $i = 1, \dots, h$ e $j = 1, \dots, k$ sono chiamate frequenze congiunte relative.

Distribuzione marginale

- Le distribuzioni marginali per x e per y sono invece pari a

Variabile x	c_1	$\dots$	c_i	$\dots$	c_h	**Totale**
Frequenze assolute	n_{1+}	$\dots$	n_{i+}	$\dots$	n_{h+}	n
Frequenze relative	f_{1+}	$\dots$	f_{i+}	$\dots$	f_{h+}	1

Variabile y	d_1	$\dots$	d_j	$\dots$	d_k	**Totale**
Frequenze assolute	n_{+1}	$\dots$	n_{+j}	$\dots$	n_{+k}	n
Frequenze relative	f_{+1}	$\dots$	f_{+j}	$\dots$	f_{+k}	1

Terminologia & notazione

Distribuzione condizionata $(x \mid y = d_j)$

- La j-esima colonna mostra la distribuzione di x **condizionata** ad $y = d_j$ oppure, equivalentemente, la distribuzione di x dato $y = d_j$.

Distribuzione $x \mid y = d_j$	c_1	$\dots$	c_i	$\dots$	c_h	**Totale**
Frequenze assolute	n_{1j}	$\dots$	n_{ij}	$\dots$	n_{hj}	n_{+j}
Frequenze relative	n_{1j}/n_{+j}	$\dots$	n_{ij}/n_{+j}	$\dots$	n_{hj}/n_{+j}	1

Distribuzione condizionata $(y \mid x = c_i)$

- La i-esima colonna mostra la distribuzione di y **condizionata** ad $x = c_i$ oppure, equivalentemente, la distribuzione di y dato $x = c_i$.

Distribuzione $y \mid x = c_i$	d_1	$\dots$	d_j	$\dots$	d_k	**Totale**
Frequenze assolute	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{ik}	n_{i+}
Frequenze relative	n_{i1}/n_{i+}	$\dots$	n_{ij}/n_{i+}	$\dots$	n_{ik}/n_{i+}	1

Il disastro del Titanic, distribuzioni condizionate

(**Esito** $\mid$ **Classe** $= I$)	Salvato	Non Salvato	**Totale**
Frequenze assolute	203	122	325
Frequenze relative	0.625	0.375	1

(**Esito** $\mid$ **Classe** $= II$)	Salvato	Non Salvato	**Totale**
Frequenze assolute	118	167	285
Frequenze relative	0.41	0.59	1

(**Esito** $\mid$ **Classe** $= III$)	Salvato	Non Salvato	**Totale**
Frequenze assolute	178	528	706
Frequenze relative	0.25	0.75	1

Il disastro del Titanic, distribuzioni condizionate

(**Classe** | **Esito** = Salvato)	I	II	III	**Totale**
Frequenze assolute	203	118	178	499
Frequenze relative	0.41	0.24	0.36	1

(**Classe** | **Esito** = Non salvato)	I	II	III	**Totale**
Frequenze assolute	122	167	528	817
Frequenze relative	0.15	0.20	0.65	1

Sommario

Distribuzione congiunta

- La distribuzione congiunta è il "nucleo" della tabella. Comprende il numero di osservazioni che presentano una modalità della prima variabile contemporaneamente (**congiuntamente**) ad una modalità della seconda variabile.

Distribuzione condizionata

- Le distribuzioni condizionate considerano le frequenze della prima variabile solamente (**condizionatamente**) per certi valori della seconda variabile.

Distribuzione marginale

- Le distribuzioni marginali considerano le frequenze della prima variabile a prescindere (**marginalmente**) dall'esito della seconda variabile.

La dipendenza tra variabili

- Ri-consideriamo i dati del disastro del Titanic.

- Abbiamo notato che la sopravvivenza **dipende** dalla classe in cui viaggiava il passeggero visto che la frazione di sopravvissuti all'incidente varia al variare della classe.

- Indichiamo con y la **Classe** e con x l'**Esito**. Diremo quindi che la variabile x **dipende** dalla variabile y.

- **Dipendenza di x dato y**. La variabile x **dipende** dalla variabile y se le distribuzioni condizionate di x dato y sono tra loro diverse in termini di **frequenze relative**.

- La dipendenza di y dato x ovviamente è definita in maniera speculare.

L'indipendenza tra variabili

- Supponiamo che la distribuzione congiunta sia la seguente.

Esito	**Classe** I	II	III	Totale
Salvato	150	200	300	650
Non salvato	300	400	600	1300
Totale	450	600	900	1950

- Nonostante le frequenze assolute delle distribuzioni condizionate (**Esito** | **Classe**) siano diverse tra loro, le frequenze relative risultano invece uguali.

Esito	**Classe** I	II	III
Salvato	0.33	0.33	0.33
Non salvato	0.67	0.67	0.67
Totale	1	1	1

Indipendenza tra variabili

- Nel caso della tabella precedente è quindi ragionevole affermare non esiste dipendenza di x da y.

- Si ricordi che n_{ij}/n_{+j} è la frequenza relativa di c_i nella distribuzione di x condizionata a $y = d_j$.

- **Indipendenza di x da y**. La variabile x è **indipendente in distribuzione** da y se per ogni $i = 1, \dots, h$ vale che
$$\frac{n_{i1}}{n_{+1}} = \cdots = \frac{n_{ij}}{n_{+j}} = \cdots = \frac{n_{ik}}{n_{+k}}.$$
Viceversa, diremo che x dipende in distribuzione da y.

- In altri termini, x è indipendente da y se tutte le distribuzioni condizionate $(x \mid y = d_j)$ sono uguali in termini di frequenze relative, per ogni $j = 1, \dots, k$.

Indipendenza, distribuzione marginale

- **Proprietà**. Se x è indipendente da y, allora le distribuzioni condizionate sono tutte uguali e pari alla distribuzione marginale di x. In altri termini

$$f_{i+} = \frac{n_{i+}}{n} = \frac{n_{ij}}{n_{+j}}, \qquad i = 1, \dots, h,$$

per ogni valore di $j = 1, \dots, k$.

- Per dimostrare questa proprietà, si noti anzitutto che l'indipendenza implica che

$$\frac{n_{ij'}}{n_{+j'}} = \frac{n_{ij}}{n_{+j}} \implies n_{ij'} = n_{ij}\frac{n_{+j'}}{n_{+j}},$$

per ogni $j, j' = 1, \dots, k$ e $i = 1, \dots, h$. Quindi:

$$\frac{n_{i+}}{n} = \frac{1}{n}\sum_{j'=1}^{k} n_{ij'} = \frac{1}{n}\sum_{j'=1}^{k} \frac{n_{ij}n_{+j'}}{n_{+j}} = \frac{n_{ij}}{n_{+j}}\sum_{j'=1}^{k}\frac{n_{+j'}}{n} = \frac{n_{ij}}{n_{+j}}\frac{n}{n} = \frac{n_{ij}}{n_{+j}}.$$

Indipendenza tra variabili

- **Proprietà**. Se x è indipendente in distribuzione da y, allora y è indipendente in distribuzione da x. In altri termini, per ogni $j = 1, \dots, k$ vale anche che

$$\frac{n_{1j}}{n_{1+}} = \cdots = \frac{n_{ij}}{n_{i+}} = \cdots = \frac{n_{hj}}{n_{h+}} = \frac{n_{+j}}{n} = f_{+j}.$$

- **Nota**. L'indipendenza è pertanto un concetto **simmetrico**. Possiamo quindi parlare di indipendenza tra x ed y senza dover indicare necessariamente una direzione.

- Per dimostrare questa proprietà, si noti che se x è indipendente da y, allora per la proprietà precedente

$$f_{i+} = \frac{n_{i+}}{n} = \frac{n_{ij}}{n_{+j}}, \qquad i = 1, \dots, h, \quad j = 1, \dots, k.$$

  Di conseguenza, avremo che

$$\frac{n_{ij}}{n_{i+}} = \frac{n_{+j}}{n} = f_{+j}, \qquad i = 1, \dots, h, \quad j = 1, \dots, k.$$

Le frequenze attese

- Supponiamo siano note le **distribuzioni marginali** delle variabili x ed y.
- Inoltre, supponiamo che x ed y siano indipendenti in distribuzione. Le frequenze congiunte pertanto devono essere necessariamente pari a $n_{ij} = (n_{i+}n_{+j})/n$.
- **Frequenze attese**. Le frequenze attese (assolute e relative) sono definite a partire dalle frequenze marginali come segue

$$\hat{n}_{ij} = \frac{n_{i+}n_{j+}}{n}, \qquad \hat{f}_{ij} = \frac{\hat{n}_{ij}}{n} = f_{i+}f_{+j}, \qquad i = 1, \dots, h, \quad j = 1, \dots, k.$$

- In altri termini, le frequenze attese sono le frequenze **congiunte** che è lecito attendersi sotto l'**ipotesi di indipendenza** tra le variabili x ed y.

Disastro del Titanic, frequenze attese

- Frequenze attese

Esito	**Classe** I	II	III	Totale
Salvato	123.2	108.1	267.7	499
Non salvato	201.8	176.9	438.3	817
Totale	325	285	706	1316

- Frequenze attese relative

Esito	**Classe** I	II	III	Totale
Salvato	0.094	0.082	0.203	0.38
Non salvato	0.153	0.134	0.333	0.62
Totale	0.247	0.217	0.536	1

- **Esercizio**. Si verifichi che le frequenze condizionate relative sono tutte uguali.

La massima dipendenza

- La **massima dipendenza** di x dato y si verifica, viceversa, quando la conoscenza della variabile y determina univocamente la variabile x.

- Supponiamo di osservare il seguente insieme di dati

	Classe			
Esito	I	II	III	Totale
Salvato	325	285	0	610
Non salvato	0	0	706	706
Totale	325	285	706	1316

- Condizionatamente alla variabile **Classe**, la variabile **Esito** è univocamente determinata.

- Il viceversa non è vero. Conoscere l'**Esito** non determina univocamente la **Classe**.

- **Nota**. La (perfetta) dipendenza è quindi un concetto **asimmetrico**. La perfetta dipendenza di x dato y non implica il contrario.

Connessione e indice χ^2

- Siamo interessati a trovare un indice di **connessione**, ovvero un indice utilizzato per **quantificare la dipendenza** tra due variabili x ed y.

- È ragionevole basare tale indice sulle **contingenze**, ovvero sulle differenze

$$(\text{contingenza}_{ij}) = n_{ij} - \hat{n}_{ij}, \qquad i = 1, \dots, h, \quad j = 1, \dots, k.$$

- **Indice χ^2 di Pearson**. L'indice di connessione χ^2 è definito come

$$\chi^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = n \left(\sum_{i=1}^{h} \sum_{j=1}^{k} \frac{f_{ij}^2}{f_{i+} f_{+j}} - 1 \right).$$

- L'indice χ^2 è pertanto sempre maggiore o uguale a zero. È pari a zero in particolare in caso di indipendenza.

Indice χ^2 normalizzato

Teorema (senza dimostrazione)

- L'indice χ^2 di Pearson è tale che

$$\chi^2 \leq n \min\{h-1, k-1\}.$$

- Se $k \leq h$ l'indice raggiunge il suo massimo in caso di dipendenza perfetta di x dato y.
- Se $h \leq k$ l'indice raggiunge il suo massimo in caso di dipendenza perfetta di y dato x.

- Il precedente teorema consente di definire un indice χ^2 **normalizzato**, ovvero

$$\chi^2_{\text{norm}} = \frac{\chi^2}{(\text{massimo valore di } \chi^2)} = \frac{\chi^2}{n \min\{h-1, k-1\}},$$

che è ovviamente tale che $0 \leq \chi^2_{\text{norm}} \leq 1$.

- Utilizzando i dati del Titanic, si ottiene $\chi^2 = 133.05$ e $\chi^2_{\text{norm}} = 0.1011$.