STATISTICA DESCRITTIVA - SCHEDA N. 1
VARIABILI QUALITATIVE

1. Le variabili qualitative

Una caratteristica (o variabile) si dice **qualitativa** se è un attributo non misurabile. Ad esempio: il genere, il colore degli occhi (a livello macroscopico), il livello di scolarità, etc.
Precisiamo che talvolta una variabile misurabile può essere considerata qualitativa quando non si utilizzano le 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.

I risultati assunti (es M e F per il genere) si chiamano **modalità** o **livelli**. Spesso si codificano con valori numerici Ad esempio M$\rightarrow$1 e F$\rightarrow$2 per il genere, oppure analfabeta $\rightarrow$ 1, elementare $\rightarrow$ 2, media $\rightarrow$ 3, superiore $\rightarrow$ 4, università $\rightarrow$ 5 per il livello di scolarità. Mentre nel secondo caso la codifica numerica corrisponde a un ordine crescente di livello di scolarità, le modalità della variabile genere non sono ordinabili. Se le modalità hanno un ordine intrinseco, le variabili si dicono **ordinali**, altrimenti si dicono **nominali**.

2. La distribuzione di una variabile qualitativa e le sue rappresentazioni: le tabelle di contingenza e i diagrammi a barre

*Le rappresentazioni usuali per le variabili qualitative sono le **tabelle di contingenza** (o semplicemente **tabelle**) e i **diagrammi a barre** (o **istogrammi**).*

Osserviamo che per le variabili nominali l'ordine delle modalità nelle tabelle e nelle rappresentazioni grafiche è arbitrario.

Nelle tabelle di contingenza ad ogni valore *i* della variabile è associato il numero n_i delle volte in cui tale valore si riscontra nelle n osservazioni oppure la sua frequenza relativa (n_i/n). La tabella con le frequenze relative viene anche detta tabella della **distribuzione della variabile.**

I diagrammi a barre sono rappresentazioni grafiche in cui nelle ascisse sono riportati i valori assunti dalla variabile e in ordinata i conteggi o le frequenze.

ESEMPIO: Consideriamo la suddivisione dei gruppi sanguigni (A; B; AB;0) in una popolazione caucasica.

	A	B	AB	0	
Conteggi	60	16	7	66	149
Frequenze percentuali	40.3	10.7	4.7	44.3	100

Numero di persone con il gruppo A

Frequenza percentuale delle unità che hanno AB come gruppo

Numero di osservazioni effettuate

Tabella 1. Rappresentazione della distribuzione di una variabile qualitativa

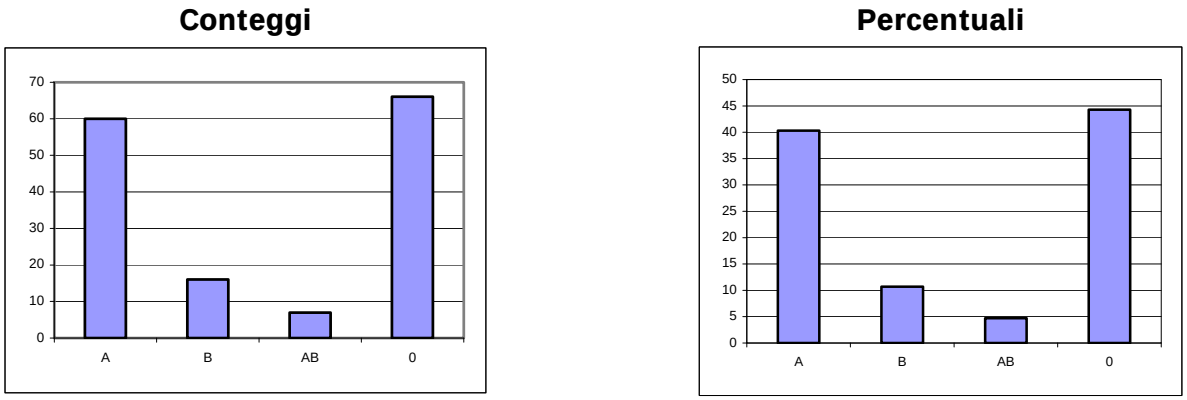


Figura 1. Diagrammi a barre per i conteggi e la distribuzione di una variabile qualitativa

Osserviamo che le due rappresentazioni grafiche sono diverse solo per quanto riguarda la scala delle ordinate.

Per approndire le notazioni sulle tabelle di contingenza per la distribuzione delle variabili vedi Appendice 1.

3. La distribuzione congiunta di due variabili qualitative: le tabelle di contingenza a due entrate e alcune diagrammi a barre

I risultati della rilevazione di due caratteristiche qualitative X e Y sulla stessa popolazione di numerosità n possono essere schematizzati con tabelle di contingenza "a due entrate", cioè tabelle in cui il numero nella posizione ij indica il conteggio n_{ij} oppure la frequenza f_{ij} (= n_{ij} /n) dell'osservazione X=i e Y=j . La tabella con le frequenze relative viene anche detta tabella della ***distribuzione congiunta di X e Y.***

ESEMPIO: La tabella seguente mostra i dati *espressi in percentuale* di 160 laureati in Matematica presso l'Università di Genova negli anni 1990-1993 e il tempo di attesa della prima occupazione. E' da notare che il questionario è stato fatto nel 1994 e quindi i dati dei quattro anni non sono omogenei tra loro (ad esempio non ci possono essere laureati del 1993 che hanno trovato lavoro dopo un anno).

Percentuale degli intervistati laureati nel 1992 **E** che hanno trovato lavoro entro 6 mesi. Quindi il conteggio assoluto sarà 16 (=10*160/100)

Colonna delle frequenze delle classi ANNI (ottenuta sommando le righe)

ANNI\TEMPO	<6 mesi	6-12 mesi	>12 mesi	Disoccup.	TOTALE
1990	12.5	2.5	0.0	3.5	18.5
1991	15.5	2.5	2.0	7.5	27.5
1992	10.0	3.5	3.5	10.0	27.0
1993	4.0	4.0	0.0	19.0	27.0
TOTALE	42.0	12.5	5.5	40.0	100

Riga delle frequenze delle classi TEMPO DI ATTESA (ottenuta sommando le colonne)

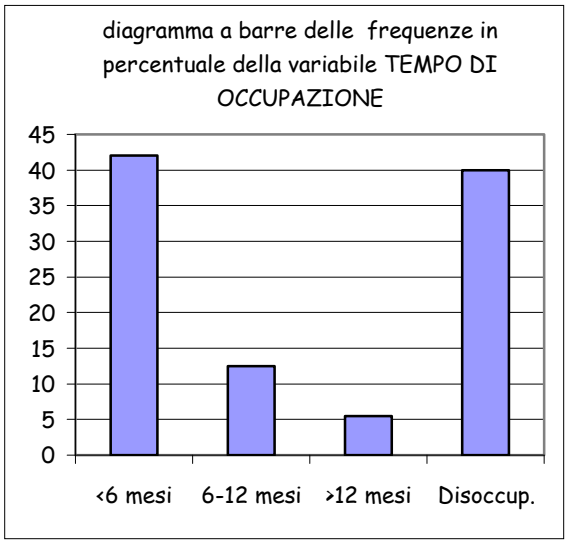
Percentuale degli intervistati che hanno trovato lavoro dopo un anno dalla laurea

Tabella 2. Rappresentazione della distribuzione congiunta di due variabili qualitative

L'ultima riga e l'ultima colonna sono dette **distribuzioni marginali** (o totali) delle caratteristiche qualitative studiate.

Per approfondire le notazioni sulle tabelle di contingenza "a due vie" vedi Appendice 2.

Si possono tracciare differenti *diagrammi a barre* a seconda di ciò che si vuole evidenziare.



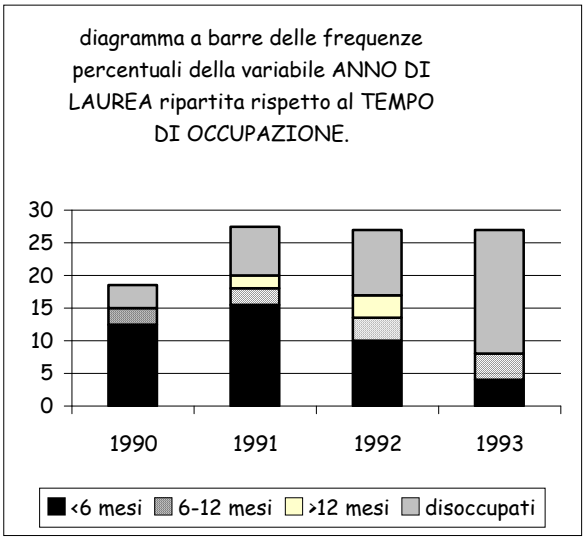


Figura 2. Alcuni diagrammi a barre per la distribuzione di due variabili qualitative

4. I profili riga e i profili colonna e le loro rappresentazioni

Uno studio completo di due variabili qualitative X e Y comprende anche l'esame del comportamento di una variabile rispetto all'altra.
Una lettura approssimativa della tabella di contingenza potrebbe condurre a conclusioni non giuste. Nell'esempio dei laureati, la percentuale rispetto al totale di chi ha trovato lavoro da 6 a 12 mesi dopo la laurea è la stessa per i laureati nel 1990 (2.5%) e nel 1991 (2.5%). Ma per confrontare i tempi di attesa della prima occupazione nei diversi anni bisogna tener conto anche di quante persone si sono laureate in ciascun anno e quindi è opportuno confrontare i valori con la percentuale dei laureati nei due anni.

La frequenza relativa dell'osservazione *Y=j*, conoscendo *X=i* è il rapporto fra il la frequenza di f_{ij} e la frequenza totale f_i delle osservazioni *X=i*. Questi dati si possono visualizzare in una nuova tabella (**tabella dei profili riga**).

Riprendiamo l'esempio dei laureati. La tabella dei profili riga *espressi in percentuale* diventa:

ANNI\TEMPO	<6 mesi	6-12 mesi	>12 mesi	Disoccup.	TOTALE
1990	67.6	13.,5		18.9	100
1991	56.4	9.1	7.3	27.3	100
1992	37.0	13.0	13.0		100
1993			0.0	70.4	100

Nel 1990 la percentuale di laureati che hanno trovato lavoro fra 6-12 mesi è 13.5%, nel 1991 è 9.1%

Tabella 3. Rappresentazione dei profili riga

ESERCIZO Completa la tabella calcolando i quattro valori mancanti
Se si considera la variabile X condizionata da Y si costruiscono in maniera analoga le **tabelle dei profili colonna**. Nell'esempio:

ANNI\TEMPO	<6 mesi	6-12 mesi	>12 mesi	Disoccup.
1990	29.8		0.0	8.8
1991	36.9	20.0	36.4	18.8
1992	23.8	28.0		25.0
1993	9.5	32.0	0.0	
TOTALE	100	100	100	100

Percentuale di laureati nel 1991 tra quelli che hanno trovato lavoro da 6 a 12 mesi dopo la laurea

ESERCIZO. Completa la tabella calcolando i tre valori mancanti

Tabella 4. Rappresentazione dei profili colonna

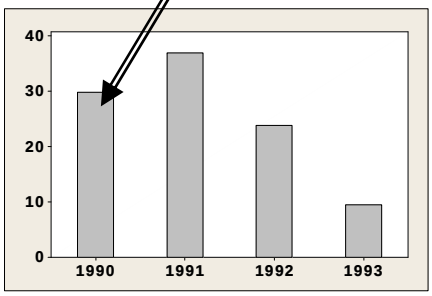


Figura 3. Diagramma a barre del profilo colonna relativo a chi ha trovato lavoro in meno di 6 mesi.

I profili riga permettono di evidenziare se la variabile Y risente del condizionamento della variabile X. In particolare, se le righe della tabella (o i corrispondenti *diagrammi a barre*) sono simili si può ipotizzare che le due variabili non si condizionino. Analogo discorso vale per i profili colonna.

Nell'esempio dei laureati, il confronto dei *diagrammi a barre* delle frequenze del tempo di attesa della prima occupazione relativi ai singoli anni (cioè la rappresentazione grafica profili riga, vedi Figura 4) mostra una dipendenza dal tempo di attesa e l'anno di laurea.
Ricordiamo però che l'indagine è stata fatta nel 1994 e quindi le ultime due colonne dei quattro anni non sono omogenei fra loro. In generale, quando si interpretano i dati, è sempre opportuno riferirsi all'ambito nel quale sono stati raccolti.

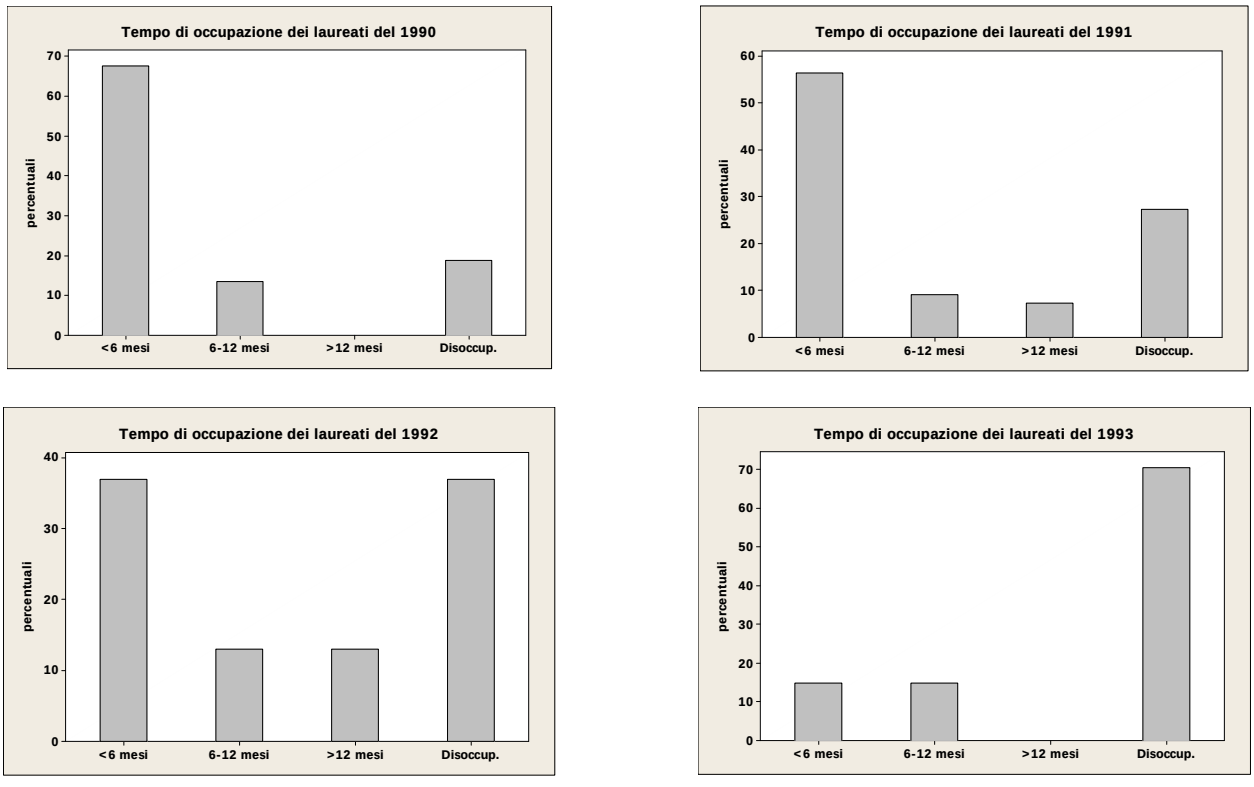


Figura 4. Diagrammi a barre dei profili riga.

Questi diagrammi a barre si possono anche confrontare con quello dei tempi di attesa del totale dei laureati intervistati (ottenuto dall'ultima riga della tabella di contingenza)

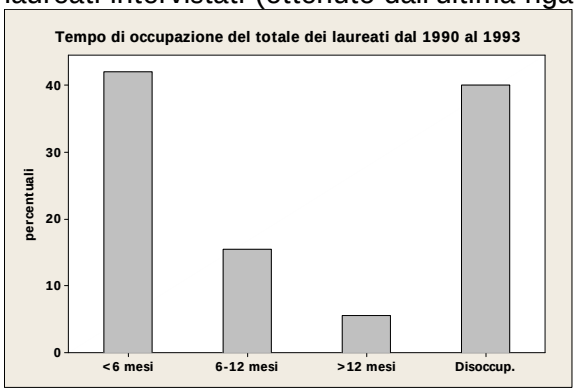


Figura 5. Diagramma a barre del totale colonna.

Tale diagramma si può pensare come "grafico medio" della popolazione dei laureati considerati, dove la media è "pesata" rispetto alla percentuale dei laureati nei diversi anni.

Vediamo, ad esempio, come si può calcolare la percentuale di chi ha trovato lavoro entro 6 mesi **rispetto al totale della popolazione** a partire dai profili riga e dal totale per anno.

- nel 1990 sono il 67.6% del 18.5% del totale dei laureati, cioè il 12.74 (= 0.676 x 0.185)
- nel 1991 sono il 56.4% del 27.5% del totale dei laureati, cioè il 15.55
- e così via

ANNI	<6 mesi	TOTALE
1990	67.6	18.5
1991	56.4	27.5
1992	37.0	27.0
1993	14.8	27.0

Sommando questi valori si ha la percentuale di chi ha trovato lavoro entro 6 mesi rispetto al totale della popolazione:

$$0.676 \times 0.185 + 0.564 \times 0.275 + 0.370 \times 0.270 + 0.148 \times 0.270 = 0.42$$

Sarebbe stato sbagliato fare una media non pesata dei valori dei profili riga, cioè:

$$(0.676 + 0.564 + 0.370 + 0.148)/4 = 0.4394$$

Essendo medie pesate dei profili, le distribuzioni marginali sono dette anche **distibuzioni medie.**

Per confrontare meglio i profili riga con il totale (o media) generale dei laureati si possono considerare le **differenze dei profili dal totale**:

ANNI\TEMPO	<6 mesi	6-12 mesi	>12 mesi	Disoccup.
1990	25.6	1.0	-5.5	-21.1
1991	14.4	-3.4	1.8	-12.7
1992	-5.0	0.5	7.5	-3.0
1993	-27.2	2.3	-5.5	30.4

Tabella 5. Rappresentazione delle differenze dei profili riga dalla distribuzione totale (o marginale o media) della variabile colonna

Osserviamo che la somma per riga delle deviazioni dei profili riga dal totale è 0.

Anche i valori di questa tabella possono essere rappresentati graficamente, come si vede nella seguente figura.

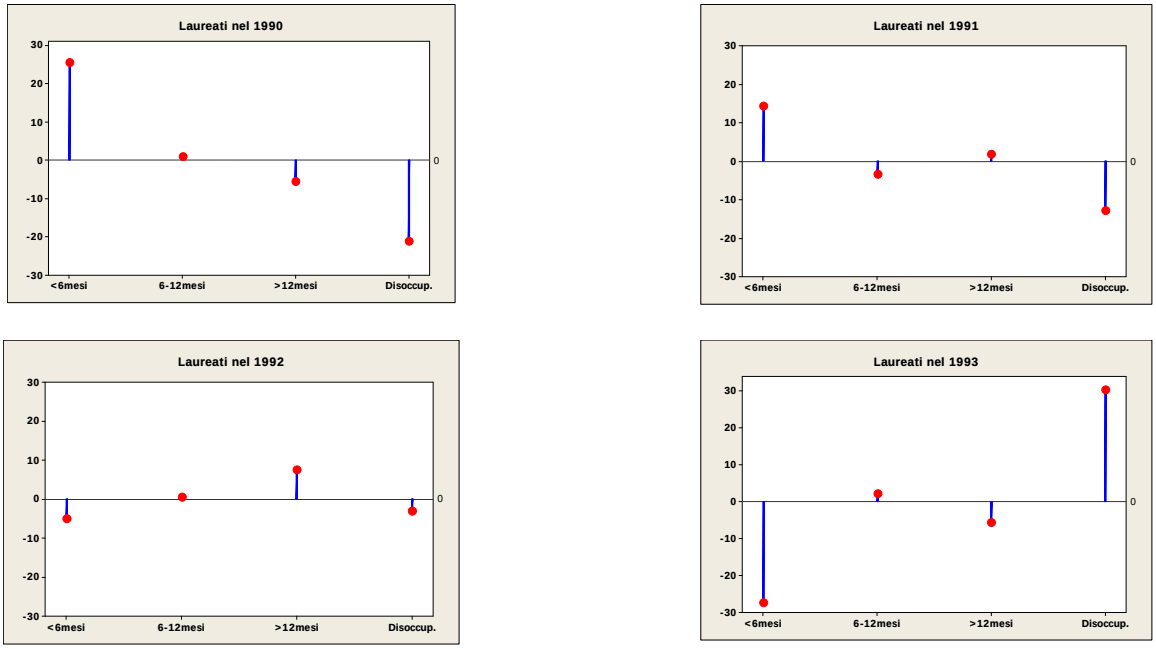


Figura 6. Diagrammi a barre delle differenze dei profili riga dalla distribuzione totale (o marginale o media) della variabile colonna.

5. L'indipendenza

Vediamo un altro modo per studiare la mancanza di legami fra le variabili.
Se sono fissate le distribuzioni marginali, come deve essere la tabella di contingenza della distribuzione congiunta se non ci sono legami fra gli anni di laurea e il tempo di attesa della prima occupazione?

Ad esempio, la frequenza degli studenti che si sono laureati nel 1990 e che hanno trovato lavoro entro 6 mesi sarà il 18.5% del 42%, ovvero il 7.7%.
In generale, in caso di **assenza di legami** (o di **indipendenza**), in ogni cella della tabella della **distribuzione congiunta** ci dovrebbe essere il prodotto dei marginali corrispondenti.

ANNI\TEMPO	<6 mesi	6-12 mesi	>12 mesi	Disoccup.	TOTALE
1990	7.7	2.3	1.0	7.5	18.5
1991	11.5	3.5	1.5	11.0	27.5
1992	11.4	3.4	1.5	10.7	27.0
1993	11.4	3.4	1.5	10.7	27.0
TOTALE	42.0	12.5	5.5	40.0	100.0

12.5x27.5/100

Tabella 6. Rappresentazione della distribuzione congiunta di due variabili qualitative nel caso di indipendenza

Naturalmente una tabella di questo tipo sarà difficilmente ottenibile nelle rilevazioni sperimentali, ma - come nel caso dei profili in cui si osserva se c'è *somiglianza* con il profilo mariginale – permette di avere un elemento di confronto rispetto ai dati osservati.
DOMANDA: Come si potrebbe misurare la "distanza" fra la tabella dei dati osservati e la tabella dell'indipendenza?

Si osservi che "legame" o "condizionamento" di due variabili non significa che una variabile è causa dell'altra.

Nel caso degli anni di laurea e il tempo di attesa della prima occuapazione, si può dire " i tempi di attesa variano (o no) *a seconda* degli anni di laurea"; ma in generale i dati – da soli - non forniscono informazioni sulla causalità dei fenomeni. Vediamo un altro esempio.

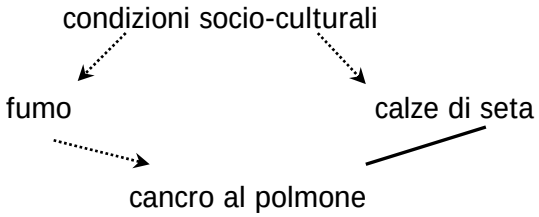
ESEMPIO: In una indagine svolta in modo accurato su un campione numeroso di donne negli Stati Uniti negli anni '30 si è trovato un forte legame fra l'avere il cancro ai polmoni e il portare le calze di seta. Che cosa se ne deduce?

"visto che non può essere che il cancro ai polmoni induca le donne a portare le calze di seta"

allora

"il portare le calze di seta favorisce il cancro al polmone"

Che cosa c'è dietro?



Dai dati si osserva il legame indicato con linea unita senza freccia; eventuali causalità, indicate con linee tratteggiate, possono essere individuate studiando a fondo il problema.

ESERCIZI

1) A fianco sono riportati i dati raccolti su 15 soggetti per due variabili che hanno le modalità codificate con 0 e 1.

Costruire due tabelle di contingenza a due entrate con i valori congiunti delle due variabili, una con i valori assoluti e l'altra con quelli percentuali

A	B
1	1
1	0
1	0
1	1
1	1
0	0
0	0
0	1
1	0
1	0
0	0
1	1
1	1
1	1
1	1

2) Nella tabella a fianco sono riportati i dati (conteggi) riguardanti a una seconda indagine sui tempi di attesa della prima occupazione di laureati in Matematica a Genova. In questo caso l'indagine è stata condotta alla fine del 2000.

	<6 mesi	6-12 mesi	>12 mesi	Disoccup.
1996	15	4	6	1
1997	18	7	2	3
1998	17	8	3	0
1999	27	8	1	2
2000	11	1	0	4

a) Costruire una tabella con la distribuzione congiunta dell'anno di laurea e del tempo di attesa e le due distribuzioni marginali.
b) Costruire i profili riga e opportune rappresentazioni grafiche.
c) Consideriamo il tempo di attesa della prima occupazione. Utilizzando la distribuzione (marginale) del tempo di attesa dell'indagine 1990-1994 e quella dell'indagine 1996-2000, calcolare la distribuzione totale per gli 8 anni insieme.
d) Analizziamo i legami fra le tabelle delle due indagini separate e le corrispondenti tabelle con i dati di tutti i 9 anni. Dire se sono uguali o diverse
    a. la tabella della distribuzione congiunta anno/tempo di attesa
    b. la tabella dei profili riga
    c. la tabella dei profili colonna

3) La seguente tabella riporta la distribuzione della popolazione residente in Italia al Censimento del 1981 secondo due caratteri: il *titolo di studio* (*Y*) e il *ramo di attività* (*X*):

X *Y*	laureati	diplomati	licenza media	licenza elementare	totali
Agricoltura	13	77	323	1221	1634
Industria	120	951	2714	3731	7516
Commercio	63	497	1280	1489	3329
trasporti e comunic.	19	255	464	476	1214
credito e assicuraz.	176	447	224	70	917
pubblica amminis.	784	1302	1146	1013	4245
Totali	1175	3529	6151	8000	18855

I dati in migliaia sono di abitanti, fonte ISTAT. Nella tabella non sono considerati i residenti privi di titolo di studio

a) Costruite la tabella di contingenza della distribuzione congiunta (*X,Y)*..

b) Costruite le due distribuzioni di frequenza marginali e due diagrammi a barre che le rappresentino. Rispondete poi alla seguente domanda: "data la distribuzione delle frequenze congiunte hai visto come sia possibile ottenere le distribuzioni marginali. È vero che date le due distribuzioni marginali è possibile, in generale, risalire alla distribuzione delle frequenze congiunte? Perché?" In generale, conoscendo le distribuzioni marginali, con quante celle vuote possiamo ancora ricostruire la tabella?"
c) Costruite le tabelle dei profili riga e dei profili colonna delle due variabili *X* e Y.
d) Che cosa pensate si possa dire relativamente al condizionamento di una delle due variabili rispetto all'altra? E relativamente alla dipendenza causale di *Y* da *X* e di *X* da *Y*? Giustificate la vostra risposta.
e) Discutete su questo esempio la frase: "l'analisi dei profili riga e quella dei profili colonna porta alle stesse conclusioni, cambia solo l'ottica con cui studiare il fenomeno".

4) La seguente tabella riporta la distribuzione della popolazione residente in Italia al Censimento del 2001 secondo due caratteri: il *titolo di studio* (*Y*) e il *ramo di attività* (*X*).
(dati in rete sul sito ISTAT).

SEZIONI DI ATTIVITÀ ECONOMICA	Grado di istruzione						
	Laurea	Diploma univers. o simile	Diploma di scuola sec. sup.	Licenza di scuola media	Licenza di scuola elementare	Nessun titolo di studio	Totale
Agricoltura	26.300	4.050	215.565	471.119	374.408	62.236	1.153.678
Industria	344.850	43.886	2.406.304	3.194.435	948.994	90.512	7.028.981
Commercio	128.317	31.457	1.630.412	1.697.209	463.231	35.912	3.986.538
Trasporti e comunicazioni	51.239	6.878	436.271	379.164	97.698	7.779	979.029
Credito e assicurazioni, servizi alle imprese, noleggio	552.449	31.792	1.169.820	255.016	39.723	3.881	2.052.681
Altre attivita'	1.304.837	279.338	2.426.284	1.410.038	335.906	36.422	5.792.825
Totale	2.407.992	397.401	8.284.656	7.406.981	2.259.960	236.742	20.993.732

a) Ricavare da questi dati una la tabella il più simile possibile a quella dell'esercizio precedente e – scegliendo una analisi per riga o per colonna – effettuare confronti sui cambiamenti avvenuti nei 20 anni considerati.

5) La seguente tabella riporta la distribuzione dei decessi per fasce di età (Y) e per sesso (X) della popolazione italiana nell'anno 2002 (fonte ISTAT):

Morti per età e sesso – Anno 2002			
	M	F	TOT
0-4 anni	1563	1211	2774
5-19 anni	1343	572	1915
20-39 anni	8379	3266	11645
40-54 anni	16320	9238	25558
55-69 anni	57232	30715	87947
70-79 anni	88331	64105	152436
80-89 anni	79272	105664	184936
90 e oltre	26856	66323	93179
TOTALE	279296	281094	560390

a) Costruire la tabella di contingenza con i valori percentuali della distribuzione congiunta (X,Y).
b) Costruire le due distribuzioni di frequenza marginali e due diagrammi a barre che le rappresentino.
c) Date le due distribuzioni marginali è possibile, risalire alla distribuzione delle frequenze congiunte? Come si fa?

APPENDICE: NOTAZIONI

1. Le liste o le tabelle "a una entrata"

I valori di una lista (o tabella "a una entrata") sono spesso indicati utilizzando la posizione che occupano. Ad esempio nella seguente tabella i conteggi sono indicati con n_1, n_2, n_3, n_4 e le frequenze percentuali con f_1, f_2, f_3, f_4.

A	B	AB	0
60	16	7	66
40.3	10.7	4.7	44.3

Un generico elemento è indicato con n_i o con f_i. Bisogna precisare quali valori può assumere l'indice i; nel nostro caso va da 1 a 4 e si scrive: i = 1, ..., 4. Per una tabella generica con I celle (o caselle) si scrive: i = 1, ..., I. Attenzione alle lettere maiuscole e minuscole.

2. Le posizioni e gli elementi di una tabella "a due entrate"

Le celle di una tabella sono spesso indicate con una coppia di valori che corrispondono alla loro posizione (come nel gioco della battaglia navale, ma con numeri sia per le coordinate orizzontali che per le coordinate verticali).
Il **primo** valore si riferisce alla posizione sulle **righe** e il **secondo** alla posizione sulle **colonne**.
Ad esempio la cella di posizione (3, 2) è nella terza riga e nella seconda colonna (escluse la colonna e la riga di intestazione), quindi è la cella

ANNI\TEMPO	<6 mesi	6-12 mesi	>12 mesi	Disoccup.
1990	12.5	2.5	0.0	3.5
1991	15.5	2.5	2.0	7.5
1992	10.0	3.5	3.5	10.0
1993	4.0	4.0	0.0	19.0

In generale la cella di posizione (i, j) si trova nella generica riga i e nella generica colonna j, con i = 1, ..., I e con j = 1, ..., J. Nel nostro esempio con i = 1, ..., 4 e con j = 1, ..., 4.

La tabella che stiamo considerando contiene le frequenze relative percentuali: indichiamo con $f_{3,2}$ il valore 3.5 che corrisponde alla percentuale di coloro che si sono laureati nel 1992 (3 riga) e hanno trovato lavoro dopo 6-12 mesi.
In generale il contenuto delle celle di questa tabella è indicato con $f_{i,j}$ oppure più semplicemente con f_{ij} (senza la virgola) con i = 1, ..., I e con j = 1, ..., J.

3. Le somme

Per indicare somme di tutti gli elementi di una lista si scrive ad esempio:

$\sum_{i=1}^{4} f_i$ che si legge "somma per i che va da 1 a 4 di f_i" e corrisponde a $f_1 + f_2 + f_3 + f_4$

Sotto il segno $\sum$ di somma (o sommatoria) si mette la lettera con l'indice e, dopo il segno "uguale", il valore iniziale, sopra a $\sum$ si mette il valore finale.

Quindi possiamo scrivere: $\sum_{i=1}^{4} f_i$ = 100 (oppure = 1) e $\sum_{i=1}^{4} n_i = n$.

Le cose si complicano un po' quando abbiamo due indici come in una tabella a doppia entrata.
Consideriamo nuovamente la tabella

ANNI\TEMPO	<6 mesi	6-12 mesi	>12 mesi	Disoccup.
1990	12.5	2.5	0.0	3.5
1991	15.5	2.5	2.0	7.5
1992	10.0	3.5	3.5	10.0
1993	4.0	4.0	0.0	19.0

Che cosa vuol dire e quanto vale: $\sum_{j=1}^{4} f_{3j}$?

Vuol dire che sto considerando la terza riga (primo indice uguale a 3 fissato) e faccio la somma dei valori di tutte le colonne di questa riga; quindi il risultato è 27.0.
Talvolta questo valore, cioè il totale marginale, si indica anche con $f_{3\bullet}$ oppure con f_{3+}

Che cosa vuol dire e quanto vale: $\sum_{i=1}^{4} f_{i4}$?

Ancora più complicato: possiamo avere anche due somme una dentro l'altra:

$$\sum_{i=1}^{4} \left(\sum_{j=1}^{4} f_{ij} \right) \text{ che si scrive anche } \sum_{i=1}^{4} \sum_{j=1}^{4} f_{ij}$$

Questo corrispondere a:

$$\sum_{j=1}^{4} f_{1j} + \sum_{j=1}^{4} f_{2j} + \sum_{j=1}^{4} f_{3j} + \sum_{j=1}^{4} f_{4j}$$

quindi complessivamente:

$$\left(f_{11} + f_{12} + f_{13} + f_{14}\right) + \left(f_{21} + f_{22} + f_{23} + f_{24}\right) + \left(f_{31} + f_{32} + f_{33} + f_{34}\right) + \left(f_{41} + f_{42} + f_{43} + f_{44}\right)$$

Quanto vale il risultato?