01 – Descrivere (statisticamente) un fenomeno — Unità n° 02

La descrizione quantitativa di un fenomeno passa attraverso due fasi:

la formazione dei dati | ***la sintesi dei dati***

- La ***formazione del dato*** statistico prevede:
  1) l'osservazione del fenomeno oggetto di studio sulle unità del collettivo
  2) l'annotazione sistematica, unità per unità, della modalità rilevate

Per ogni unità statistica si dispone, in generale, di un'ingente mole di informazioni che occorre organizzare sistematicamente al fine di renderne agevole l'elaborazione

Il processo di raccolta dei dati sulle unità statistiche può essere realizzata ad esempio con la compilazione di questionari

- La ***sintesi dei dati*** avviene attraverso l'uso di strumenti matematico/statistici

02 – Rappresentazione statistica dei dati — Unità n° 02

Quando parliamo di ***rappresentazione statistica*** dei dati stiamo considerando in che modo organizzare i diversi modi di manifestarsi del carattere oggetto di studio nel collettivo

Da un punto di vista formale potremmo usare un foglio di calcolo per rappresentare i diversi dati

Codice intervista (unità statistiche)	Genere	Età	Altezza	Provincia	
1	F	23	156	CS	…
2	F	26	171	NA	…
3	M	23	175	CS	…
4	F	28	163	KR	…
5	M	21	170	KR	…
6	M	24	184	CS	…
7	M	28	178	RC	…
8	F	20	165	CS	…
9	F	19	166	RC	…
10	M	22	180	KR	…
…	…	…	…	…	…

In questa tabella abbiamo allo stesso tempo più caratteri, sia quantitativi che qualitativi

Ogni cella rappresenta l'osservazione di quel carattere per l'unità indicata sulla riga corrispondente

Genericamente si mutua a volte il termine **dataset**

La ***distribuzione statistica*** descrive il modo in cui uno o più caratteri, rappresentativi di un certo fenomeno, si manifestano (secondo la terminologia comune "si distribuiscono") in una popolazione oggetto di studio

03 – Le distribuzioni unitarie

L'elenco delle modalità osservate unità per unità costituisce una **distribuzione unitaria**

A seconda di quanti caratteri statistici prendiamo in considerazione:

un singolo carattere statistico (quantitativo o qualitativo) -> **distribuzione unitaria semplice**

più caratteri statistici (quantitativi o qualitativi) -> **distribuzione unitaria multipla**
[un caso particolare è quello in cui consideriamo solo due caratteri]

A seconda del numero di caratteri studiati dobbiamo utilizzare approcci diversi per analizzare il fenomeno: si parla di **statistica univariata** per un solo carattere, di **statistica bivariata** per due caratteri, di **statistica multivariata** per più caratteri

in questo corso si approfondiranno alcuni strumenti della statistica univariata e bivariata

04 – Esempio

In questa tabella è possibile osservare, per ciascun Corso di Laurea della Facoltà di Economia, quanti sono gli studenti attivi iscritti nell'A.A. 2011/2012:

Corso di Laurea della Facoltà di Economia	N° Iscritti 2011/2012
GIURISPRUDENZA	1662
DISCIPLINE ECONOMICHE E SOCIALI PER LO SVILUPPO	214
ECONOMIA	549
ECONOMIA AZIENDALE	1148
STATISTICA PER LE AZIENDE E LE ASSICURAZIONI	139
SCIENZE TURISTICHE	467
VALORIZZAZIONE DEI SISTEMI TURISTICO CULTURALI	68
ECONOMIA AZIENDALE	403
ECONOMIA APPLICATA	100
DISCIPLINE ECONOMICHE E SOCIALI PER LO SVILUPPO E LA COOPERAZIONE	49
STATISTICA E INFORMATICA PER L'AZIENDA E LA FINANZA	50

→ *Qual è il collettivo oggetto di studio?*

→ *Qual è il carattere oggetto di studio?*

05 – Serie storiche e territoriali

Unità n° 02

È possibile considerare un particolare tipo di distribuzione unitaria, comunemente utilizzata per studiare in che modo un certo fenomeno si è evoluto/manifestato in tempi o luoghi differenti

Se il nostro riferimento è il tempo parliamo di **serie storica**: rappresenta l'evoluzione temporale di un carattere quantitativo (ad es. il prezzo di un prodotto rilevato anno per anno in un dato arco di tempo, le quotazioni di un titolo nelle diverse giornate di contrattazione in Borsa, ecc.)

Se il nostro riferimento è lo spazio parliamo di **serie territoriale**: rappresenta la manifestazione di un carattere quantitativo in un dato istante in luoghi differenti (ad es. il numero di turisti in un dato anno rilevati per ogni regione d'Italia, il numero di addetti dei diversi stabilimenti posseduti da una certa industria, ecc.)

Per studiare tali distribuzioni esistono degli strumenti specifici, ma in generale per descrivere i dati è possibile utilizzare anche tutti gli strumenti tipici della Statistica Descrittiva

06 – Esempio

Nella tabella seguente sono riportati i *prezzi medi nazionali al consumo*, in Euro per Litro, della benzina senza piombo dal 1996 al 2011 (fonte: Ministero dello Sviluppo Economico)

ANNO	€/L.
1996	0.92
1997	0.94
1998	0.91
1999	0.96
2000	1.08
2001	1.05
2002	1.05
2003	1.06
2004	1.12
2005	1.22
2006	1.28
2007	1.30
2008	1.38
2009	1.23
2010	1.36
2011	1.55

Qual è il collettivo?

07 – Ordinamento delle modalità

Unità n° 02

Quando si rilevano caratteri qualitativi ordinabili o quantitativi talvolta risulta essere utile un **ordinamento** delle modalità in senso crescente, dall'unità che ha manifestato con minore intensità il fenomeno studiato a quella che lo ha invece manifestato con maggiore intensità (o seguendo la gerarchia o la logica degli attributi dal meno importante al più importante)

x_1	x_2	x_3	…	x_{i-1}	x_i	x_{i+1}	…	x_{N-1}	x_N

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	…	$x_{(i-1)}$	$x_{(i)}$	$x_{(i+1)}$	…	$x_{(N-1)}$	$x_{(N)}$

Nel primo caso x_1 rappresenta la modalità osservata sulla prima delle unità del collettivo, nel secondo caso $x_{(1)}$ rappresenta la modalità con intensità più bassa osservata nel collettivo (o quella gerarchicamente o logicamente inferiore rispetto a tutte le altre)

8 – Da distribuzioni di quantità a distribuzioni di variazioni

Unità n° 02

Nelle serie storiche è possibile operare una trasformazione dei dati che consente di leggere in modo diverso il fenomeno indagato: è infatti possibile esprimere le quantità osservate in termini di variazioni relative (o di tassi di variazione) e costruire quindi una distribuzione del carattere *variazione di X tra t e t+1*

x_1	x_2	x_3	...	x_{i-1}	x_i	x_{i+1}	...	x_{N-1}	x_N

↓

-	$\frac{x_2}{x_1}$	$\frac{x_3}{x_2}$	...	$\frac{x_{i-1}}{x_{i-2}}$	$\frac{x_i}{x_{i-1}}$	$\frac{x_{i+1}}{x_i}$	...	$\frac{x_{N-1}}{x_{N-2}}$	$\frac{x_N}{x_{N-1}}$

Bisogna fare attenzione al fatto che il numero di unità statistiche (e di conseguenza delle modalità osservate) è passato da N a N-1, e che si tratta adesso di *coppie* di tempi

9 – Esempio

Consideriamo i *prezzi medi nazionali al consumo*, in Euro per Litro, della benzina senza piombo dal 1996 al 2011, e costruiamo la distribuzione delle variazioni relative di prezzo

ANNO	€/L.
1996	0.92
1997	0.94
1998	0.91
1999	0.96
2000	1.08
2001	1.05
2002	1.05
2003	1.06
2004	1.12
2005	1.22
2006	1.28
2007	1.30
2008	1.38
2009	1.23
2010	1.36
2011	1.55

ANNO	var.rel	var.%
1996/1997	1.02	0.02
1997/1998	0.97	-0.03
1998/1999	1.05	0.05
1999/2000	1.13	0.13
2000/2001	0.97	-0.03
2001/2002	1.00	0.00
2002/2003	1.01	0.01
2003/2004	1.06	0.06
2004/2005	1.09	0.09
2005/2006	1.05	0.05
2006/2007	1.02	0.02
2007/2008	1.06	0.06
2008/2009	0.89	-0.11
2009/2010	1.11	0.11
2010/2011	1.14	0.14

10 – Distribuzioni di frequenza

Unità n° 02

Quando abbiamo una popolazione molto numerosa può non essere conveniente avere una lunga lista con tutte le modalità osservate sulle diverse unità statistiche

Nell'osservare ad es. un collettivo di aziende proviamo a rispondere alle seguenti domande:

- *Qual è l'assetto societario più frequente?*
- *Qual è la percentuale di aziende che hanno un n° di addetti inferiore a 15?*

In tali casi è necessario utilizzare una rappresentazione dei dati più "compatta", nota come **distribuzione di frequenza**: per rappresentare i dati innanzi tutto si deve costruire un elenco di tutte le modalità che sono state osservate nel collettivo, quindi contare su quante unità statistiche abbiamo osservato una ad una le specifiche modalità

Dobbiamo però prendere in considerazione un carattere alla volta: se la nostra distribuzione unitaria multipla è composta da 4 caratteri è necessario costruire per ciascuno una separata distribuzione di frequenza

11 – Notazione

Una distribuzione di frequenza per un carattere con *k* modalità distinte si presenta in forma tabellare come un elenco delle diverse modalità e delle corrispondenti frequenze

X	**n**
x_1	n_1
x_2	n_2
...	...
x_i	n_i
...	...
x_k	n_k
totale	**N**

$\mathbf{x_i}$ è la generica modalità **i** del carattere **X** (con i=1,2,…,k)
$\mathbf{n_i}$ è la i-esima frequenza, corrispondente alla modalità $\mathbf{x_i}$

La frequenza può essere letta indifferentemente come:
1) il numero di volte che la modalità è stata rilevata sul collettivo
2) il numero di unità statistiche che presentano la stessa modalità

$$N=\sum_{i=1}^{k} n_i = n_1 + n_2 + \ldots + n_i + \ldots + n_k$$

Si legge "sommatoria per i che va da 1 a k di n con i"

12 – Esempio

Su un collettivo costituito da 50 famiglie è stato rilevato il carattere *numero di figli*, ottenendo la seguente distribuzione unitaria semplice

3 1 3 2 2 0 2 1 5 4 2 2 3 1 1 2 2 0 2 1 4 2 1 2 1
4 3 2 1 3 0 4 3 2 0 3 2 2 1 2 3 1 0 2 2 1 2 2 1 3

Numero di figli	Numero di famiglie
0	5
1	12
2	19
3	9
4	4
5	1
Totale	50

Modalità

Frequenze

Numerosità del collettivo

13 – Esempio

In questa tabella è possibile osservare, per ciascuno studente iscritto nell'A.A. 2011/2012 alla Facoltà di Economia, il Corso di Laurea frequentato:

Corso di Laurea della Facoltà di Economia	N° Iscritti 2011/2012
GIURISPRUDENZA	1662
DISCIPLINE ECONOMICHE E SOCIALI PER LO SVILUPPO	214
ECONOMIA	549
ECONOMIA AZIENDALE	1148
STATISTICA PER LE AZIENDE E LE ASSICURAZIONI	139
SCIENZE TURISTICHE	467
VALORIZZAZIONE DEI SISTEMI TURISTICO CULTURALI	68
ECONOMIA AZIENDALE	403
ECONOMIA APPLICATA	100
DISCIPLINE ECONOMICHE E SOCIALI PER LO SVILUPPO E LA COOPERAZIONE	49
STATISTICA E INFORMATICA PER L'AZIENDA E LA FINANZA	50

- *Qual è il collettivo oggetto di studio?*
- *Qual è il carattere oggetto di studio?*
- *Che tipo di distribuzione abbiamo?*

14 – Migliorare la leggibilità dei dati

Abbiamo visto come in presenza di tantissime osservazioni sia conveniente trasformare la lista dei dati, che abbiamo chiamato distribuzione unitaria (semplice per un solo carattere, multipla per più caratteri), in una distribuzione di frequenza

In tal modo abbiamo una rappresentazione compatta di tutte i dati raccolti: qual è il carattere, quali sono le modalità, quanto grande è il collettivo, quante unità statistiche hanno presentato ciascuna delle diverse modalità

Questa rappresentazione in taluni casi non è sufficiente. Supponiamo infatti di considerare per due collettivi la distribuzione di frequenza del carattere Genere:

Genere	**n**
Maschile	?
Femminile	55

Genere	**n**
Maschile	?
Femminile	55

Dalla lettura delle due tabelle vediamo come nei due collettivi ci sia lo stesso n° di soggetti di genere femminile: la modalità *Femminile* ha la stessa importanza nel descrivere come si è manifestato il fenomeno?

15 – Frequenze assolute e relative — Unità n° 02

Se nella lettura della tabella precedente non teniamo conto della diversa numerosità del collettivo siamo portati a credere che l'importanza del genere femminile sia esattamente la stessa quando invece non è così

Dobbiamo allora considerare il numero di unità statistiche che presentano una certa modalità in rapporto alla dimensione del collettivo che stiamo esaminando

Il conteggio delle unità statistiche che presentano una certa modalità, che abbiamo indicato come frequenza, è definito più correttamente come **frequenza assoluta**, per distinguerlo dal conteggio delle unità statistiche che presentano una certa modalità del carattere in relazione alla numerosità del collettivo, detto **frequenza relativa**

I dati riportati nella tabella (distribuzione di frequenza con frequenze assolute) sono trasformati dividendo ciascuna frequenza assoluta per la numerosità del collettivo, ottenendo così una nuova rappresentazione dei dati (distribuzione di frequenza con frequenze relative)

In tal modo nella lettura dei dati possiamo indicare anche il peso che quella modalità ha nella descrizione del collettivo rispetto al carattere che ci interessa

16 – Notazione

Una distribuzione di frequenza per un carattere con *k* modalità distinte si presenta in forma tabellare come un elenco delle diverse modalità e delle corrispondenti frequenze relative

X	**f**
x_1	f_1
x_2	f_2
...	...
x_i	f_i
...	...
x_k	f_k
totale	**1**

$\mathbf{x_i}$ è la generica modalità **i** del carattere **X** (con i=1,2,…,k)

$\mathbf{f_i}$ è la i-esima frequenza relativa, corrispondente alla modalità $\mathbf{x_i}$

$$f_i = \frac{n_i}{N} \qquad (i=1,2,\ldots,k) \qquad 0 \leq f_i \leq 1$$

La frequenza relativa indica:

1) *l'importanza della i-esima modalità nel collettivo studiato (in termini relativi)*
2) *il numero di unità statistiche che presentano la stessa modalità rispetto alla dimensione del collettivo*

$$\sum_{i=1}^{k} f_i = f_1 + f_2 + \ldots + f_i + \ldots + f_k = 1$$

17 – Esempio

Consideriamo di voler studiare il n° di figli per famiglia in un collettivo di 50 famiglie, indicato con **A**, e in un secondo collettivo di 100 famiglie, indicato con **B**

COLLETTIVO A

X	n
0	5
1	12
2	19
3	9
4	4
5	1
Totale	50

COLLETTIVO B

X	n
0	20
1	10
2	35
3	15
4	10
5	10
Totale	100

- *È corretto affermare che le famiglie con un figlio sono più importanti nel collettivo A che in B?*

- *È corretto affermare che le famiglie con due figli sono più importanti nel collettivo A che in B?*

- *In A le famiglie con un figlio sono lo 0,24 del totale delle famiglie, mentre in B sono lo 0,10. Le famiglie con un solo figlio sono più numerose in A*

- *In A le famiglie con due figli sono lo 0,38 del totale delle famiglie, mentre in B sono lo 0,35. Le famiglie con due figli sono più numerose in A*

	A		B	
X	n	f	n	f
0	5	0,10	20	0,20
1	12	0,24	10	0,10
2	19	0,38	35	0,35
3	9	0,18	15	0,15
4	4	0,08	10	0,10
5	1	0,02	10	0,10
Totale	50	1	100	1

18 – Frequenze percentuali

È possibile una ulteriore trasformazione dei dati che ne facilita la comprensione e li rende di fatto fruibili ad un pubblico più ampio degli "addetti ai lavori "

Moltiplicando le frequenze relative per cento si rappresentano i dati in termini **percentuali**: è come se stessimo considerando un ipotetico collettivo di 100 unità statistiche nel quale il numero di unità che presentano una specifica modalità è proporzionale a quello che si è osservato nella realtà e sul quale abbiamo rilevato i dati

Sesso	**Fr. relativa**	**%**
Maschile	45/100=0,45	45%
Femminile	55/100=0,55	55%
Totale	100/100=1	100%

Sesso	**Fr. relativa**	**%**
Maschile	75/130=0,58	58%
Femminile	55/130=0,42	42%
Totale	130/130=1	100%

Il 45% dei soggetti del primo collettivo è di sesso maschile, a fronte di un 55% di sesso femminile. Nel secondo le percentuali sono rispettivamente del 58% e del 42%

Dalla lettura delle due tabelle si ricava che il peso relativo dei soggetti di sesso maschile è più basso nel secondo collettivo: se questo fosse stato composto da 100 unità come il primo avremmo dovuto osservato solo 42 unità invece di 55 per avere la stessa importanza

19 – Notazione

Una distribuzione di frequenza per un carattere con *k* modalità distinte si presenta in forma tabellare come un elenco delle diverse modalità e delle corrispondenti frequenze relative

X	**p**
x_1	p_1
x_2	p_2
...	...
x_i	p_i
...	...
x_k	p_k
totale	**100**

$\mathbf{x_i}$ è la generica modalità **i** del carattere **X** (con i=1,2,…,k)

$\mathbf{p_i}$ è la i-esima frequenza percentuale, corrispondente alla modalità $\mathbf{x_i}$

$$p_i = \frac{n_i}{N} \times 100 = f_i \times 100 \quad (i=1,2,...,k) \quad 0 \leq p_i \leq 100$$

La frequenza percentuale indica:

1) *l'importanza della modalità in un ipotetico collettivo di 100 unità*
2) *il numero di unità statistiche che presentano la stessa modalità in un ipotetico collettivo di 100 unità statistiche*

$$\sum_{i=1}^{k} p_i = p_1 + p_2 + ... + p_i + ... + p_k = 100$$

20 – Esercizio

È stato effettuato un sondaggio tra i consumatori di una marca di succo di frutta per rilevare il gusto preferito. Di seguito sono riportate le preferenze registrate:

PERA PESCA ALBICOCCA ANANAS MELA ARANCIA MELA PESCA
ALBICOCCA MELA PESCA MELA PESCA ALBICOCCA ANANAS PERA
PESCA ALBICOCCA ANANAS PERA ALBICOCCA PESCA ARANCIA PERA
PERA ANANAS MELA PERA ALBICOCCA MELA ARANCIA PERA MELA
ARANCIA ALBICOCCA PERA PESCA MELA ANANAS ARANCIA PESCA
PESCA ARANCIA PERA PERA ANANAS MELA PERA ALBICOCCA MELA

1) Indicare inoltre qual è il collettivo, la sua numerosità, la natura del carattere studiato

2) Costruire la distribuzione di frequenza del carattere “gusto preferito”

3) Quanti sono i consumatori che preferiscono il succo di frutta alla pera?

21 – Distribuzione in classi

Quando si analizza un fenomeno che può essere espresso per mezzo di un carattere discreto con numerose modalità (ES. età in anni compiuti), oppure quando si usano caratteri continui (ES. peso, fatturato), è possibile che le distribuzioni di frequenza assolute o relative non siano idonee e non migliorino la comprensione dei dati

In questi casi può essere adoperata un'altra rappresentazione dei dati: le modalità (discrete o continue) sono organizzate in intervalli di valori detti **classi**, e le frequenze associate a ciascun intervallo rappresentano il n° di unità sulle quali è osservato/misurato un valore appartenente all'intervallo stesso

Bisogna dire che se la rappresentazione in classi presenta la stessa facilità di lettura di una qualsiasi distribuzione di frequenza (assoluta o relativa) non è però altrettanto immediata e di facile costruzione a partire dalla distribuzione unitaria dei dati. È infatti necessario tenere in considerazione diversi elementi: il numero di classi adeguato al problema, l'ampiezza delle diverse classi, la possibilità di includere tutte le modalità del carattere, e così via…

22 – Determinazione del numero di classi

Non esiste un modo univoco per determinare il numero di classi: molte volte la scelta è lasciata all'esperienza di chi effettua lo studio in base alla natura del fenomeno d'interesse

La regola da seguire è che non bisogna scegliere un numero di classi eccessivamente piccolo per non perdere dettaglio nella rappresentazione del fenomeno, ma allo stesso tempo non bisogna scegliere un n° di classi eccessivamente grande e "sacrificare" la leggibilità della distribuzione

Nel corso degli anni sono state proposte diverse soluzioni per determinare in modo oggettivo il numero di classi ideale per una popolazione di numerosità pari a N: una possibile soluzione è quella di considerare il numero k di classi ottenuto dalla **formula di Sturges**

$$k \cong 1 + 3{,}322 \log_{10}(N)$$

Operativamente, il n° di classi ritenuto adeguato è compreso tra almeno 4-5 e non più di 15-20

23 – Ampiezza delle classi (1) — Unità n° 02

L'ampiezza delle classi può essere sempre costante oppure di volta in volta differente: nel primo caso si parla di classi **equiampie**, nel secondo caso si parla di classi **non equiampie**

La scelta di un tipo dipende dalle scelte soggettive del ricercatore e dalla natura del fenomeno

Nel caso in cui si considerino classi di ampiezza diversa bisogna chiaramente procedere ad una scelta coerente con il fenomeno che si sta analizzando

Età	**Criterio**
Da 0 a 5 anni	*Età prescolare*
Da 6 a 10 anni	*Scuola elementare*
Da 11 a 13 anni	*Scuola media*
Da 14 a 18 anni	*Scuola superiore*
…	…

In questo caso la suddivisione in classi del carattere è dettata da un criterio esterno che fornisce comunque un interessante punto di vista rispetto al fenomeno che si sta studiando

24 – Ampiezza delle classi (2)

Se invece si considerano classi di ampiezza uguale allora è necessario trovare un modo per determinare in modo pratico e veloce la quantità che si assume costante per ogni intervallo

Tale quantità può essere ottenuta facilmente considerando l'ampiezza della distribuzione, a partire dalla differenza della modalità più grande e della modalità più piccola osservata nella distribuzione unitaria dei dati e dividendo per il numero di classi definito precedentemente:

$$\frac{x_{(N)} - x_{(1)}}{k} \cong \omega$$

La lettera omega dell'alfabeto greco è utilizzata per convenzione per indicare l'ampiezza della classe: va chiaramente approssimata al numero intero più vicino

25 – Classi aperte e classi chiuse

È possibile parlare di classi **aperte** o **chiuse** a seconda che gli estremi siano inclusi o meno nell'intervallo: la modalità più piccola della classe è detta *estremo inferiore*, la modalità più grande è detta invece *estremo superiore*

Se l'estremo inferiore è incluso nello classe mentre non lo è quello superiore allora si parla di classe **chiusa a sinistra e aperta a destra**; se invece l'estremo inferiore della classe non è incluso nella classe mentre lo è quello superiore si parla di classe **aperta a sinistra e chiusa a destra**. Se includiamo sia l'estremo inferiore che superiore allora parliamo genericamente di classe **chiusa**: questo tipo di classi è però idoneo per rappresentare i soli caratteri discreti

La scelta di includere o meno uno degli estremi è univoca: se decidiamo che la prima classe della distribuzione è chiusa a sinistra e aperta a destra (o viceversa), allora tutte le classi della distribuzione saranno dello stesso tipo

Un particolare tipo di classi sono quelle **non limitate inferiormente o superiormente**: in tal caso si utilizza la notazione matematica < (minore di) e > (maggiore di), oppure si ricorre ad esempio a locuzioni del tipo "fino a" (<) o "più di" (>)

26 – Notazione (1)

In generale una distribuzione in classi per un carattere con *k* classi distinte si presenta come:

X	**n**
$x_0 - x_1$	n_1
$x_1 - x_2$	n_2
...	...
$x_{i-1} - x_i$	n_i
...	...
$x_{k-1} - x_k$	n_k
totale	**N**

$\mathbf{x_{i-1}\text{-}x_i}$ è la generica i-esima classe di modalità del carattere X
$\mathbf{n_i}$ è la frequenza corrispondente alla classe $\mathbf{x_{i-1}\text{-}x_i}$

La frequenza indica in modo equivalente:
1) il numero di volte che la classe di modalità è stata rilevata sul collettivo
2) il numero di unità statistiche che appartengono alla data classe

→ Analogamente a quanto visto è possibile calcolare per ogni classe sia le frequenze relative sia le frequenze percentuali

N.B.: le classi non devono mai essere vuote (cioè con 0 unità statistiche)

27 – Notazione (2)

Unità n° 02

In generale per indicare se una classe è aperta o chiusa a destra o a sinistra si utilizza la seguente notazione:

x_{i-1} -| x_i oppure **$(x_{i-1}$, $x_i]$** -> la classe è chiusa a destra e aperta a sinistra (le unità che presentano x_{i-1} non sono incluse nella classe, quelle che presentano x_i invece lo sono)

x_{i-1} |- x_i oppure **$[x_{i-1}$, $x_i)$** -> la classe è aperta a destra e chiusa a sinistra (le unità che presentano x_{i-1} sono incluse nella classe, quelle che presentano x_i invece non lo sono)

x_{i-1} - x_i oppure **$[x_{i-1}$, $x_i]$** -> la classe è chiusa a destra e sinistra (sia le unità con x_{i-1} che quelle che presentano x_i sono incluse nella classe)

Classe $x_i - \mid x_{i+1}$	n_i	*Ampiezza* a_i
70 -| 100	20	30
100 -| 120	7	20
120 -| 140	18	20
140 -| 170	65	30
170 -| 180	21	10
180 -| 200	45	20
200 -| 220	24	20
Totale	**200**	

28 – Rappresentazione dei dati

Una volta ottenute le classi e "contate" quante sono le unità statistiche appartenente ad ogni classe abbiamo di fatto ottenuto una distribuzione di frequenza assoluta, con la differenza che non abbiamo tutte le modalità osservate ma intervalli di modalità

In tutti i casi in cui è necessario effettuare delle operazioni sulle distribuzioni in classe risulta difficile ritornare ad una distribuzione di frequenza o unitaria: a tal scopo per convenzione si fa riferimento ad un valore rappresentativo dell'intera classe, detto **valore centrale**, calcolato dalla semisomma degli estremi inferiore e superiore di ciascuna classe

$$\frac{\text{estr. inferiore} + \text{estr. superiore}}{2} = \text{valore centrale}$$

29 – Ampiezza costante e valore centrale

Una volta definito il numero delle classi e l'ampiezza di ciascuna di esse per ottenere gli estremi inferiore e superiore di ciascuna di esse si procede come segue: innanzi tutto si ordinano tutte le modalità in senso crescente, dalla più piccola alla più grande

1ª classe -> $x_0 -| x_1 = \quad x_0 -| x_0 + \omega$

2ª classe -> $x_1 -| x_2 = \quad x_0 + \omega -| x_0 + 2\omega$ oppure $x_1 -| x_1 + \omega$

3ª classe -> $x_2 -| x_3 = x_0 + 2\omega -| x_0 + 3\omega$ oppure $x_2 -| x_2 + \omega$

…

In generale -> $x_{i-1} -| x_i = x_0 + (i-1)\cdot\omega -| x_0 + i\cdot\omega$

Quindi una volta individuato l'estremo inferiore è possibile ottenere l'estremo superiore della classe aggiungendo la quantità relativa all'ampiezza

Per calcolare il valore centrale di ciascuna classe è sufficiente aggiungere all'estremo inferiore delle classi la metà dell'ampiezza **$\omega/2$**

30 – Esempio

Consideriamo la distribuzione unitaria di un carattere X per un collettivo formato da 200 unità statistiche

81,46	73,02	87,89	96,97	96,54	98,75	70,43	82,22	95,90	74,16
199,43	200,17	190,08	196,08	204,52	209,08	208,04	196,49	195,45	194,70
183,18	174,61	168,53	169,05	175,86	167,22	144,52	161,84	146,02	128,10
182,67	159,04	139,88	197,58	207,99	166,99	149,72	140,62	139,87	173,42
176,50	183,91	158,27	121,00	157,38	176,95	187,96	177,16	164,94	171,75
203,27	198,59	200,71	199,29	191,21	195,01	207,38	201,73	205,98	196,20
165,69	117,11	184,28	147,32	154,99	141,96	200,60	157,27	140,33	144,39
208,43	128,65	181,60	145,80	141,88	127,86	199,38	199,03	165,53	190,84
165,00	161,63	166,90	163,46	174,68	185,09	185,24	186,48	158,18	142,17
128,92	119,61	155,29	178,83	168,23	147,93	112,49	128,74	163,55	121,86
77,35	71,36	70,97	74,92	76,59	70,51	78,55	80,29	86,61	80,72
176,85	179,01	165,26	171,93	213,43	216,30	181,64	154,06	177,58	162,62
163,94	166,20	177,60	165,01	128,75	201,33	162,90	170,66	156,95	201,21
199,69	147,06	155,00	167,72	179,37	156,51	208,00	197,84	158,19	212,91
188,48	165,99	215,25	183,18	129,08	116,86	153,66	133,90	189,07	174,83
192,04	208,89	203,64	198,20	203,88	191,26	208,52	190,57	196,71	209,99
192,21	138,04	147,00	172,53	169,92	167,42	139,43	150,04	139,08	196,55
149,78	178,11	181,38	194,63	157,36	163,88	195,21	167,63	162,88	119,97
155,16	144,50	144,12	123,98	188,78	166,56	188,45	186,68	169,16	172,41
126,57	146,26	161,36	114,21	123,79	190,42	184,53	170,87	107,27	169,40

non equiampie

X	n	ω
70 -l 100	20	30
100 -l 120	7	20
120 -l 140	18	20
140 -l 170	65	30
170 -l 180	21	10
180 -l 200	45	20
200 -l 220	24	20
Totale	**200**	

equiampie

X	n	ω
70 -l 85	14	15
85 -l 100	6	15
100 -l 115	3	15
115 - l 130	16	15
130 -l 145	15	15
145 -l 160	25	15
160 -l 175	41	15
175 -l 190	29	15
190 -l 205	37	15
205 -l 220	14	15
Totale	**200**	

31 – Frequenze cumulate

Unità n° 02

Nel caso in cui le modalità del carattere in esame sono ordinate può essere interessante studiare la frequenza con cui si presentano nel collettivo in esame modalità inferiori o uguali ad un certa soglia. Le **frequenze cumulate** sono utili quando vogliamo fissare una delle modalità e leggere i dati della distribuzione rispetto a questa

Ricarica telefonica	frequenza assoluta	fr. assoluta cumulata
10	10	10
50	6	16
100	5	21
Totale	21	-

Se vogliamo sapere quanti individui hanno acquistato una ricarica con un taglio inferiore o uguale ad una certa soglia basta leggere la frequenza cumulata in corrispondenza della modalità che ci interessa: ad es. se vogliamo il numero di unità statistiche che hanno ricaricato massimo (al più) 50 € (minore o uguale) è pari a 16 (10+6)

Se vogliamo sapere quanti individui hanno acquistato una ricarica con un taglio inferiore a una certa soglia basta leggere la frequenza cumulata della modalità precedente a quella che ci interessa: ad es. le unità che hanno ricaricato meno di 50 € sono 10

32 – Notazione

È possibile calcolare le frequenze cumulate a partire dalle frequenze assolute, relative o percentuali. Per distinguere le frequenze cumulate vengono indicate con la lettera maiuscola corrispondente

X	N	F	P
x_1	N_1	F_1	P_1
x_2	N_2	F_2	P_2
...	...	...	...
x_i	N_i	F_i	P_i
...	...	...	...
x_k	N_k	F_k	P_k

x_i è la generica i-esima modalità del carattere X (con i=1,2,…,k)

N_i è la i-esima frequenza assoluta cumulata delle prime i modalità
F_i è la i-esima frequenza relativa cumulata delle prime i modalità
P_i è la i-esima frequenza percentuale cumulata delle prime i modalità

$$N_i = \sum_{h=1}^{i} n_h \quad , \quad i = 1,2,...,k \qquad F_i = \sum_{h=1}^{i} f_h \quad , \quad i = 1,2,...,k$$

$$P_i = \sum_{h=1}^{i} p_h \quad , \quad i = 1,2,...,k$$

33 – Esercizio

Il responsabile del settore personale del Comune di Cosenza conosce la distribuzione degli impiegati secondo la qualifica funzionale

Qualifica	Impiegati
I	58
II	308
III	287
IV	71
V	52
VI	28
VII	12
	816

Il Comune ha bandito un concorso per quattro posti riservati agli interni con qualifica non inferiore alla V

Qual è la percentuale dei possibili candidati al concorso?

Qual è il collettivo statistico e qual è il carattere oggetto di studio?

Come traduciamo in termini statistici il quesito del responsabile del personale?

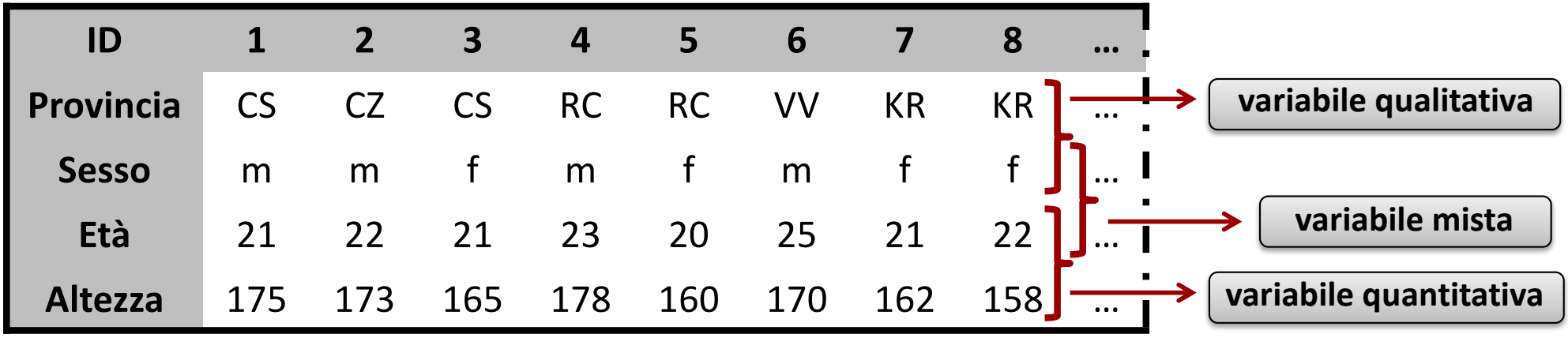
34 – Lavorare con più variabili

Unità n° 02

Sono molti i casi in cui è possibile osservare o misurare su ognuna delle unità statistica di un collettivo più di una variabile contemporaneamente => si parla di **distribuzioni multiple**

Quando studiamo congiuntamente due variabili statistiche si parla in generale di ***variabile doppia*** e quindi conseguentemente di una ***distribuzione unitaria doppia***

ID	1	2	3	4	5	6	7	8	…
Provincia	CS	CZ	CS	RC	RC	VV	KR	KR	…
Sesso	m	m	f	m	f	m	f	f	…
Età	21	22	21	23	20	25	21	22	…
Altezza	175	173	165	178	160	170	162	158	…



35 – Notazione

La distribuzione unitaria doppia per i caratteri X e Y può essere vista come un insieme di N coppie di modalità osservate congiuntamente sulle unità del collettivo oggetto di studio

X	**Y**
x_1	y_1
x_2	y_2
...	...
x_i	y_i
...	...
x_N	y_N

$\mathbf{x_i}$ è la generica **i-ma** modalità del carattere **X** (con i=1,2,…,N)
$\mathbf{y_i}$ è la generica **i-ma** modalità del carattere **Y** (con i=1,2,…,N)

$\mathbf{(x_i,y_j)}$ è la coppia di modalità che osserviamo per ciascuna unità del collettivo

Poiché i due caratteri sono legati insieme nella distribuzione non è possibile effettuare contemporaneamente un ordinamento di X e Y: qualora ciò sia necessario dovrà essere effettuato una volta per variabile (per non perdere il riferimento all'unità corrispondente)

36 – Distribuzioni doppie di frequenza

NOME	GENERE	REGIONE
M. Rossi	M	Marche
A. Bianchi	F	Calabria
A. Franchi	F	Umbria
B. Gini	M	Piemonte
A. Grandi	F	Marche
P. Lini	F	Umbria

		Regione			
		Calabria	Marche	Umbria	Piemonte
Sesso	M	0	1	0	1
	F	1	1	2	0

Possiamo rappresentare le distribuzioni unitarie doppie come **distribuzioni doppie di frequenza** contando le unità statistiche che presentano contemporaneamente una modalità di una variabile e una modalità dell'altra variabile considerata

ESEMPIO

Anno 1991	Gruppo di corsi di laurea			
Condizione occupazionale	**gruppo medico**	**gruppo economico**	**gruppo letterario**	**Totale**
occupati stabilmente	6.816	7.328	7.705	21.849
occupati precariamente	4.666	720	5.858	11.244
non lavorano	1.183	181	1.476	2.840
Totale	12.665	8.229	15.039	35.933

37 – Notazione

TABELLA A DOPPIA ENTRATA

	y_1	…	y_j	…	y_c	TOT
x_1	n_{11}	…	n_{1j}	…	n_{1c}	$n_{1.}$
…	…	…	…	…	…	…
x_i	n_{i1}	…	n_{ij}	…	n_{ic}	$n_{i.}$
…	…	…	…	…	…	…
x_r	n_{r1}	…	n_{rj}	…	n_{rc}	$n_{r.}$
TOT	$n_{.1}$	…	$n_{.j}$	…	$n_{.c}$	$n_{..}$

distribuzione marginale di riga

distribuzione marginale di colonna

$n_{..}$ → gran totale

Consideriamo una variabile doppia **(X,Y)** e supponiamo che sia stata organizzata in una tabella: sulle righe le **r** modalità di X e sulle colonne le **c** modalità di Y

$$n_{i.} = \sum_{j=1}^{c} n_{ij} = n_{i1} + n_{i2} + \ldots + n_{ic}$$

FREQ. MARGINALE DI RIGA

$$n_{.j} = \sum_{i=1}^{r} n_{ij} = n_{1j} + n_{2j} + \ldots + n_{rj}$$

FREQ. MARGINALE DI COLONNA

Su ciascuna delle righe abbiamo la distribuzione semplice di ogni modalità *i* della X rispetto a tutte quelle della variabile Y; su ciascuna colonna abbiamo la distribuzione semplice di ogni modalità *j* della Y rispetto a tutte quelle della variabile X

L'elemento generico n_{ij} rappresenta il numero di unità che presentano allo stesso tempo le modalità x_i e y_j e viene detto *frequenza congiunta*

38 – Esempio

Consideriamo la distribuzione doppia di frequenza del ***tipo di birra*** preferito e dell'***età*** di un collettivo di consumatori

		Età			
		18 - 22	*23 - 26*	*27 - 30*	
Tipo di Birra	*Bionda*	12	22	11	**45**
	Rossa	5	9	14	**28**
	Scura	3	15	18	**36**
		20	**46**	**43**	**109**

La frequenza congiunta n_{21} ci dice quanti sono i consumatori tra i 18 e i 22 anni che preferiscono la birra rossa (=> 5)

Il gran totale ci dice qual è la dimensione del collettivo

Birra	
Bionda	**45**
Rossa	**28**
Scura	**36**
	109

Età	
18 - 22	**20**
23 - 26	**46**
27 - 30	**43**
	109

Le distribuzioni marginali rappresentano di fatto le distribuzioni di una variabile per volta: la distr. marginale di riga è la distribuzione di frequenza della variabile birra a prescindere dall'età; allo stesso modo la distr. marginale di colonna è la distribuzione della variabile età a prescindere dal tipo di birra preferito

39 – Trasformazione dei dati: frequenze relative

frequenze assolute

	y_1	…	y_j	…	y_c	TOT
x_1	n_{11}	…	n_{1j}	…	n_{1c}	$n_{1.}$
…	…	…	…	…	…	…
x_i	n_{i1}	…	n_{ij}	…	n_{ic}	$n_{i.}$
…	…	…	…	…	…	…
x_r	n_{r1}	…	n_{rj}	…	n_{rc}	$n_{r.}$
TOT	$n_{.1}$	…	$n_{.j}$	…	$n_{.c}$	$n_{..}$

frequenze relative

	y_1	…	y_j	…	y_c	TOT
x_1	f_{11}	…	f_{1j}	…	f_{1c}	$f_{1.}$
…	…	…	…	…	…	…
x_i	f_{i1}	…	f_{ij}	…	f_{ic}	$f_{i.}$
…	…	…	…	…	…	…
x_r	f_{r1}	…	f_{rj}	…	f_{rc}	$f_{r.}$
TOT	$f_{.1}$	…	$f_{.j}$	…	$f_{.c}$	1

Per ottenere le frequenze relative da una tabella a doppia entrata è sufficiente dividere ogni elemento della matrice per il *gran totale* n :

$$f_{ij} = \frac{n_{ij}}{n}$$ → ***frequenza congiunta relativa***

È sempre un numero tra 0 e 1 → $0 \le f_{ij} \le 1$

frequenze marginali relative →

$$f_{i.} = \sum_{j=1}^{c} f_{ij}$$

$$f_{.j} = \sum_{i=1}^{r} f_{ij}$$

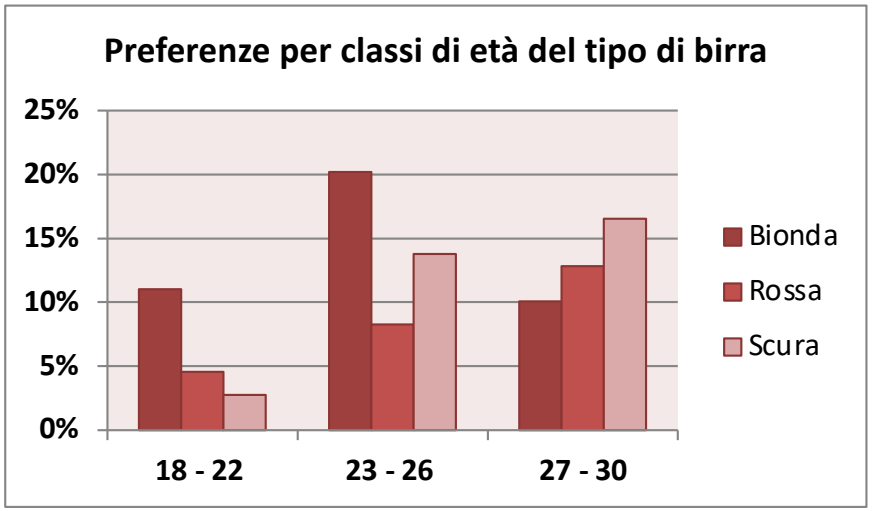
40 – Esempio

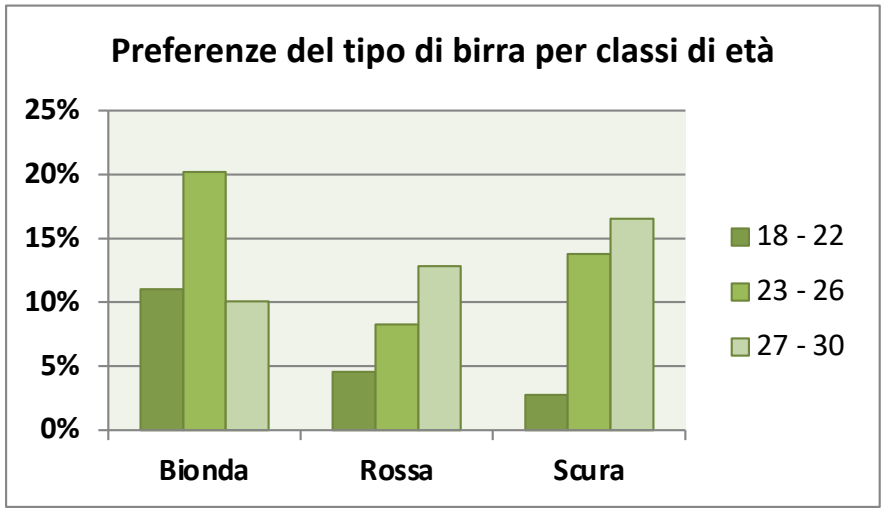
Consideriamo ancora la distribuzione doppia di frequenza del ***tipo di birra*** preferito e dell'***età*** di un collettivo di consumatori, e calcoliamo le frequenze relative

Tipo di Birra	Età: *18 - 22*	*23 - 26*	*27 - 30*	
Bionda	12	22	11	**45**
Rossa	5	9	14	**28**
Scura	3	15	18	**36**
	20	**46**	**43**	**109**

Tipo di Birra	Età: *18 - 22*	*23 - 26*	*27 - 30*	
Bionda	0.11	0.20	0.10	**0.41**
Rossa	0.05	0.08	0.13	**0.26**
Scura	0.03	0.14	0.17	**0.33**
	0.18	**0.42**	**0.39**	**1.00**

Per facilitare la lettura spesso è preferibile esprimere le frequenze relative in termini percentuali





41 – Tabelle doppie e rapporti di composizione

Utilizzando i rapporti di composizione è possibile leggere dalle tabelle doppie altre informazioni:

in particolare, possiamo utilizzare i rapporti per esplorare la tabella, calcolando la percentuale di unità statistiche che hanno una o più modalità della variabile in riga tra quelle che ne hanno una o più della variabile in colonna (e viceversa)

		Età			
		18 - 22	*23 - 26*	*27 - 30*	
Tipo di Birra	*Bionda*	12	22	11	**45**
	Rossa	5	9	14	**28**
	Scura	3	15	18	**36**
		20	**46**	**43**	**109**

(1) ***Tra tutti coloro che hanno meno di 27 anni qual è la percentuale di quelli che non gradiscono la birra scura?***

(2) ***Tra tutti quelli che preferiscono la birra scura e la birra rossa qual è la percentuale di quelli che hanno più di 26 anni?***

I consumatori con meno di 27 anni sono le unità appartenenti alle classi *18 - 22* e *23 - 26* => 20 + 46

Tra questi consumatori, quelli che non preferiscono la birra scura sono 12 + 5 e 22 + 9

La risposta al primo quesito è (12+5+22+9)/(20+46) = 0,73 => 73%

I consumatori che preferiscono la birra scura e quella rossa sono 28 + 36

Tra questi consumatori, hanno più di 26 anni quelli della classe 27 - 30 => 14 + 18

La risposta al secondo quesito è (14+18)/(28+36) = 0,50 => 50%

42 – Trasformazione dei dati: frequenze condizionate

Unità n° 02

	y_1	…	y_j	…	y_c	TOT
x_1	f_{11}	…	f_{1j}	…	f_{1c}	$f_{1.}$
…	…	…	…	…	…	…
x_i	f_{i1}	…	f_{ij}	…	f_{ic}	$f_{i.}$
…	…	…	…	…	…	…
x_r	f_{r1}	…	f_{rj}	…	f_{rc}	$f_{r.}$
TOT	$f_{.1}$	…	$f_{.j}$	…	$f_{.c}$	1

In questo modo otteniamo la cosiddetta **distribuzione condizionata** di Y rispetto alla modalità x_1 di X: questa distribuzione prende il nome di ***profilo riga***; allo stesso modo possiamo ottenere il ***profilo colonna***

Consideriamo una variabile doppia **(X,Y)** e supponiamo di voler studiare la distr. della variabile Y rispetto ad un prefissato valore di $X=x_i$

	y_1	…	y_j	…	y_h	
x_1	f_{11}	…	f_{1j}	…	f_{1c}	$f_{1.}$

$$f(Y|x_1)=\frac{f(X=x_1,Y=y_j)}{f(X=x_1)}=\frac{f_{1j}}{f_{1.}}$$

profilo riga

$$f(Y=y_j|X=x_i)=\frac{f(X=x_i,Y=y_j)}{f(X=x_i)};\quad j=1,2,\ldots,c$$

profilo colonna

$$f(X=x_i|Y=y_j)=\frac{f(X=x_i,Y=y_j)}{f(Y=y_j)};\quad i=1,2,\ldots,r$$

43 – Esempio

Tipo di Birra	Età: 18 - 22	23 - 26	27 - 30	
Bionda	0.11	0.20	0.10	0.41
Rossa	0.05	0.08	0.13	0.26
Scura	0.03	0.14	0.17	0.33
	0.18	0.42	0.39	1.00

Consideriamo di nuovo la distribuzione doppie di frequenza relativa del tipo di birra preferito e dell'età di un collettivo di consumatori

Tipo di Birra	Età: 18 - 22	23 - 26	27 - 30	
Bionda	0.27	0.49	0.24	1.00
Rossa	0.18	0.32	0.50	1.00
Scura	0.08	0.42	0.50	1.00
	0.18	0.42	0.39	1.00

Se dividiamo ogni elemento sulle righe per il totale otteniamo le distribuzioni condizionate $Y|x_i$: mostrano rispetto alle modalità in riga la composizione rispetto alla variabile in colonna

Tipo di Birra	Età: 18 - 22	23 - 26	27 - 30	
Bionda	0.60	0.47	0.26	0.41
Rossa	0.25	0.20	0.33	0.26
Scura	0.15	0.33	0.42	0.33
	1.00	1.00	1.00	1.00

Se dividiamo ogni elemento sulle colonne per il totale otteniamo le distribuzioni condizionate $X|y_j$: mostrano rispetto alle modalità in colonna la composizione rispetto alla variabile in riga