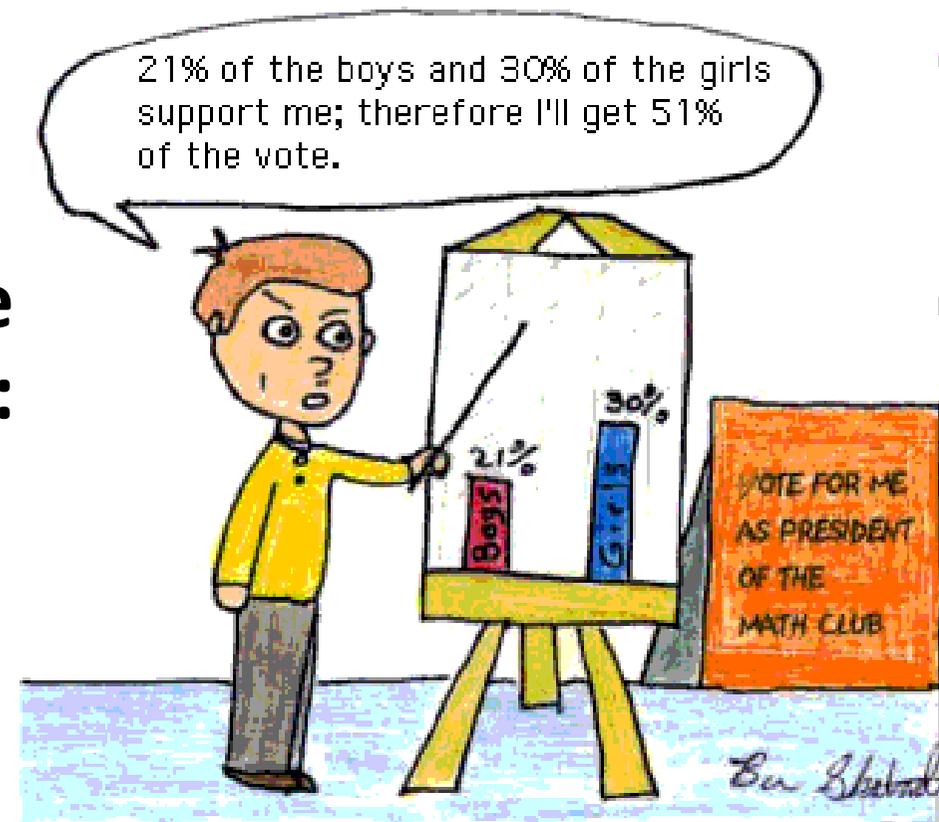


Analisi delle distribuzioni doppie di caratteri di natura qualsiasi: dipendenza



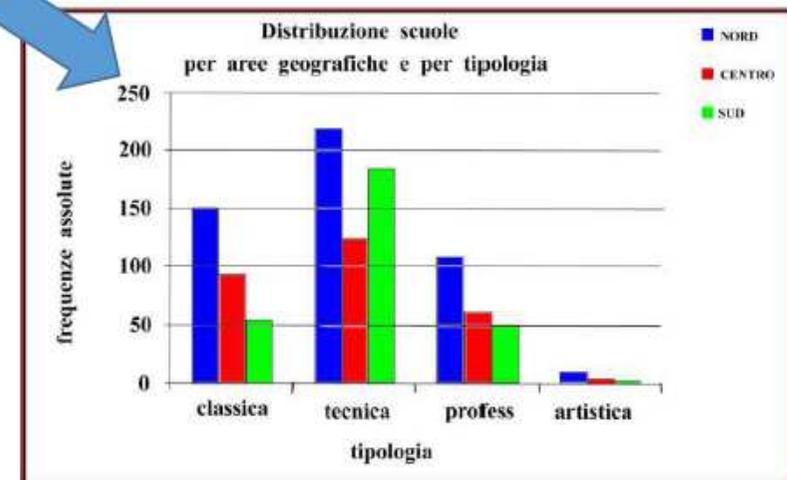
Statistica univariata

Tutto ciò che abbiamo visto fino ad ora aveva a che fare con la **statistica univariata**. E che significa? Significa che **analizzavamo una variabile alla volta**.

Esempi

- ✓ Istogramma peso (1 variabile: peso),
- ✓ diagramma a torta genere (1 variabile: genere),
- ✓ varianza del numero esami sostenuti

Esempio di più variabili analizzate allo stesso tempo



Da 1 a 2 variabili cambiano gli obiettivi

Obiettivi

Univariata

1 sola variabile. Vogliamo:

- 1) Sintetizzarla
- 2) Descriverla
- 3) Grafici maggiormente utilizzati: boxplot-istogrammi, grafici a barre semplici, grafici a torte

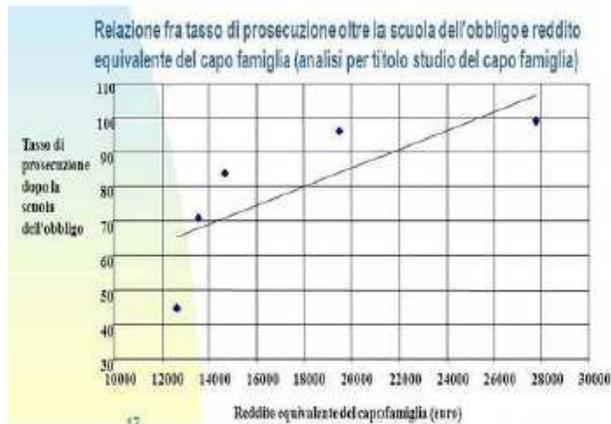
Bivariata

2 variabili. Vogliamo:

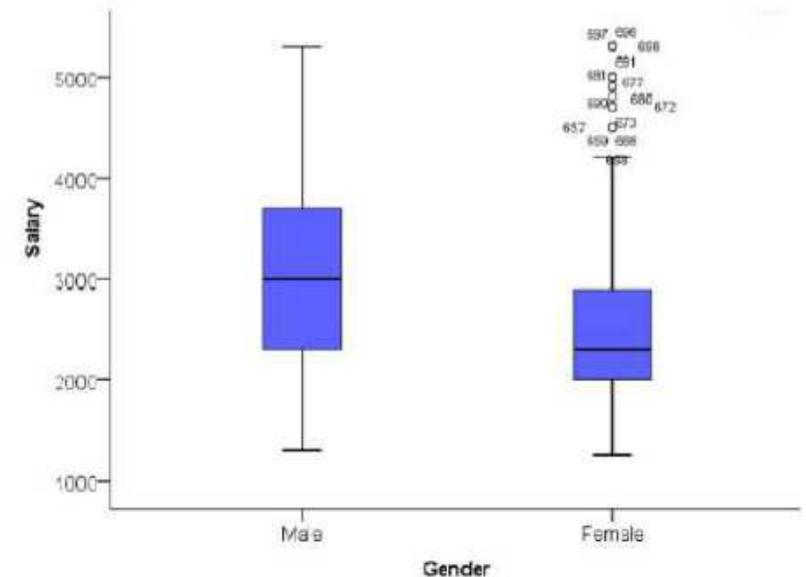
- 1) Capire se hanno una relazione
- 2) Capire se una causa l'altra
- 3) Grafici maggiormente utilizzati: grafici a barre sovrapposte o affiancate, boxplot condizionati, **diagrammi a dispersione**

Si passa da una idea puramente **descrittiva** di un fenomeno al cercare la sua **spiegazione** (introducendo una nuova variabile)

Che tipo di analisi bivariate possiamo condurre?



Oss	maschi	femmine	Totale marginale di riga
ripetenti	12	4	16
non ripetenti	7	16	23
Totale marginale di colonna	19	20	39



Entrambe **quantitative** Entrambe **qualitative** (incroci di categorie)

Una **numerica** e una **qualitativa**

Che tipo di analisi bivariate possiamo condurre?

*Nel caso di **variabili qualitative**, si può parlare di:*

- o1) Tabelle di contigenza e misure di associazione

*Nel caso di **variabili quantitative**, si può parlare di:*

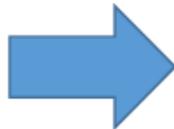
- o1) Covarianza
- o2) Correlazione
- o3) Diagramma a dispersione
- o4) Analisi della dipendenza

Tabella di contingenza: due variabili qualitative che si incrociano

		<u>Fumo</u>	
		Sì	No
<u>Genere</u>	M	25	35
	F	20	30

La tabella generata si chiama **tabella di contingenza**

Essa incrocia le modalità di **2 variabili**, formando delle celle che sono dei veri e propri incroci, cioè individui che presentano contemporaneamente **la modalità in riga e quella in colonna**



Si lavora sulle **frequenze che si incrociano**

La tabella di contingenza in dettaglio

Variabile Fumo Modalità fumo: si o no?

Ogni cella, frequenza congiunta

Variabile Genere

		<u>Fumo</u>		
		Sì	No	
<u>Genere</u>	M	25	35	60
	F	20	30	50
		45	65	

Modalità genere: M o F?

Ogni cella esterna frequenza marginale (o di riga o di colonna)

La somma delle frequenze marginali, o di riga o di colonna, darà sempre il totale, che qui è uguale a 110

Quindi, N=110

Perché si costruisce?

- Vogliamo valutare la **relazione** (presente o meno) tra due variabili qualitative
- Vogliamo comunque vedere la **distribuzione congiunta** delle due insieme
- Congiunta, cioè **insieme**: frequenze congiunte nelle celle frutto di incroci
- Vogliamo valutare la **distribuzione condizionata** di una rispetto all'altra
- Condizionata, **cioè una data l'altra**: una sola colonna o una sola riga
- La relazione tra due variabili qualitative prende **il nome di associazione**

Esempi di tabelle di contingenza

	Soda	Coffee	Tea	Water	Total
20-29	10	8	5	2	25
30-39	11	9	2	3	25
40-49	8	9	1	7	25
50-59	9	8	3	5	25
Total	38	34	11	17	100



Anche variabili quantitative discretizzate (età) possono essere usate in tabelle di contingenza

Health status	Gender	Test result		
		Positive	Negative	Equivocal
Diseased	Male	a	b	c
	Female	d	e	f
Healthy	Male	g	h	i
	Female	j	k	l

Here, one variable depicts health status, another depicts gender, and the third depicts the outcome in the form a test result



Oltre che due variabili, è possibile addirittura incrociarne ben 3!!
La tabella si complica ma fino a 3 variabili è ancora leggibile

Tabella di contingenza

Tabella di contingenza: sinonimo di **distribuzione doppia di frequenze**, ossia di distribuzione di frequenze secondo due caratteri.

Età	Grado d'istruzione				Totale
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 4 anni	
da 25 a 34	4459	11562	10693	11071	37785
da 35 a 54	9174	26455	22647	23160	81436
oltre 55	14226	20060	11125	10597	56008
Totale	27859	58077	44465	44828	175229

Tabella di contingenza: distribuzione marginale(1)

Le frequenze che appaiono nell'ultima riga, pari ai totali di colonna, si riferiscono unicamente al carattere "Grado d'istruzione"; esse configurano, insieme alle modalità, la **distribuzione marginale del carattere** in questione.

Età	Grado d'istruzione				Totale
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 4 anni	
da 25 a 34	4459	11562	10693	11071	37785
da 35 a 54	9174	26455	22647	23160	81436
oltre 55	14226	20060	11125	10597	56008
Totale	27859	58077	44465	44828	175229

Tabella di contingenza: distribuzione marginale(2)

Le frequenze che appaiono nell'ultima colonna, pari ai totali di riga, si riferiscono unicamente al carattere "Età"; esse configurano, insieme alle modalità, la distribuzione marginale del carattere in questione.

Età	Grado d'istruzione				Totale
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 4 anni	
da 25 a 34	4459	11562	10693	11071	37785
da 35 a 54	9174	26455	22647	23160	81436
oltre 55	14226	20060	11125	10597	56008
Totale	27859	58077	44465	44828	175229

Tabella di contingenza in simboli

Carattere X	Carattere Y						Totale
	y_1	y_2	...	y_j	...	y_t	
x_1	n_{11}	n_{12}	⋮	n_{1j}	⋮	n_{1t}	n_{10}
x_2	n_{21}	n_{22}	⋮	n_{2j}	⋮	n_{2t}	n_{20}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	n_{i1}	n_{i2}	⋮	n_{ij}	⋮	n_{it}	n_{i0}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_s	n_{s1}	n_{s2}	⋮	n_{sj}	⋮	n_{st}	n_{s0}
Totale	n_{01}	n_{02}	⋮	n_{0j}	⋮	n_{0t}	N

- ❑ questi simboli indicano modalità di caratteri qualitativi, oppure valori o classi di caratteri quantitativi (s è il numero di modalità di X, t il numero di modalità di Y)
- ❑ n_{ij} è la frequenza della coppia di modalità (x_i, y_j) -FREQUENZE CONGIUNTE
- ❑ n_{i0} e n_{0j} sono i totali di riga e di colonna -FREQUENZE MARGINALI

Tabella di contingenza: percentuali sul totale

Età	Grado d'istruzione				Totale
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 3 anni	
da 25 a 34	2.5%	6.6%	6.1%	6.3%	21.6%
da 35 a 54	5.2%	15.1%	12.9%	13.2%	46.5%
oltre 55	8.1%	11.4%	6.3%	6.0%	32.0%
Totale	15.9%	33.1%	25.4%	25.6%	100%

La percentuale che appare in una data casella indica la frequenza percentuale delle unità del collettivo in esame che presentano le modalità che corrispondono a tale casella (in alto e nel margine sinistro).

Ad esempio il 15.1% delle unità statistiche del collettivo statistico in esame hanno un'età compresa fra i 35 e 54 ed il diploma di scuola superiore

Tabella di contingenza: percentuali sul totale in simboli

Carattere X	Carattere Y						Totale
	y_1	y_2	...	y_j	...	y_t	
x_1	p_{11}	p_{12}	⋮	p_{1j}	⋮	p_{1t}	p_{10}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	p_{i1}	p_{i2}	⋮	p_{ij}	⋮	p_{it}	p_{i0}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_s	p_{s1}	p_{s2}	⋮	p_{sj}	⋮	p_{st}	p_{s0}
Totale	p_{01}	p_{02}	⋮	p_{0j}	⋮	p_{0t}	100

$$p_{ij} = \frac{n_{ij}}{N} \times 100$$

Frequenza percentuale congiunta
della modalità (x_i, y_j)

$$p_{i0} = \frac{n_{i0}}{N} \times 100$$

Frequenza percentuale marginale
della modalità x_i

$$p_{0j} = \frac{n_{0j}}{N} \times 100$$

Frequenza percentuale marginale
della modalità y_j

$$\sum_{i=1}^s \sum_{j=1}^t p_{ij} = 100$$

$$\sum_{i=1}^s p_{i0} = 100$$

$$\sum_{j=1}^t p_{0j} = 100$$

Tabella di contingenza: distribuzioni condizionate (1)

Età	Grado d'istruzione				Totale
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 4 anni	
da 25 a 34	4459	11562	10693	11071	37785
da 35 a 54	9174	26455	22647	23160	81436
oltre 55	14226	20060	11125	10597	56008
Totale	27859	58077	44465	44828	175229

Prendere una distribuzione condizionata equivale a considerare la distribuzione del carattere "**Grado d'istruzione**" limitatamente ai casi in cui il **carattere età è compreso ad esempio nell'intervallo 35-54.**

Tabella di contingenza: distribuzioni condizionate (2)

Se, invece, associamo alle modalità del carattere “Età” le frequenze di una colonna interna della tabella, otteniamo una **distribuzione condizionata**.

Età	Grado d'istruzione				Totale
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 4 anni	
da 25 a 34	4459	11562	10693	11071	37785
da 35 a 54	9174	26455	22647	23160	81436
oltre 55	14226	20060	11125	10597	56008
Totale	27859	58077	44465	44828	175229

Prendere una distribuzione condizionata equivale a considerare la distribuzione del carattere “Età” limitatamente ai casi in cui il carattere Grado di istruzione è uguale ad esempio al diploma di scuola superiore.

Distribuzione marginale e distribuzioni condizionate di Y in simboli

Le due righe segnalate in rosso configurano la distribuzione marginale del carattere Y. Le due righe segnalate in giallo configurano la generica distribuzione condizionata di Y rispetto ad una data modalità di X.

Carattere X	Carattere Y						Totale
	y_1	y_2	...	y_i	...	y_t	
x_1	n_{11}	n_{12}	⋮	n_{1j}	⋮	n_{1t}	n_{10}
x_2	n_{21}	n_{22}	⋮	n_{2j}	⋮	n_{2t}	n_{20}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	n_{i1}	n_{i2}	⋮	n_{ij}	⋮	n_{it}	n_{i0}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_s	n_{s1}	n_{s2}	⋮	n_{sj}	⋮	n_{st}	n_{s0}
Totale	n_{01}	n_{02}	⋮	n_{0j}	⋮	n_{0t}	N

Distribuzione marginale e distribuzioni condizionate di X in simboli

Le due righe segnalate in rosso configurano la **distribuzione marginale del carattere X**.
 Le due righe segnalate in giallo configurano la **generica distribuzione condizionata di X rispetto ad una data modalità di Y**.

Carattere X	Carattere Y						Totale
	y_1	y_2	...	y_j	...	y_t	
x_1	n_{11}	n_{12}	\vdots	n_{1j}	\vdots	n_{1t}	n_{10}
x_2	n_{21}	n_{22}	\vdots	n_{2j}	\vdots	n_{2t}	n_{20}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_i	n_{i1}	n_{i2}	\vdots	n_{ij}	\vdots	n_{it}	n_{i0}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_s	n_{s1}	n_{s2}	\vdots	n_{sj}	\vdots	n_{st}	n_{s0}
Totale	n_{01}	n_{02}	\vdots	n_{0j}	\vdots	n_{0t}	N

Tabella di contingenza: confronto delle distribuzioni condizionate

L'idea che sarà sviluppata è quella di confrontare le distribuzioni condizionate di un carattere rispetto alle modalità dell'altro per stabilire se i due caratteri sono legati.

Questo confronto **non può che essere effettuato prendendo le frequenze relative o percentuali di riga (o di colonna)** delle distribuzioni da confrontare.

Una distribuzione marginale o condizionata in cui si considerano le frequenze relative o percentuali (*sul totale di riga o di colonna*) verrà qualificata come **normalizzata** (*ad 1 se si considerano le frequenze relative, a 100 se si considerano le frequenze percentuali*).

I caratteri possono essere di natura qualsiasi.

Tabella di contingenza: confronto delle distribuzioni condizionate- esempio

Esempio: si vuole stabilire se il carattere “Grado d’istruzione” sia legato all’età; in altre parole, si vuole accertare se l’appartenenza a una data classe di età influisca, determini, in qualche misura, i livelli di istruzione degli individui di quella classe.

Non è possibile confrontare direttamente le frequenze assolute registrate da una data modalità del grado di istruzione nelle diverse classi di età perché per ogni classe di età si registra un numero diverso di unità statistiche (frequenza marginale di riga).

Per effettuare il confronto si possono calcolare le **percentuali** (o le **proporzioni**) sul totale di riga **normalizzando** in tal modo i sotto-collettivi distinti per classe di età:

si considera un sotto-collettivo di numerosità 100 (o 1) per ogni classe di età

Tabella di contingenza

Età	Grado d'istruzione				Totale
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 4 anni	
da 25 a 34	4459	11562	10693	11071	37785
da 35 a 54	9174	26455	22647	23160	81436
oltre 55	14226	20060	11125	10597	56008
Totale	27859	58077	44465	44828	175229

Tabella di contingenza: percentuali sul totale di riga
(distribuzione percentuale di Y condizionata ad X)

esempio

Età	Grado d'istruzione				Totale
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 3 anni	
da 25 a 34	11.8%	30.6%	28.3%	29.3%	100%
da 35 a 54	11.3%	32.5%	27.8%	28.4%	100%
oltre 55	25.4%	35.8%	19.9%	18.9%	100%
Distribuzione marginale % del grado di istruzione	15.9%	33.1%	25.4%	25.6%	100%

La percentuale che appare in una data casella indica la percentuale delle unità del sotto-collettivo di età compresa nella classe indicata in riga che presentano il grado di istruzione indicato in colonna.

Ad esempio il 32.5% delle unità statistiche che hanno un'età compresa fra i 35 e 54 ha il diploma di scuola superiore.

Distribuzione percentuale di Y condizionata ad X in simboli

Carattere X	Carattere Y						Totale
	y_1	y_2	...	y_j	...	y_t	
x_1	$p_{1/x1}$	$p_{2/x1}$	⋮	$p_{j/x1}$	⋮	$p_{t/x1}$	100
x_2	$p_{1/x2}$	$p_{2/x2}$	⋮	$p_{j/x2}$	⋮	$p_{t/x2}$	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	$p_{1/xi}$	$p_{2/xi}$	⋮	$p_{j/xi}$	⋮	$p_{t/xi}$	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_s	$p_{1/xs}$	$p_{2/xs}$	⋮	$p_{j/xs}$	⋮	$p_{t/xs}$	100
Distribuzione % marginale di y	p_{01}	p_{02}	⋮	p_{0j}	⋮	p_{0t}	100

$$p_{j/x_i} = \frac{n_{ij}}{n_{i0}} \times 100$$

Frequenza percentuale della modalità y_j condizionata alla modalità x_i

$$p_{0j} = \frac{n_{0j}}{N} \times 100$$

Frequenza percentuale marginale della modalità y_j

$$\sum_{j=1}^t p_{j/x_i} = 100, \quad i = 1, 2, \dots, s$$

Tabella di contingenza: confronto delle distribuzioni condizionate- esempio (2)

Esempio:

se si volesse confrontare la composizione in termini di età di ogni sotto-collettivo omogeneo rispetto al titolo di studio, si dovrebbero calcolare le **percentuali** (o le *proporzioni*) sul totale di colonna **normalizzando** in tal modo i sotto-collettivi distinti per grado di istruzione: **si considera un sotto-collettivo di numerosità 100 (o 1) per ogni livello di istruzione**

Tabella di contingenza

Età	Grado d'istruzione				Totale
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 4 anni	
da 25 a 34	4459	11562	10693	11071	37785
da 35 a 54	9174	26455	22647	23160	81436
oltre 55	14226	20060	11125	10597	56008
Totale	27859	58077	44465	44828	175229

Tabella di contingenza: percentuali sul totale di colonna (distribuzione percentuale di X condizionata ad Y)

esempio

Età	Grado d'istruzione				Distribuzione % marginale dell'età
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 3 anni	
da 25 a 34	16.0%	19.9%	24.0%	24.7%	21.6%
da 35 a 54	32.9%	45.6%	50.9%	51.7%	46.5%
oltre 55	51.1%	34.5%	25.0%	23.6%	32.0%
Totale	100%	100%	100%	100%	100%

La percentuale che appare in una data casella indica la percentuale delle unità del sotto-collettivo omogeneo rispetto al grado di istruzione indicato in colonna che hanno età nella classe indicata in riga.

Ad esempio il 45.6% delle unità statistiche aventi il diploma di scuola superiore ha un'età compresa fra 35 e 54 anni.

Distribuzione percentuale di X condizionata ad Y in simboli

Carattere X	Carattere Y						Distribuzione % marginale di X
	y_1	y_2	...	y_j	...	y_t	
x_1	$p_{1/y1}$	$p_{1/y2}$	⋮	$p_{1/yj}$	⋮	$p_{1/yt}$	p_{10}
x_2	$p_{2/y1}$	$p_{2/y2}$	⋮	$p_{2/yj}$	⋮	$p_{2/yt}$	p_{20}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	$p_{i/y1}$	$p_{i/y2}$	⋮	$p_{i/yj}$	⋮	$p_{i/yt}$	p_{i0}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_s	$p_{s/y1}$	$p_{s/y2}$	⋮	$p_{s/yj}$	⋮	$p_{s/yt}$	p_{s0}
Totale	100	100	⋮	100	⋮	100	100

$$p_{i/y_j} = \frac{n_{ij}}{n_{0j}} \times 100$$

Frequenza percentuale della modalità x_i condizionata alla modalità y_j

$$p_{i0} = \frac{n_{i0}}{N} \times 100$$

Frequenza percentuale marginale della modalità y_j

$$\sum_{i=1}^s p_{i/y_j} = 100, \quad j = 1, 2, \dots, t$$

Confronto delle distribuzioni condizionate- esempio

Gruppo di lauree	Numero di Laureati			Percentuali riga			Percentuali colonna		
	Uomini	Donne	Totale	Uomini	Donne	Totale	Uomini	Donne	Totale
agrarario	2002	1664	3666	54.6%	45.4%	100%	2.7%	1.5%	2.0%
architettura	4544	5022	9566	47.5%	52.5%	100%	6.2%	4.5%	5.2%
chimico-farmaceutico	1954	3299	5253	37.2%	62.8%	100%	2.7%	3.0%	2.8%
economico-statistico	12286	13417	25703	47.8%	52.2%	100%	16.7%	12.1%	13.9%
educazione fisica	2002	1341	3343	59.9%	40.1%	100%	2.7%	1.2%	1.8%
geo-biologico	3264	6627	9891	33.0%	67.0%	100%	4.4%	6.0%	5.4%
giuridico	3948	6413	10361	38.1%	61.9%	100%	5.4%	5.8%	5.6%
ingegneria	16832	5315	22147	76.0%	24.0%	100%	22.9%	4.8%	12.0%
insegnamento	672	8789	9461	7.1%	92.9%	100%	0.9%	7.9%	5.1%
letterario	4757	11702	16459	28.9%	71.1%	100%	6.5%	10.5%	8.9%
linguistico	1570	9335	10905	14.4%	85.6%	100%	2.1%	8.4%	5.9%
medico	6802	13503	20305	33.5%	66.5%	100%	9.3%	12.2%	11.0%
politico-sociale	7500	13990	21490	34.9%	65.1%	100%	10.2%	12.6%	11.6%
psicologico	1677	8677	10354	16.2%	83.8%	100%	2.3%	7.8%	5.6%
scientifico	3680	1896	5576	66.0%	34.0%	100%	5.0%	1.7%	3.0%
Totale	73490	110990	184480	39.8%	60.2%	100%	100%	100%	100%

Elaborazione su dati Almalaurea -XV Indagine (2013) - Condizione occupazionale dei laureati

Analisi delle relazioni di una distribuzione doppia con caratteri di natura qualsiasi

Data la distribuzione congiunta di due caratteri di natura qualsiasi, i quesiti a cui la statistica deve rispondere sono

- ✓ esiste **dipendenza** o **indipendenza** fra i due caratteri?
- ✓ se esiste dipendenza, come sono associate le modalità dei due caratteri?

Analisi delle relazioni di una distribuzione doppia con caratteri di natura qualsiasi: indipendenza

Su un collettivo di 180 studenti della d'Annunzio, distinti per sesso, è stata rilevata la frequenza con cui accedono al servizio mensa

Indipendenza

	Frequenza della mensa			
Sesso	Raramente	A volte	Spesso	Totale
Maschio	10	20	30	60
Femmina	20	40	60	120
Totale	30	60	90	180

	Frequenza della mensa			
Sesso	Raramente	A volte	Spesso	Totale
Maschio	16.7%	33.3%	50.0%	100%
Femmina	16.7%	33.3%	50.0%	100%
Totale	16.7%	33.3%	50.0%	100%

La frequenza con cui accede al servizio mensa dipende dal sesso dello studente?

Dalla tabella delle percentuali sul totale di riga si vede come per ogni modalità di frequenza del servizio mensa le percentuali di maschi e di femmine sono uguali:
la frequenza non dipende dal sesso dello studente

Analisi delle relazioni di una distribuzione doppia con caratteri di natura qualsiasi: dipendenza perfetta

Su un collettivo di 150 lavoratori viene registrato il titolo di studio e la qualifica professionale all'interno dell'azienda

Dipendenza perfetta				
	Qualifica professionale			
Titolo di studio	Bassa	Media	Alta	Totale
Licenza elementare	10	0	0	10
Licenza media	20	0	0	20
Diploma	0	70	0	70
Laurea	0	0	50	50
Totale	30	70	50	150

La qualifica professionale dipende dal titolo di studio?

Si registra una dipendenza perfetta della qualifica professionale dal titolo di studio

Un carattere Y dipende perfettamente da X quando ad ogni modalità di X è associata una sola modalità di Y. Non è un relazione bidirezionale.

Analisi delle relazioni di una distribuzione doppia con caratteri di natura qualsiasi: interdipendenza perfetta

Su un collettivo di 45 dipendenti di un'azienda si rileva la categoria professionale e la classe retributiva

Interdipendenza perfetta

Categoria	Classe retributiva			Totale
	25000-28000	28000-33500	33500-45000	
Operatore	25	0	0	25
Collaboratore	0	15	0	15
Funzionario	0	0	5	5
Totale	25	15	5	45

Il livello retributivo dipende dalla categoria professionale?

Si registra una interdipendenza perfetta

Tra due caratteri sussiste interdipendenza perfetta se ad ogni modalità di uno dei due caratteri corrisponde una e una sola modalità dell'altro carattere e viceversa

È una relazione bidirezionale.

Analisi delle relazioni di una distribuzione doppia con caratteri di natura qualsiasi: dipendenza

valori assoluti

Gruppo di lauree	<i>Lavora</i>	<i>Non lavora ma cerca</i>	<i>Impegnato in corso universitario/ praticantato</i>	<i>Non lavora e non cerca e non è impegnato</i>	<i>Totale</i>
agrario	1767	1111	645	143	3666
architettura	4496	2650	2085	335	9566
chimico-farmaceutico	2626	1361	1045	221	5253
economico-statistico	11412	7351	6272	668	25703
educazione fisica	2237	692	324	90	3343
geo-biologico	3037	2898	3620	336	9891
giuridico	2818	3647	3699	197	10361
ingegneria	10187	3699	7774	487	22147
insegnamento	6784	1996	407	274	9461
letterario	7160	5020	3588	691	16459
linguistico	5201	3359	1952	393	10905
medico	12467	4954	1970	914	20305
politico-sociale	11068	6855	2772	795	21490
psicologico	4255	2972	2775	352	10354
scientifico	2549	836	2007	184	5576
Totale	88064	49401	40935	6080	184480

Analisi delle relazioni di una distribuzione doppia con caratteri di natura qualsiasi: dipendenza

% riga

Gruppo di lauree	<i>Lavora</i>	<i>Non lavora ma cerca</i>	<i>Impegnato in corso universitario/ praticantato</i>	<i>Non lavora e non cerca e non è impegnato</i>	<i>Totale</i>
agrario	48.2%	30.3%	17.6%	3.9%	100%
architettura	47.0%	27.7%	21.8%	3.5%	100%
chimico-farmaceutico	50.0%	25.9%	19.9%	4.2%	100%
economico-statistico	44.4%	28.6%	24.4%	2.6%	100%
educazione fisica	66.9%	20.7%	9.7%	2.7%	100%
geo-biologico	30.7%	29.3%	36.6%	3.4%	100%
giuridico	27.2%	35.2%	35.7%	1.9%	100%
ingegneria	46.0%	16.7%	35.1%	2.2%	100%
insegnamento	71.7%	21.1%	4.3%	2.9%	100%
letterario	43.5%	30.5%	21.8%	4.2%	100%
linguistico	47.7%	30.8%	17.9%	3.6%	100%
medico	61.4%	24.4%	9.7%	4.5%	100%
politico-sociale	51.5%	31.9%	12.9%	3.7%	100%
psicologico	41.1%	28.7%	26.8%	3.4%	100%
scientifico	45.7%	15.0%	36.0%	3.3%	100%
% condizione occupazionale	47.7%	26.8%	22.2%	3.3%	100%

Elaborazione su dati Almalaurea -XV Indagine (2013) - Condizione occupazionale dei laureati

Dipendenza statistica: connessione

Alla luce delle considerazioni precedenti, possiamo dare una definizione del concetto di dipendenza statistica con riferimento a una tabella di contingenza.

Il carattere Y (X) dipende dal carattere X (Y) se le distribuzioni condizionate normalizzate sono diverse tra loro.

All'opposto:

Si dice che il carattere Y (X) non dipende dal carattere X (Y) quando le distribuzioni condizionate normalizzate sono uguali tra loro.

Frequenze congiunte in caso di indipendenza

Situazione di indipendenza

	Frequenza della mensa			Tot
	Raramente	A volte	Spesso	
M	10	20	30	60
F	20	40	60	120
Tot	30	60	90	180

	Frequenza della mensa			Totale
	Raramente	A volte	Spesso	
M	16.7%	33.3%	50.0%	100%
F	16.7%	33.3%	50.0%	100%
Tot	16.7%	33.3%	50.0%	100%

X e Y sono indipendenti se le distribuzioni delle frequenze percentuali condizionate di Y rispetto ad X sono uguali alla distribuzione percentuale marginale di Y (oppure se le distribuzioni delle frequenze percentuali condizionate di X rispetto ad Y sono uguali alla marginale di X).

Quindi nel caso di indipendenza esiste proporzionalità fra le frequenze assolute delle righe (o delle colonne).

Da tale relazione si vede come nel caso di indipendenza le frequenze congiunte possono essere espresse come

$$p_{j/x_i} = p_{0j}$$

$$\frac{n_{ij}}{n_{i0}} 100 = \frac{n_{0j}}{N} 100 \quad \forall i, j$$

$$n_{ij} : n_{i0} = n_{0j} : N \quad \forall i, j$$

$$n_{ij} = \frac{n_{i0} \cdot n_{0j}}{N} \quad \forall i, j$$

Misura della dipendenza

L'intensità della dipendenza viene misurata come allontanamento dalla condizione di indipendenza.

Per far questo si costruisce la tabella teorica di indipendenza che presenta in ogni casella la frequenza teorica di indipendenza.

$$\hat{n}_{ij} = \frac{n_{i0} \cdot n_{0j}}{N} \quad \forall i, j$$

Per misurare la dipendenza si mette a confronto la tabella effettiva con quella teorica di indipendenza tramite le differenze

$$\text{frequenza effettiva} - \text{frequenza teorica} = n_{ij} - \hat{n}_{ij}$$

Costruzione della tabella teorica di indipendenza

Calcolo delle frequenze teoriche.

$$\hat{n}_{ij} = \frac{n_{i0} \cdot n_{0j}}{N} \quad \forall i, j$$

$$\frac{37785 \cdot 27859}{175229}$$

Età	Grado d'istruzione				Totale
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 4 anni	
da 25 a 34	6007.3				37785
da 35 a 54					81436
oltre 55					56008
Totale	27859	58077	44465	44828	175229

Per calcolare le frequenze teoriche si riscrive la tabella riportando solo i marginali di riga e di colonna ed il totale e poi per ogni casella si effettua il prodotto tra il totale di riga e il totale di colonna diviso per il totale generale.

Costruzione della tabella teorica di indipendenza

Nella tabella che segue sono indicate in rosso le frequenze che si avrebbero se vi fosse indipendenza.

$$\hat{n}_{ij} = \frac{n_{i0} \cdot n_{0j}}{N} \quad \forall i, j$$

Età	Grado d'istruzione				Totale
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 4 anni	
da 25 a 34	6007.3 4459	12523.3 11562	9588.1 10693	9666.4 11071	37785
da 35 a 54	12947.2 9174	26990.7 20455	20664.7 22647	20833.4 23160	81436
oltre 55	8904.5 14226	18563.0 20060	14212.2 11125	14328.3 10597	56008
Totale	27859	58077	44465	44828	175229

Le frequenze teoriche sono state ottenute nel modo seguente: in ogni casella si è posto il numero risultante dal prodotto tra il totale di riga e il totale di colonna diviso per il totale generale.

Indice di associazione chi-quadrato di Pearson

$$\chi^2 = \sum_{i=1}^s \sum_{j=1}^t \frac{(n_{ij} - n_{i0}n_{0j}/N)^2}{n_{i0}n_{0j}/N} =$$
$$= \sum_{i=1}^s \sum_{j=1}^t \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

L'indice χ^2

- è sempre non negativo
- assume valore 0 nel caso di associazione nulla
- aumenta all'aumentare della dipendenza (o associazione dei due caratteri)
- a parità di associazione l'indice aumenta al crescere di N (quindi dipende da N)

Misura della dipendenza: calcolo dell'indice

χ^2

$$\chi^2 = \sum_{i=1}^s \sum_{j=1}^t \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Età	Grado d'istruzione				Totale
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 4 anni	
da 25 a 34	6007.3 4459	12523.3 11562	9588.1 10693	9666.4 11071	37785
da 35 a 54	12947.2 9174	26990.7 26455	20664.7 22647	20833.4 23160	81436
oltre 55	8904.5 14226	18563.0 20060	14212.2 11125	14328.3 10597	56008
Total e	27859	58077	44465	44828	175229

$$\begin{aligned} \chi^2 &= \frac{(4459 - 6007.3)^2}{6007.3} + \\ &+ \frac{(11562 - 12523.3)^2}{12523.3} + \dots + \\ &+ \frac{(10597 - 14328.3)^2}{14328.3} = \\ &= 7307.8 \end{aligned}$$

Età	Grado d'istruzione			
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 4 anni
da 25 a 34	399.1	73.8	127.3	204.1
da 35 a 54	1099.6	10.6	190.2	259.8
oltre 55	3180.2	120.7	670.6	971.7

$$\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Totale=7307.8

Indice di associazione quadratica media

Siccome l'indice χ^2 dipende da N, per eliminare l'influenza della numerosità delle osservazioni si può calcolare l'indice di associazione quadratica media

$$\psi = \sqrt{\frac{1}{N} \sum_{i=1}^s \sum_{j=1}^t \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}} = \sqrt{\frac{\chi^2}{N}}$$

➤ Nell'esempio precedente si ha

$$\psi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{7307.8}{175229}} = 0.2$$

Indici di associazione relativi

L'indice ψ , e l'indice χ^2 , vanno rapportati al massimo che possono assumere per poter esprimere un giudizio sul **grado di dipendenza**.

Si può dimostrare che il massimo di ψ è pari a

$$\max(\psi) = \sqrt{\min(s-1, t-1)}$$

Il massimo di χ^2 è

$$\max(\chi^2) = N \cdot \min(s-1, t-1)$$

Nella [tabella di interdipendenza perfetta](#) (dove si deve avere $s=t$) si ottiene il valore massimo

Interdipendenza perfetta: un esempio

X	Y				Tot
	y ₁	y ₂	y ₃	y ₄	
x ₁	45	0	0	0	45
	10.4	4.6	8.6	21.3	
x ₂	0	20	0	0	20
	4.6	2.1	3.8	9.5	
x ₃	0	0	0	92	92
	21.3	9.5	17.5	43.6	
x ₄	0	0	37	0	37
	8.6	3.8	7.1	17.5	
Tot	45	20	37	92	194

X	Y				Totale
	y ₁	y ₂	y ₃	y ₄	
x ₁	10.4	4.7	8.6	21.3	
x ₂	4.7	156.1	3.8	9.5	
x ₃	21.3	9.5	17.5	53.6	
x ₄	8.6	3.8	127.1	17.5	
Totale	582				582

$$\max(\chi^2) = N \cdot \min(s - 1, t - 1)$$

$$\max(\psi) = \sqrt{\min(s - 1, t - 1)}$$

$$\chi^2 = \sum_{i=1}^s \sum_{j=1}^t \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} =$$

$$= 582$$

$$\psi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{582}{194}} = \sqrt{3}$$

$$\max(\chi^2) = N \min(s - 1, t - 1) = 194 \cdot \min(4 - 1, 4 - 1) = 194 \cdot 3 = 582$$

$$\max(\psi) = \sqrt{\min(s - 1, t - 1)} = \sqrt{\min(4 - 1, 4 - 1)} = \sqrt{3}$$

Indice simmetrico di connessione

Un indice normalizzato di connessione è l'**indice di Cramér**

$$C = \frac{\psi}{\sqrt{\min[(s-1), (t-1)]}}$$

dove $\min[(s-1), (t-1)]$ indica il minimo tra le due quantità tra parentesi quadra.

L'indice C

- è sempre compreso tra 0 e 1

vale 0 nel caso di indipendenza

vale 1 se

i due caratteri sono perfettamente associati

e $s=t$

X dipende perfettamente da Y e $s < t$

Y dipende perfettamente da X e $s > t$

Misura normalizzata di dipendenza unilaterale

Se si considera la variabile Y come logicamente dipendente si può considerare una dipendenza unilaterale e rapportare l'indice ψ al suo massimo, ottenendo l'indice normalizzato di dipendenza di Y da X

$$C_y = \frac{\psi}{\sqrt{t-1}}$$

Analogamente, se si considera la variabile X come logicamente dipendente si può considerare una dipendenza unilaterale e rapportare l'indice ψ al suo massimo, ottenendo l'indice normalizzato di dipendenza di X da Y

$$C_x = \frac{\psi}{\sqrt{s-1}}$$

Misura normalizzata di dipendenza : esempio

Calcolo dell'indice C_Y per la tabella di contingenza

	Grado d'istruzione				Totale
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 4 anni	
da 25 a 34	4459	11562	10693	11071	37785
da 35 a 54	9174	26455	22647	23160	81436
oltre 55	14226	20060	11125	10597	56008
Totale	27859	58077	44465	44828	175229

$$\psi = 0.20$$

$$C_Y = \frac{0.20}{\sqrt{3}} = 0.12$$

Non si ha una forte dipendenza di Y da X

$$\max(\psi) = \sqrt{t-1} = \sqrt{3}$$

Interpretazione dell'associazione

Se non c'è indipendenza, per veder come si associano le modalità dei due caratteri si possono confrontare le frequenze osservate con le frequenze teoriche.

Età	Grado d'istruzione			
	Non ha terminato le superiori	Diploma di scuola superiore	Università da 1 a 3 anni	Università oltre i 4 anni
da 25 a 34	Neg. 6007.3 4459	Neg. 12523.3 11562	Pos. 9588.1 10693	Pos. 9666.4 11071
da 35 a 54	Neg. 12947.2 9174	Neg. 26990.7 26455	Pos. 20664.7 22647	Pos. 20833.4 23160
oltre 55	Pos. 8904.5 14226	Pos. 18563.0 20060	Neg. 14212.2 11125	Neg. 14328.3 10597

$n_{ij} > \hat{n}_{ij}$ Associazione positiva fra le modalità x_i e y_j

$n_{ij} < \hat{n}_{ij}$ Associazione negativa fra le modalità x_i e y_j

Interpretazione dell'associazione: esempio

Inclinazione politica	<i>Livello di accordo con l'affermazione: È molto importante per me essere Italiano</i>				Totale
	<i>per nulla d'accordo</i>	<i>poco d'accordo</i>	<i>abbastanza d'accordo</i>	<i>molto d'accordo</i>	
sinistra	Pos. 59 37.5	Pos. 110 84.5	Pos. 51 64.6	14 47.4	234
centro	22 39.4	77 88.8	84 68.0	Pos. 63 49.8	246
destra	6 10.1	9 22.7	15 17.4	Pos. 33 12.8	63
Totale	87	196	150	110	543

Interpretazione dell'associazione: esempio

	<u>Livello di accordo con l'affermazione:</u> È molto importante per me essere Italiano				
Inclinazione politica	<i>per nulla d'accordo</i>	<i>poco d'accordo</i>	<i>abbastanza d'accordo</i>	<i>molto d'accordo</i>	Totale
sinistra	25.2%	47.0%	21.8%	6.0%	100%
centro	8.9%	31.3%	34.1%	25.6%	100%
destra	9.5%	14.3%	23.8%	52.4%	100%
Distribuzione % del livello di accordo	16.0%	36.1%	27.6%	20.3%	100%