Il test di indipendenza in una tabella di contingenza

Lezioni di Psicometria

Giovanni Battista Flebus

Tabelle di contingenza

- La misurazione più semplice è la conta delle frequenze, per esempio il genere, il tipo di scuola frequentato, la provenienza geografica ecc.

- La frequenza o il conteggio sono delle informazioni che raramente presentano un interesse per la ricerca.

- Solitamente è più interessante rilevare la contemporanea presenza di due variabili (per es, genere e provenienza, professione del padre e scuola frequentata ecc) e verificare se fra le due caratteristiche ci sono delle interdipendenze. I dati si presentano nelle **tabelle di contingenza**.

TABELLE DI CONTINGENZA a un criterio

- La tabella di contingenza può contenere dati riferiti alle categorie di **una sola** variabile.
- Se, ad esempio, in una domanda di un questionario si chiede all'intervistato qual è il suo orientamento politico, si potrebbe avere questa distribuzione

	conteggi	percentuali
1 Nessun orientamento politico (NOP)	221	14,2
2 sinistra	446	28,8
3 centro	448	28,9
4 destra	436	28,1
Totale	1551	100

TABELLE DI CONTINGENZA a un criterio

- Possiamo anche chiedere se fa le vacanze nello stesso posto, e otterremmo questa distribuzione

Fai le vacanze nello stesso posto ?	conteggi	percentuali
1 Mai o quasi mai	1058	67,8
2 spesso-sempre	503	32,2
Totale	1561	100

Tabelle di contingenza a due criteri .

- Se, invece di esaminare semplicemente la distribuzione di un campione rispetto alle categorie di una variabile, vogliamo vedere se esiste una **relazione tra due variabili**, incrociamo queste due variabili in una tabella di contingenza **a due criteri**.
- In ogni casella della tabella troviamo il numero di persone che presenta una particolare combinazione delle categorie delle due variabili.
- Considerando i dati della ricerca evidenziata, avremo una tabella di contingenza di questo tipo:

Vacan Vacanze nello stesso posto * Orpol Orientamento politico Crosstabulation

Count

		Orpol Orientamento politico				
		1 Nessun orientamento politico	2 sinistra	3 centro	4 destra	Total
Vacanze nello stesso posto	1 quasi mai	127	347	286	290	1050
	3 spesso-sempre	94	99	162	146	501
Total		221	446	448	436	1551

Relazione tra variabili visibile dalle tabelle di contingenza

Si può ipotizzare una relazione tra le variabili esaminando una tabella di contingenza?

In altre parole, esiste una relazione fra orientamento politico e scelta del posto dove andare in vacanza?

- Per rispondere, ricordiamo il concetto di totali marginali

- Per ogni riga (e per ogni colonna) otteniamo un conteggio, che può essere trasformato in percentuale

	1 NOP	2 sinistra	3 centro	4 destra		
1 quasi mai	127	347	286	290	1050	67,7 %
3 spesso-sempre	94	99	162	146	501	32,3 %
Count	221	446	448	436	1551	
	14,2%	28,8%	28,9%	28,1%		

Chi cambia spesso sono 501 su 1551, pari al 32,3 %
Chi non ha orientamento politico sono 221, pari al 14,2 %
Entrambi sono **totali marginali (o percentuali) marginali**

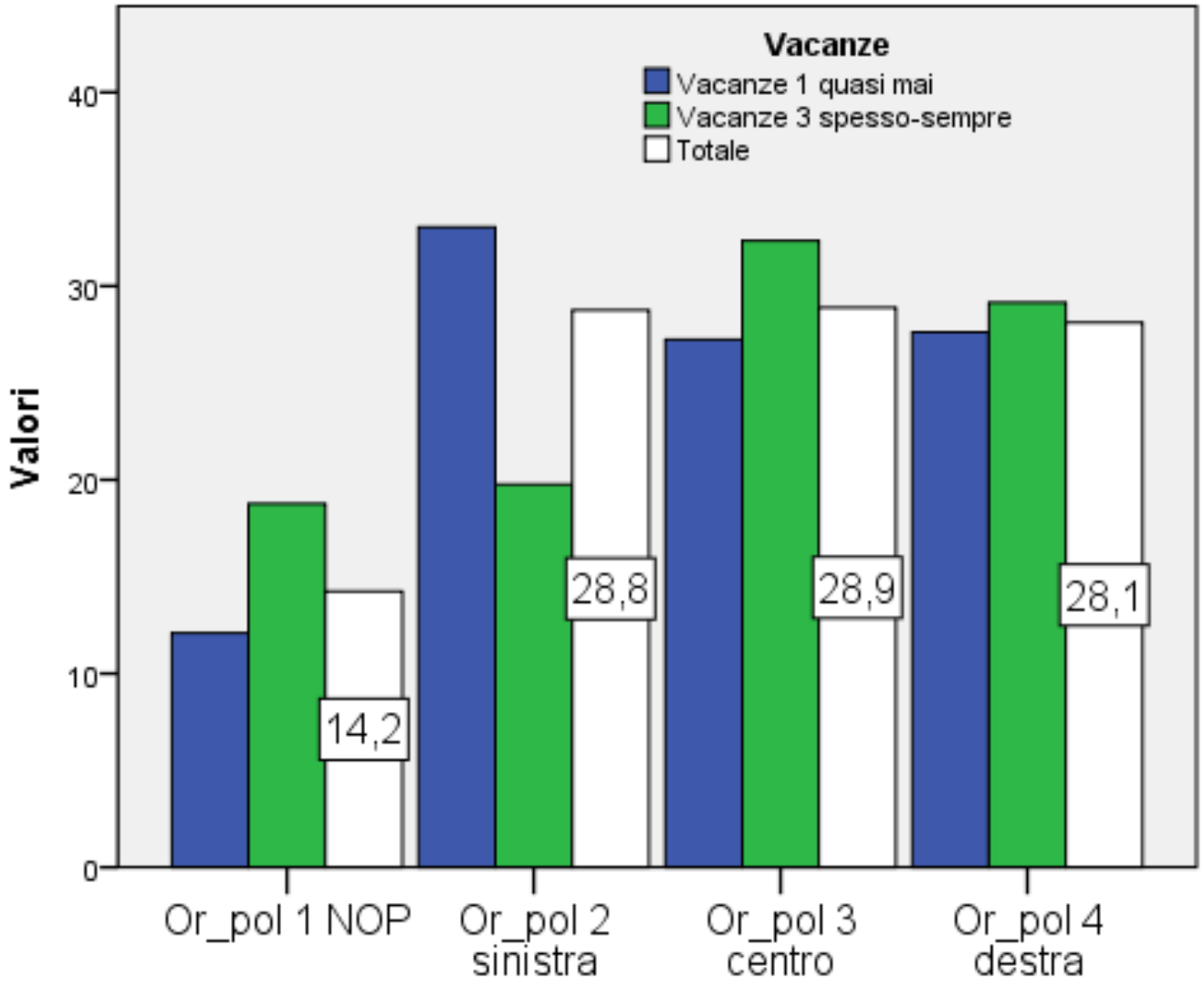
L'ipotesi di indipendenza

- Se le due misurazioni sono indipendenti, ovvero se non c'è nessuna relazione fra cambiare luogo di vacanza e avere un certo orientamento politico, le percentuali totali di riga dovrebbero essere uguali in tutte le colonne, e le percentuali totali nelle colonne dovrebbero essere uguali in tutte le righe.
- Uguali va inteso come approssimativamente uguali, ovvero, entro margini accettabili di variabilità stocastica.

	1 NOP	2 sinistra	3 centro	4 destra		
1 quasi mai	14,2	28,8	28,9	28,1	1050	67,7 %
3 spesso-sempre	14,2	28,8	28,9	28,1	501	32,3 %
Count	221	446	448	436	1551	
	14,2%	28,8%	28,9%	28,1%		

Nell'ipotesi di indipendenza, all'interno di ciascuna colonna ci aspettiamo di trovare le stesse percentuali marginali di riga.
Per esempio, se quelli di destra sono il 28,1 %, dovremmo trovare 28 % fra quelli che cambiano sempre e quelli che non cambiano mai

Tavola di contingenza
Vacanze *Orientamento politico

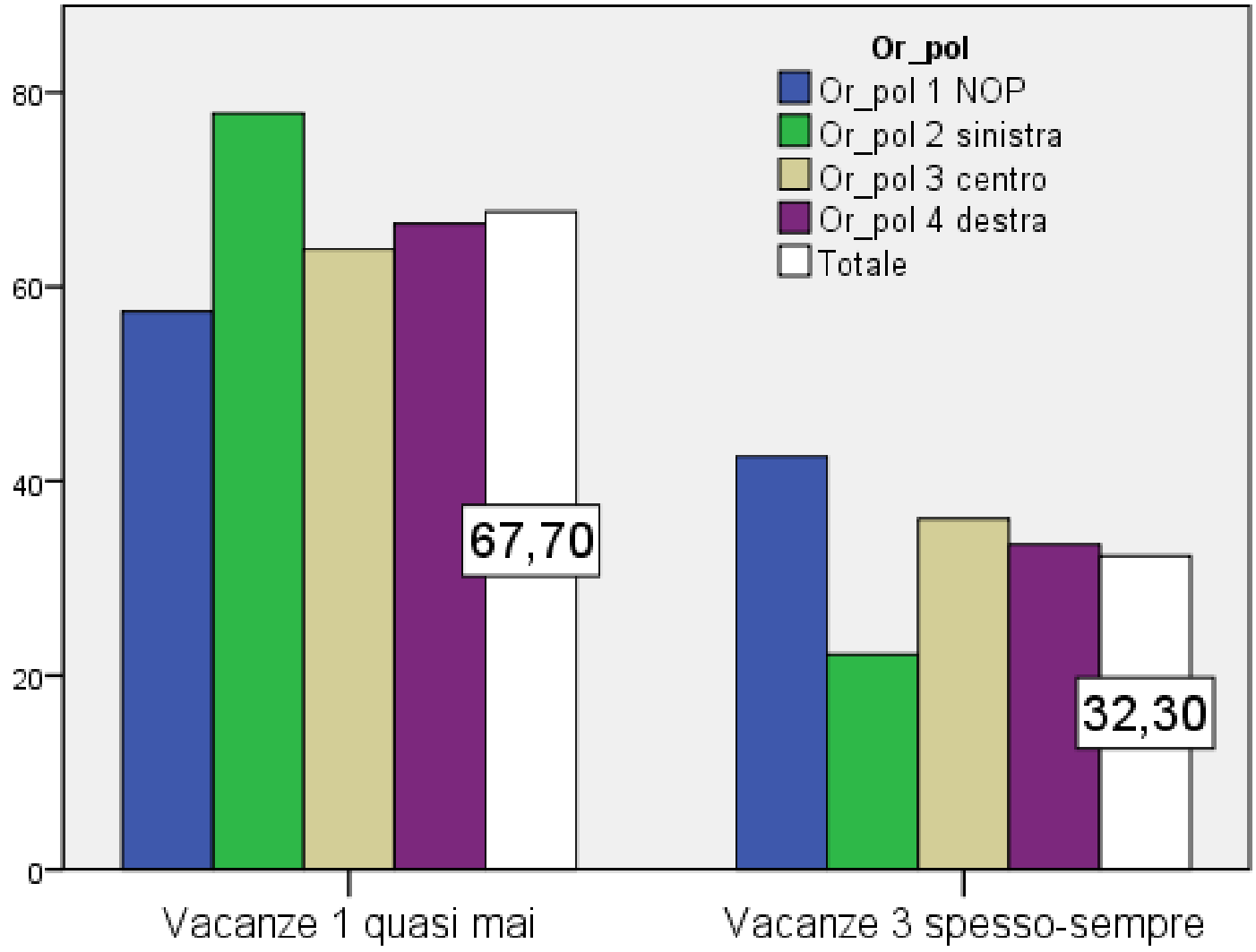


	1 NOP	2 sinistra	3 centro	4 destra		
1 quasi mai	67,7	67,7	67,7	67,7	1050	67,7 %
3 spesso-sempre	32,3	32,3	32,3	32,3	501	32,3 %
Count	221	446	448	436	1551	
	14,2%	28,8%	28,9%	28,1%		

Nell'ipotesi di indipendenza, all'interno di ciascuna riga ci aspettiamo di trovare le stesse percentuali marginali di colonna.

Per esempio, se quelli che non cambiamo mai sono il 67,7, allora dovremmo trovare in nei quattro orientamenti politici sempre la stessa percentuale di 67,7

Tavola di contingenza
Orientamento politico x



Dalle percentuali ai valori attesi…

	1 NOP	2 sinistra	3 centro	4 destra		
1 quasi mai	67,7	67,7	67,7	67,7	1050	67,7 %
3 spesso-sempre	32,3	32,3	32,3	32,3	501	32,3 %
Count	221	446	448	436	1551	
	14,2%	28,8%	28,9%	28,1%		

Quelli che non cambiano mai sono il 67,7. Possiamo tradurre il la percentuale in frequenza: il 67,7 di coloro che non hanno orientamento politico è uguale a 67,7 diviso 100 moltiplicato per 221, pari a 149,6. Questa è la frequenza teorica o frequenza attesa.
Quelli che cambiano spesso sono invece circa un terzo di 221, pertanto la frequenza attesa è uguale a 32,3 /100 x 221= 71,3. Naturalmente le due frequenze teoriche sono uguali al totale marginale: 149,6+71,3=221.

Frequenza attesa

È il conteggio teorico che ci aspettiamo di trovare in ogni cella, in base all'ipotesi di indipendenza

Il suo valore, combinando e semplificando le percentuali, si ottiene con la formula seguente:

$$freq.attesa = \frac{tot.riga \times tot.colonna}{tot.generale}$$

Per ciascuna cella della tavola di contingenza si può ottenere una frequenza attesa (indicata con la lettera A), che si può confrontare con la frequenza osservata (indicata con la lettera O)

Vacan Vacanze nello stesso posto * Orpol Orientamento politico Crosstabulation

Count

		Orpol Orientamento politico				
		1 Nessun orientamento politico	2 sinistra	3 centro	4 destra	Total
Vacanze nello stesso posto	1 quasi mai	127	347	286	290	1050
	3 spesso-sempre	94	99	162	146	501
Total		221	446	448	436	1551

$$freq.attesa \frac{1050 \times 221}{1551} = 149{,}6$$

Calcolo della discrepanza

- La differenza fra le due quantità deve essere però aggiustata in basa alla grandezza della frequenza attesa (una differenza di 5 può essere considerevole se la frequenza attesa è 2, ma trascurabile se la frequenza attesa è 200).
- La discrepanza fra frequenza attesa e frequenza osservata va elevata al **quadrato**, in modo che qualsiasi discrepanza in negativo **non compensi** una discrepanza in positivo.

Calcolo della discrepanza complessiva

- per ogni cella si calcola la discrepanza al quadrato e si sommano tutte
- Il totale o la somma delle discrepanze si chiama chi quadrato

$$x^2 = \sum \frac{(O - A)^2}{A}$$

Valutare le discrepanze

- Se le due variabili sono indipendenti, la discrepanza è piccola, poiché le frequenze osservate sono simili a quelle teoriche.
- La fluttuazione è casuale, e quindi è limitata
- Se al contrario la somma delle discrepanze è grande, le due variabili non sono indipendenti: c'è associazione fra alcuni modalità delle due variabili.
- Il test del chi quadrato ci aiuta a prendere la decisione

I gradi di libertà

- I gradi di libertà fanno riferimento al numero di celle libere di variare in una tabelle a doppia entrata.
- Con i totali marginali fissi, nella tabella accanto possiamo inserire la frequenze in due celle in due righe diverse. Non una di più.
- Tutte le altre restano automaticamente fissate, tenuto conto dei totali marginali. La tabella ha appunto (2-1)x (3-1) =2 gradi di libertà

30	50		100
			200
50	100	150	300

TEST DEL CHI QUADRATO

- Si usa la variabile casuale del chi quadrato (ottenuta sommando k variabili gaussiane standard elevate al quadrato).
- Permette di valutare la rarità di un valore (sempre positivo). Valori bassi sono frequenti, valori elevati sono rari.
- Permette quindi di valutare se il valore discrepante che abbiamo ottenuto è raro oppure comune

Procedura per il test del chi quadrato

1) Calcolare l'indice di discrepanza (chi quadrato calcolato)
2) Calcolare il numero di gradi di libertà
3) Consultare le tavole del chi quadrato con il livello di significatività desiderato (es: 0,05)
4) Individuare il valore del chi quadrato teorico con k gradi di libertà corrispondente allo 0,05
5) Confrontare il chi quadrato calcolato con il valore critico
6) Se il valore calcolate è inferiore al chi quadrato critico, accettare l'ipotesi nulla
7) Se il valore calcolate è superiore a quello critico, rifiutare l'ipotesi nulla e concludere che le due variabili non sono indipendenti

Verifica dell'associazione tra le variabili

innanzitutto, stabiliamo la nostra ipotesi nulla e quella alternativa (sono quasi sempre le stesse):

$\mathbf{H_0}$: C'è indipendenza (non c'è associazione) tra orientamento politico e scelta del posto di vacanza

.

$\mathbf{H_1}$: c'è associazione tra orientamento politico e scelta del posto di vacanza

Per decidere se accettare o scartare l'ipotesi nulla, dovremo calcolare un "valore test" statistico, e confrontare questo valore con il valore critico corrispondente.

O = valori osservati
A = valori attesi

Calcolo del CHI QUADRATO

$$x^2 = \sum \frac{(O - A)^2}{A}$$

$$Val.\,atteso = \frac{tot.colonna \times tot.riga}{tot.generale}$$

osservato	tot marg	tot margi	totale	atteso	residuo	quadrato
127	1050	221	1551	149,61	-1,85	3,42
347	1050	446	1551	301,93	2,59	6,73
286	1050	448	1551	303,29	-0,99	0,99
290	1050	436	1551	295,16	-0,30	0,09
94	501	221	1551	71,39	2,68	7,16
99	501	446	1551	144,07	-3,75	14,10
162	501	448	1551	144,71	1,44	2,07
146	501	436	1551	140,84	0,44	0,19
						34,74

Trovare il valore critico e i gradi di libertà

- Il valore critico (valore fisso) si trova sulle tavole di contingenza del chi quadrato incrociando il livello di significatività (di solito posto a 5%, per cui si guarda la colonna dello 0,05) con il numero dei gradi di libertà (gl) che si trovano con la seguente formula:

gl = (numero di righe - 1) x (numero colonne - 1)

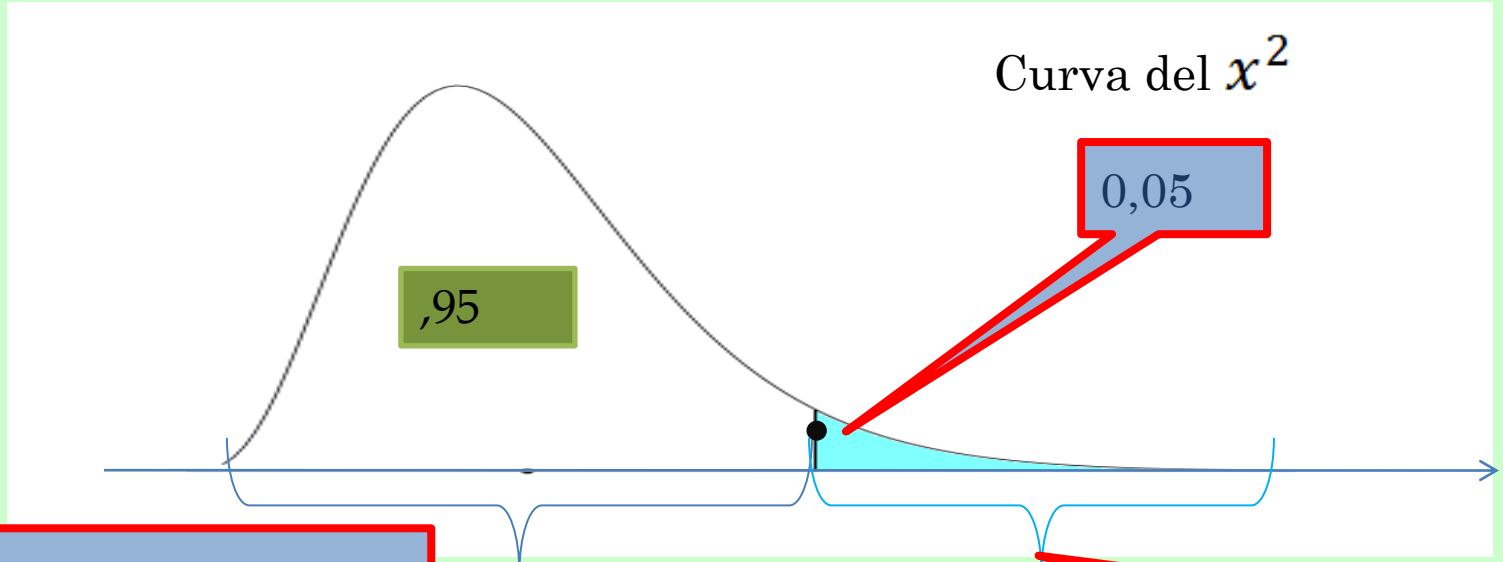
ESEMPIO:

Se la tabella ha due righe e quattro colonne, il numero dei gradi di libertà è uguale a : (2 – 1) x (4 – 1) = 1 x 3 = 3

Quindi nelle tavole di contingenza troveremo il punto dove gl = 3 e andremo fino alla casella della colonna dove il livello di significatività è di 5%: il valore trovato è **7,81**, **valore critico del chi quadrato per questo test.**

Confrontare e trarre le conclusioni

- L'ipotesi nulla può essere scartata se il valore della statistica del chi quadrato calcolato è più grande del valore critico. In tal caso, accetteremmo l'ipotesi alternativa e potremmo affermare l'esistenza di una relazione tra le due variabili prese in considerazione.



Ancora sul significato della verifica delle ipotesi

- Ecco un altro modo di interpretare il test statistico
- Dopo aver stabilito le due ipotesi (H-zero: le due caratteristiche sono indipendenti, le percentuali di riga sono uguali in tutte le colonne ; H-alternativa: vi è associazione fra le due variabili, per esempio quelli di sinistra cambiano spesso luogo di vacanza), quantifichiamo la discrepanza (il chi quadrato) fra ipotesi nulla e valori osservati: se la discrepanza è piccola, accettiamo l'ipotesi nulla, se la discrepanza è grande, la rifiutiamo.
- Che vuol dire grande discrepanza ? Si interpreta in questo modo.
- Di solito, le frequenze osservate nelle celle sono uguali a quelle attese. A volte, solo per il caso, le discrepanze sono grandi. Generalmente sono piccole, qualche rara volta queste discrepanze sono grandi. Allora , se osserviamo una discrepanza piccola, diciamo che non vi è relazione fra le due variabili. Se però osserviamo una discrepanza grande, che si manifesta per caso solo raramente (meno del cinque per cento delle volte) concludiamo una cosa diversa

- Un valore di discrepanza che ha poche probabilità di manifestarsi non lo consideriamo come un valore raro con l'ipotesi nulla.
- Lo consideriamo come un valore probabile perché c'è dipendenza fra le due variabili, ossia consideriamo valida l'ipotesi alternativa (le due variabili sono correlate). Infatti con le due variabili correlate, i valori attesi non sono quelli che abbiamo calcolato noi, sono diversi, a noi sconosciuti, ma non sono quelli del modello di indipendenza

Restrizioni sul test del x^2

- Il test del chi quadrato può essere usato solamente per le tabelle contenenti conteggi; affinché i risultati del test siano validi, è necessario che le frequenze attese ***abbiano un valore maggiore o uguale a cinque.***
- Se vi sono delle caselle con valori minori di cinque, bisogna ***raccogliere più dati***, oppure ***ridurre le categorie della tabella***.

Risultato di SPSS

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	34,735[a]	3	,000
Likelihood Ratio	35,649	3	,000
Linear-by-Linear Association	,070	1	,791
N of Valid Cases	1551		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 71,39.

Il chi quadrato è risultato significativo

- Perciò l'orientamento politico non è indipendente dalla scelta del posto dove andare in vacanza
- Si può anche dire che…
- la scelta del posto delle vacanze è in relazione con l'orientamento politico.
- Possiamo essere più specifici ?
- Guardiamo i residui…

I RESIDUI : come approfondire le analisi

- I residui standardizzati (r) sono calcolati per ciascuna casella di una tabella con la seguente formula:

$$r = \frac{O - A}{\sqrt{A}}$$

DI CUI:
O = valore osservato
A = valore atteso

INTERPRETAZIONE DEI RESIDUI

- A partire da una tabella, l'interpretazione comincia con l'osservare **se i residui di ciascuna casella sono positivi o negativi:**

Se il residuo è positivo → **Il valore osservato è più grande del valore atteso** (indica, quindi, che ci sono più persone in quella casella di quelle che avevamo supposto ci sarebbero state se non ci fosse stata associazione)

Se il residuo è negativo → **Il valore atteso è più grande del valore osservato** (indica che ci sono meno persone di quanto ci saremmo aspettati)

- Successivamente si considera la **grandezza** dei residui per vedere se la **differenza tra i valori osservati e quelli attesi è significativa.**

per determinare la significatività al 5%, il valore di ciascun residuo deve essere confrontato con 1,96 o – 1,96, quindi **si usano 2 e – 2.**

Se r > 2 o r < - 2 → il valore è significativo e richiede spiegazione e interpretazione

Se 2 > r > - 2 → il valore non è significativo e non ha bisogno di ulteriore interpretazione

Rivediamo la tabella delle vacanze e dell'orientamento politico

- Usiamo la tabella prodotta da SPSS con frequenze osservate, attese, residui e residui standardizzati
- Poiché il chi quadrato è significativo, è legittimo ricercare quelle celle che contribuiscono più di altre a rendere significativo il valore del chi quadrato

Residui positivi: quelli di sinistra tendono a non cambiare mai luogo di vacanze,

Residui positivi: quelli senza orientamento tendono a cambiare sempre luogo di vacanze

Vacanze * Or_pol Crosstabulation

			Or_pol				Total
			1 NOP	2 sinistra	3 centro	4 destra	
Vacanze	1 quasi mai	Count	127	347	286	290	1050
		Expected Count	149,6	301,9	303,3	295,2	1050,0
		Residual	-22,6	45,1	-17,3	-5,2	
		Std. Residual	-1,8	2,6	-1,0	-,3	
	3 spesso-sempre	Count	94	99	162	146	501
		Expected Count	71,4	144,1	144,7	140,8	501,0
		Residual	22,6	-45,1	17,3	5,2	
		Std. Residual	2,7	-3,8	1,4	,4	
Total		Count	221	446	448	436	1551
		Expected Count	221,0	446,0	448,0	436,0	1551,0

Vacanze * Or_pol Crosstabulation

			Or_pol				Total
			1 NOP	2 sinistra	3 centro	4 destra	
Vacanze	1 quasi mai	Count	127	347	286	290	1050
		Expected Count	149,6	301,9	303,3	295,2	1050,0
		Residual	-22,6	45,1	-17,3	-5,2	
		Std. Residual	-1,8	2,6	-1,0	-,3	
	3 spesso-sempre	Count	94	99	162	146	501
		Expected Count	71,4	144,1	144,7	140,8	501,0
		Residual	22,6	-45,1	17,3	5,2	
		Std. Residual	2,7	-3,8	1,4	,4	
Total		Count	221	446	448	436	1551
		Expected Count	221,0	446,0	448,0	436,0	1551,0

Residui negativi: quelli di sinistra che cambiano spesso o sempre luogo di vacanza sono poco numerosi

Vacanze * Or_pol Crosstabulation

			Or_pol				Total
			1 NOP	2 sinistra	3 centro	4 destra	
Vacanze	1 quasi mai	Count	127	347	286	290	1050
		Expected Count	149,6	301,9	303,3	295,2	1050,0
		Residual	-22,6	45,1	-17,3	-5,2	
		Std. Residual	-1,8	2,6	-1,0	-,3	
	3 spesso-sempre	Count	94	99	162	146	501
		Expected Count	71,4	144,1	144,7	140,8	501,0
		Residual	22,6	-45,1	17,3	5,2	
		Std. Residual	2,7	-3,8	1,4	,4	
Total		Count	221	446	448	436	1551
		Expected Count	221,0	446,0	448,0	436,0	1551,0

Residui nulli: quelli di centro o di destra non mostrano tendenze a cambiare luogo di vacanza. In una tabella di contigenza con questi soli due sottogruppi, il chi quadrato resterebbe non significativo

Residui positivi: quelli di sinistra tendono a non cambiare mai luogo di vacanze,

Residui positivi: quelli senza orientamento tendono a cambiare sempre luogo di vacanze

Vacanze * Or_pol Crosstabulation

			Or_pol				Total
			1 NOP	2 sinistra	3 centro	4 destra	
Vacanze	1 quasi mai	Count	127	347	286	290	1050
		Expected Count	149,6	301,9	303,3	295,2	1050,0
		Residual	-22,6	45,1	-17,3	-5,2	
		Std. Residual	-1,8	2,6	-1,0	-,3	
	3 spesso-sempre	Count	94	99	162	146	501
		Expected Count	71,4	144,1	144,7	140,8	501,0
		Residual	22,6	-45,1	17,3	5,2	
		Std. Residual	2,7	-3,8	1,4	,4	
Total		Count	221	446	448	436	1551
		Expected Count	221,0	446,0	448,0	436,0	1551,0

Residui negativi: quelli di sinistra che cambiano spesso o sempre luogo di vacanza sono poco numerosi

Residui nulli: quelli di centro o di destra non mostrano tendenze a cambiare luogo di vacanza. In una tabella di contigenza con questi soli due sottogruppi, il chi quadrato resterebbe non significativo

ESEMPIO REALE

Nella seguente tabella sono mostrati i residui standardizzati tra titolo di studio del padre e il giudizio di licenza di 1000 studenti.

Tavola di contingenza h8 Titolo di studio del padre * h5 Giudizio di licenza ottenuto in terza media

			h5 Giudizio di licenza ottenuto in terza media				
			1 Sufficiente	2 Buono	3 Distinto	4 Ottimo	Totale
h8 Titolo di studio del padre	1 Licenza elementare	Conteggio	30	27	13	5	75
		Res stand.	2,6	,5	-1,1	-2,4	
	2 Licenza media	Conteggio	100	97	67	38	302
		Res stand.	2,8	-,3	-,5	-2,3	
	3 Scuola professionale	Conteggio	35	50	34	25	144
		Res stand.	-,2	,4	,0	-,3	
	4 Diploma di scuola super	Conteggio	57	105	80	72	314
		Res stand.	-2,4	,1	,6	1,9	
	5 Laurea o diploma universitario	Conteggio	28	51	43	43	165
		Res stand.	-2,1	-,5	,6	2,3	
Totale		Conteggio	250	330	237	183	1000

Il chi quadrato è 48,015, con 12 gradi di libertà p< 0,0005.

Tavola di contingenza h8 Titolo di studio del padre * h5 Giudizio di licenza ottenuto in terza media

			h5 Giudizio di licenza ottenuto in terza media				
			1 Sufficiente	2 Buono	3 Distinto	4 Ottimo	Totale
h8 Titolo di studio del padre	1 Licenza elementare	Conteggio	30	27	13	5	75
		Res stand.	2,6	,5	-1,1	-2,4	
	2 Licenza media	Conteggio	100	97	67	38	302
		Res stand.	2,8	-,3	-,5	-2,3	
	3 Scuola professionale	Conteggio	35	50	34	25	144
		Res stand.	-,2	,4	,0	-,3	
	4 Diploma di scuola super	Conteggio	57	105	80	72	314
		Res stand.	-2,4	,1	,6	1,9	
	5 Laurea o diploma universitario	Conteggio	28	51	43	43	165
		Res stand.	-2,1	-,5	,6	2,3	
Totale		Conteggio	250	330	237	183	1000

I padri con titolo di studio elevato tendono ad avere studenti con giudizio elevato e viceversa.

Solamente per i padri con la scuola professionale non si può inferire una tendenza.

Chi ha buono o distinto non presenta delle indicazioni utili per indovinare il titolo di studio del padre.

Chi ha ottimo tende ad avere un padre con laurea

Gli studenti con sufficiente tendono ad avere un padre con licenza media o elementare

L'associazione fra le due caratteristiche è accertata. MA…

- Si può dire che una delle due *causa* l'altra? (c'è un effetto di causalità o causazione ?)

Quali domande sono legittime?

- Si può affermare che il giudizio di licenza del figlio è *causa* del livello di istruzione del padre?
- Si può affermare che il livello di istruzione del padre è *causa* del giudizio di licenza del figlio?
- Si può affermare che il diploma di terza media del figlio è indipendente dal livello di istruzione del padre?

La correlazione – accertata – non implica la causazione né il suo verso: la relazione di causa-effetto deve essere sempre accertata con la teoria, e mai con una tecnica statistica

Per teoria si intende tutto l'insieme di conoscenze disponibili: ragionamento, osservazioni, sperimentazione, logica, buon senso ...