# Sono tanti, giovani e bravi… saran poi promossi?

di G. Baruzzo, A. D'Arpino, P. Ranzani

Area tematica

Dati e previsioni

Autori
G. Baruzzo, A. D'Arpino, P. Ranzani

Ordine di scuola

Scuola secondaria di secondo grado – classe IV

Tempo medio per svolgere il percorso

8/10 ore

## Sommario

## Scheda generale

**Nucleo tematico**

Dati e previsioni

**Autori**

G. Baruzzo, A. D'Arpino, P. Ranzani

**Tematica affrontata**

Distribuzione statistica doppia di frequenze, connessione distributiva, dipendenza in media.

**Descrizione**

L'attività si inserisce in un percorso pluridisciplinare (ad esempio: italiano, informatica, matematica, economia,..). Nell'ambito scientifico matematico, l'insegnante si propone di condurre gli studenti ad evidenziare l'importanza dello studio della connessione tra due caratteristiche di natura qualitativa e/o quantitativa in una distribuzione doppia di frequenze. In questa proposta sono stati utilizzati i dati su due valutazioni scolastiche prese a distanza di tempo sugli stessi studenti. I dati consentono la costruzione della distribuzione doppia di frequenze e la sua lettura, ed inoltre danno la possibilità di introdurre il concetto di indipendenza distributiva e di calcolare la misura della connessione fino all'analisi della dipendenza in media.

**Ordine di scuola**

Secondaria di secondo grado: secondo biennio (preferibilmente classe quarta).

**Tempo medio per svolgere l'attività in classe**

Indicativamente 8/10 ore

**Nodi concettuali**

- Distribuzioni statistiche rispetto a due caratteri (distribuzioni doppie): unitarie e di frequenze.
- Connessione fra due caratteri in una distribuzione doppia di frequenza: confronto fra distribuzioni condizionate e marginali.

## Indicazioni curricolari

Le attività M@t.abel hanno precisi obiettivi di apprendimento che rientrano tra quelli inseriti nelle Indicazioni nazionali attualmente in vigore (D.M. n. 211 del 07/10/2010, Direttiva n. 4 del 16/01/2012, Direttiva n. 5 del 16/01/2012) e nelle Prove INVALSI. All'inizio di ciascuna attività sono riportati, perciò, i relativi riferimenti presenti nelle Indicazioni nazionali e alcuni quesiti delle Prove Invalsi che ripropongono la situazione stimolo dell'attività considerata. Una domanda Invalsi può aiutare a valutare se gli allievi hanno sviluppato, attraverso lo svolgimento dell'attività, la capacità di utilizzare la matematica per rispondere a domande in una situazione specifica. Le domande sono tratte tra quelle presenti nei vari livelli scolastici, in quanto le attività M@t.abel sono pensate in un'ottica di verticalità.

**Indicazioni Nazionali per i Licei**

**Linee generali e competenze**

Concetti e metodi che saranno obiettivo dello studio:

- la conoscenza elementare di alcuni sviluppi della matematica moderna, in particolare degli elementi del calcolo delle probabilità e dell'analisi statistica.

**Obiettivi specifici di apprendimento**

*Dati e previsioni*

Lo studente, in ambiti via via più complessi, il cui studio sarà sviluppato il più possibile in collegamento con le altre discipline e in cui i dati potranno essere raccolti direttamente dagli studenti, apprenderà a far uso delle distribuzioni

doppie condizionate e marginali, dei concetti di deviazione standard, dipendenza, […]. Studierà la probabilità condizionata e composta, la formula di Bayes e le sue applicazioni, nonché gli elementi di base del calcolo combinatorio. In relazione con le nuove conoscenze acquisite approfondirà il concetto di modello matematico.

**Linee Guida per gli Istituti Professionali e Tecnici**

*Dati e previsioni*

Le **competenze** acquisite consentiranno allo studente di:

- utilizzare il linguaggio e i metodi della matematica per organizzare e valutare adeguatamente informazioni qualitative e quantitative;
- utilizzare i concetti e i modelli delle scienze sperimentali per investigare fenomeni sociali […] per interpretare dati.

**Conoscenze**

- Concetto e rappresentazione grafica delle distribuzioni doppie di frequenze.
- Concetti di dipendenza […].

**Abilità**

- Analizzare distribuzioni doppie di frequenze. Classificare dati secondo due caratteri, rappresentarli graficamente e riconoscere le diverse componenti delle distribuzioni doppie.

## Prove INVALSI

***a.s. 2013/2014 - Domanda D12***

*Scuola secondaria di II grado – Classe II*

Il quesito può essere considerato propedeutico alle tematiche svolte nell'unità.

**D12.** **È stato effettuato un sondaggio su un campione di 1 500 donne di età compresa tra i 25 e i 55 anni per conoscere la loro opinione su una rivista mensile dedicata alla salute. Si sono ottenuti i seguenti risultati:**

| | Occupate | Disoccupate |
|---|---|---|
| **Giudizio positivo** | 450 | 276 |
| **Giudizio negativo** | 367 | 407 |

**a.** **Quante sono le donne che hanno espresso un giudizio positivo?**

**Risposta:** .....

**b.** **Quante sono le donne disoccupate intervistate?**

**Risposta:** .....

**c.** **Scegliendo a caso una delle donne intervistate, qual è la probabilità che abbia espresso un giudizio negativo?**

**Risposta:** .....

**d.** **Scegliendo a caso una delle donne intervistate tra quelle che hanno espresso un giudizio positivo, qual è la probabilità che sia una donna occupata?**

**Risposta:** .....

**Soluzione INVALSI**

**D12_a: 726**

**D12_b: 683**

**D12_c: 774/1500 = 129/250 = 0,516 = 51,6%**

**D12_d: 450/726 = 75/121 ≈ 0,6198 cioè circa il 61.98%**

*Commento*

Lo scopo della domanda quello di saper leggere ed utilizzare dati statistici ricavati da una tabella a doppia entrata.

Per rispondere correttamente alle domande **a.** e **b.** è sufficiente saper leggere una tabella a doppia entrata ed eseguire correttamente addizioni tra numeri interi. Per rispondere correttamente alla domanda **c.** è sufficiente pensare alla probabilità come rapporto tra le donne che hanno espresso un giudizio negativo e il totale delle donne intervistate. Anche per rispondere alla domanda **d.** è sufficiente sapere calcolare un rapporto, ma in questo caso è necessario riconoscere che l'insieme universo non è più costituito da tutte le donne intervistate, ma solo dalle donne intervistate che hanno espresso un giudizio positivo. Da ciò segue il significato di evento condizionato, probabilità condizionata e loro utilizzo.

Il quesito può essere utilizzato dal docente anche per introdurre il concetto di dipendenza/connessione tra caratteri fino al calcolo dell'indice di connessione Chi-quadro di Pearson trattato nell'unità.

## Introduzione all'attività

L'insegnante, in accordo con il collega di diritto, legge in classe uno stralcio di un articolo tratto dal rapporto "La qualità Educativa in Veneto" 2012.

Fonte: Regione Veneto.

www.west-info.eu/files/rapporto-La-Qualità-Educativa-in-Veneto.pdf

Si sofferma in particolare sul fatto che, differentemente da quel che accade nei sistemi educativi di altri paesi, la scelta del percorso formativo fortemente caratterizzante avviene per i ragazzi italiani all'età di 13-14 anni, all'uscita dalla terza media.

La prima scelta che gli studenti si trovano ad affrontare, dopo il percorso formativo dell'obbligo comune, è quindi quella se studiare in una scuola superiore, o prepararsi ad una attività professionale tramite i corsi regionali di formazione professionale; nel primo caso, la difficoltà consiste nella scelta della tipologia di scuola secondaria alla quale iscriversi. Questa è una decisione centrale nello sviluppo educativo di ogni studente, poiché tende a caratterizzarne fortemente il percorso successivo.

L'insegnante fa notare che, come dice il rapporto, la domanda d'istruzione si può configurare come il risultato di un complesso numero di fattori, interni ed esterni alla realtà scolastica, che determinano la spinta sociale verso determinati percorsi formativi; una spinta scorretta o sbilanciata può quindi portare un giovane a cadere in svariate difficoltà, già subito dall'inizio del suo percorso formativo. Presenta poi i dati del quinto rapporto regionale sulla dispersione scolastica redatto dall'Ufficio Scolastico Regionale per il Veneto, mettendo in evidenza che più del 38% delle interruzioni scolastiche nelle scuole secondarie di secondo grado, per l'a.s. 2006/07, si verifica al primo anno di studi. Le interruzioni al primo anno rappresentano, nella nostra regione, il 3,6% delle iscrizioni all'anno stesso, con una forte variabilità territoriale.

**Tab. 2.3.1 – Percentuale di interruzioni di percorso scolastico al primo anno sul totale delle interruzioni e sul totale delle iscrizioni al primo anno per provincia. Veneto - A.s. 2006/2007**

| Provincia | Interruzioni al I anno sul totale delle interruzioni | Interruzioni al I anno sul totale delle iscrizioni al I anno |
|---|---|---|
| Verona | 35,3 | 2,5 |
| Vicenza | 36,6 | 3,8 |
| Belluno | 39,9 | 3,0 |
| Treviso | 34,5 | 2,4 |
| Venezia | 40,6 | 3,8 |
| Padova | 42,6 | 5,3 |
| Rovigo | 39,6 | 4,8 |
| **Veneto** | **38,2** | **3,6** |

Fonte: Elaborazioni Regione Veneto - Direzione Sistema Statistico Regionale su dati Ministero dell'Istruzione, dell'Università e della Ricerca e Ufficio Scolastico Regionale per il Veneto

Prendendo lo spunto dalle informazioni lette l'insegnante propone alla classe di fare una verifica sulla "efficacia" della scelta scolastica degli iscritti nella scuola verificando se il voto finale della scuola media influenza in qualche modo il giudizio dello scrutinio di giugno della classe prima. Chiede agli studenti di ricordare il proprio voto finale della scuola media e il giudizio dello scrutinio di giugno della classe prima e domanda alla classe se, secondo loro, può esistere un qualche "legame" tra le due valutazioni.

## Attività

### Fase 1- Distribuzione doppia di frequenze

Per affrontare il problema proposto l'insegnante osserva che è necessario rilevare il voto finale della scuola secondaria di 1° grado (carattere quantitativo che si esprime con le modalità 6, 7, 8, 9, 10) e l'esito dello scrutino di giugno (carattere qualitativo ordinato che si esprime con le modalità decrescenti: Promosso, Sospeso in 1 materia, Sospeso in 2 materie, Sospeso in 3 materie, Respinto) su un numero di studenti maggiore di quello dei presenti nella classe per avere una risposta più attendibile perché, essendo ottenuta su un numero più ampio di osservazioni, è più generalizzabile. Fornisce perciò alla classe la

matrice dei dati contenenti le informazioni oggetto di studio riferite agli studenti iscritti nelle classi prime nell'anno precedente recuperate nella segreteria didattica e chiede alla classe di organizzare i dati in modo conveniente usando, eventualmente, lo strumento delle tabelle pivot di un foglio elettronico.
Con questi dati è possibile costruire una distribuzione doppia di frequenze che si presenta come nella Tabella 1.

| **Studenti delle classi prime dell'a.s. 2010/11 per voto finale della scuola secondaria di 1° grado ed esito dello scrutino di giugno 2011 (frequenze assolute)** | | | | | |
|---|---|---|---|---|---|
| **Voto finale** | **Esito scrutino di giugno $y_j$** | | | | |
| $x_i$ | Promosso | Sosp_1 | Sosp_2 | Sosp_3 | Respinto |
| 6 | 9 | 6 | 5 | 5 | 38 |
| 7 | 19 | **4** | 11 | 8 | 11 |
| 8 | 36 | 2 | 2 | | 2 |
| 9 | 23 | 1 | | | |
| 10 | 6 | | | | |

Tabella 1

*Fonte: dati forniti dalla segreteria didattica Itis "C. Zuccante" di Mestre (Ve)*

**Nota per l'insegnante**

*Se la classe non ha mai costruito tabelle doppie di frequenza è necessario aiutare gli studenti in questa attività, prendendo spunto ad esempio dalle attività "I giovani e la musica" oppure "Pivot è bello".*

L'insegnante sollecita e guida gli studenti alla scoperta del significato degli elementi che compaiono nella Tabella 1. In particolare si tratta di riconoscere le unità statistiche osservate, i caratteri rilevati e le loro modalità, il significato di alcune frequenze congiunte. Ad esempio l'insegnante domanda cosa indica il numero 4 indicato in grassetto nella tabella e la classe dovrebbe rispondere che è il numero degli studenti che hanno avuto voto finale 7 nella scuola secondaria di 1° grado e che sono stati sospesi in una materia allo scrutinio di giugno. L'insegnante invita gli studenti ad osservare come si dispongono le frequenze nelle celle, ad esempio non tutte le celle sono piene, anzi si può dire che gli incroci sotto la diagonale principale sono, in gran parte vuoti. Questo può avvalorare l'ipotesi che ci possa essere relazione fra voto finale della scuola secondaria di 1° grado ed esito dello scrutinio di giugno .

L'insegnante suggerisce poi di fissare l'attenzione sulla modalità 8 del carattere "Voto finale della secondaria di 1° grado" e chiede: quanti sono gli studenti che hanno avuto tale valutazione? Qual è la distribuzione degli studenti che hanno conseguito 8 rispetto al carattere "Esito dello scrutinio di giugno"? Si tratta di una distribuzione semplice? Da chi dipende? L'insegnante conduce gli studenti a scrivere la distribuzione richiesta:

| Voto finale | Promosso | Sosp_1 | Sosp_2 | Sosp_3 | Respinto | Totale |
|---|---|---|---|---|---|---|
| 8 | 36 | 2 | 2 | | 2 | 42 |

Che titolo diamo a questa distribuzione? Quante sono le distribuzioni del carattere "Esito dello scrutinio di giugno" condizionatamente alle diverse modalità del "Voto finale della secondaria di 1° grado"?

L'insegnante chiama queste: **distribuzioni condizionate** in termini di frequenze assolute.

In modo analogo e come rinforzo del concetto in discussione, l'insegnante può opportunamente lavorare anche sulle 5 distribuzioni condizionate del carattere

"Voto finale nella scuola secondaria di 1° grado" rispetto al carattere "Esito dello scrutinio di giugno".

L'insegnante invita a fare altre considerazioni sulla Tabella 1 e chiede: cosa si ottiene sommando tutti gli elementi di una riga? E tutti quelli di una colonna? L'insegnante guida gli studenti a rendersi conto che: da una distribuzione congiunta di frequenze assolute associate a due caratteri, si ottengono in modo univoco le due distribuzioni semplici dei due caratteri (chiamate marginali) rispetto ai quali si sono classificate congiuntamente le unità statistiche.

In questa situazione si ottengono rispettivamente: la distribuzione degli studenti delle classi prime a.s. 2010/11 rispetto all'"Esito dello scrutinio di giugno":

| $x_i$ | **Promosso** | **Sosp_1** | **Sosp_2** | **Sosp_3** | **Respinto** | **Totale** |
|---|---|---|---|---|---|---|
| $n_i$ | 93 | 13 | 18 | 13 | 51 | 188 |

e la distribuzione degli studenti delle classi prime a.s. 2010/11 rispetto al "Voto finale della secondaria di 1° grado":

| $y_j$ | $n_j$ |
|---|---|
| 6 | 63 |
| 7 | 53 |
| 8 | 42 |
| 9 | 24 |
| 10 | 6 |
| **Totale** | 188 |

L'insegnante chiede alla classe di completare la tabella 1 con le frequenze marginali dei due caratteri osservati, posizionandole rispettivamente nell'ultima riga e nell'ultima colonna che sono indicate con l'"etichetta" Totale.

| **Studenti delle classi prime dell'a.s. 2010/11 per voto finale della scuola secondaria di 1° grado ed esito dello scrutino di giugno 2011 (frequenze assolute)** | | | | | | |
|---|---|---|---|---|---|---|
| **Voto finale** | **Esito scrutino di giugno** | | | | | |
| | **Promosso** | **Sosp_1** | **Sosp_2** | **Sosp_3** | **Respinto** | **Totale** |
| 6 | 9 | 6 | 5 | 5 | 38 | 63 |
| 7 | 19 | 4 | 11 | 8 | 11 | 53 |
| 8 | 36 | 2 | 2 | | 2 | 42 |
| 9 | 23 | 1 | | | | 24 |
| 10 | 6 | | | | | 6 |
| **Totale** | 93 | 13 | 18 | 13 | 51 | 188 |

Tabella 2

L'insegnante, allo scopo di formalizzare i concetti emersi dalla discussione, fornisce alla classe lo schema generale di una distribuzione doppia di frequenze assolute (*X, Y*).

| ***X*** | ***Y*** | | | | | | **Totale** |
|---|---|---|---|---|---|---|---|
| | $y_1$ | $y_2$ | .. | $y_J$ | .. | $y_k$ | |
| $x_1$ | $n_{11}$ | $n_{12}$ | .. | $n_{1J}$ | .. | $n_{1k}$ | $n_{1\bullet} = \sum_{j=1}^{k} n_{1j}$ |
| $x_2$ | $n_{21}$ | $n_{22}$ | .. | $n_{2J}$ | .. | $n_{2k}$ | $n_{2\bullet} = \sum_{j=1}^{k} n_{2j}$ |
| .. | .. | .. | .. | .. | .. | .. | .. |
| $x_i$ | $n_{i1}$ | $n_{i2}$ | .. | $n_{iJ}$ | .. | $n_{ik}$ | $n_{i\bullet} = \sum_{j=1}^{k} n_{ij}$ |
| .. | .. | .. | .. | .. | .. | .. | .. |
| $x_h$ | $n_{h1}$ | $n_{h2}$ | .. | $n_{hJ}$ | .. | $n_{hk}$ | $n_{h\bullet} = \sum_{j=1}^{k} n_{hj}$ |
| **Totale** | $n_{\bullet 1} = \sum_{i=1}^{h} n_{i1}$ | $n_{\bullet 2} = \sum_{i=1}^{h} n_{i2}$ | | $n_{\bullet j} = \sum_{i=1}^{h} n_{ij}$ | | $n_{\bullet k} = \sum_{i=1}^{h} n_{ik}$ | $n \bullet \bullet = n$ |

Scarica la tabella con indicazioni e nomi

**Nota per l'insegnante**

*Questo tipo di formalizzazione è nuova per gli studenti. L'uso dei deponenti con due indici che scorrono contemporaneamente può inizialmente generare confusione. Può valere la pena utilizzare questo momento iniziale per soffermarsi su questi simboli, ne potrà trarre utilità non solo lo sviluppo successivo di questa unità, ma anche l'eventuale avvio alle matrici e al calcolo matriciale, così come l'uso di un qualsiasi foglio di calcolo.*

## Fase 2 - Frequenze relative congiunte e marginali

L'insegnante chiede se abbia senso costruire una distribuzione doppia di frequenze relative. Quali informazioni essa è in grado di fornire? Gli studenti propongono la seguente tabella, dove hanno avuto cura di mantenere 4 decimali per ogni frequenza relativa:

| **Studenti delle classi prime dell'a.s. 2010/11 per voto finale della scuola secondaria di 1° grado ed esito dello scrutino di giugno 2011 (frequenza congiunta relativa $f_{i,j}$)** | | | | | | |
|---|---|---|---|---|---|---|
| **Voto finale $x_i$** | **Esito scrutino di giugno $y_j$** | | | | | **Totale** |
| | **Promosso** | **Sosp_1** | **Sosp_2** | **Sosp_3** | **Respinto** | |
| 6 | 0,0479 | 0,0319 | 0,0266 | 0,0266 | 0,2021 | 0,3351 |
| 7 | 0,1011 | **0,0213** | 0,0585 | 0,0426 | 0,0585 | 0,282 |
| 8 | 0,1915 | 0,0106 | 0,0106 | | 0,0107 | 0,2234 |
| 9 | 0,1223 | 0,0053 | | | | 0,1276 |
| 10 | 0,0319 | | | | | 0,0319 |
| **Totale** | 0,4947 | 0,0691 | 0,0957 | 0,0692 | 0,2713 | 1,0000 |

Tabella 4

L'insegnante ricorda che nella tabella 1 era stato evidenziato il valore $n_{2,2} = 4$ e chiede di spiegare il significato della frequenza relativa corrispondente: $f_{2,2} = 0{,}0213$.

Per migliorare la lettura della tabella 4 gli studenti propongono di trasformare i dati relativi in dati percentuali e, rispondendo alla domanda posta, osservano che il 2,13 percento degli studenti scrutinati sono quelli che presentano X = 7 e Y = Sosp_1.

| **Studenti delle classi prime dell'a.s. 2010/11 per voto finale della scuola secondaria di 1° grado ed esito dello scrutino di giugno 2011 (percentuali sul totale)** | | | | | | |
|---|---|---|---|---|---|---|
| **Voto finale** $x_i$ | **Esito scrutino di giugno ($y_j$)** | | | | | **Totale** |
| | **Promosso** | **Sosp_1** | **Sosp_2** | **Sosp_3** | **Respinto** | |
| 6 | 4,79 | 3,19 | 2,66 | 2,66 | 20,21 | 33,51 |
| 7 | 10,11 | **2,13** | 5,85 | 4,26 | 5,85 | 28,20 |
| 8 | 19,15 | 1,06 | 1,06 | | 1,06 | 22,34 |
| 9 | 12,23 | 0,53 | | | | 12,76 |
| 10 | 3,19 | | | | | 3,19 |
| **Totale** | 49,47 | 6,91 | 9,57 | 6,92 | 27,13 | 100,00 |

Tabella 5

L'insegnante fa notare che il calcolo delle frequenze percentuali arricchisce l'informazione fornita dalla tabella delle frequenze assolute, in quanto permette di cogliere il peso relativo di ciascuna coppia di modalità sul totale, rendendo così più agevole l'interpretazione della tabella stessa. L'insegnante invita gli studenti ad indicare il significato di altre percentuali: qual è la percentuale di studenti che hanno avuto come Voto finale 8 alla scuola secondaria di 1° grado e sono stati Promossi in seconda? Qual è la percentuale di coloro che hanno

conseguito 6 alla scuola secondaria di 1° grado e sono stati respinti? Qual è la percentuale dei "non promossi" a giugno? E la percentuale dei respinti?

## Fase 3 - Distribuzioni condizionate

L'insegnante chiede se è possibile calcolare, a partire dalla Tabella 1 ulteriori distribuzioni di frequenze relative, diverse da quelle emerse nella Fase 2 e di spiegare quali informazioni forniscono.

Alcuni studenti propongono di dividere ogni frequenza congiunta per il totale di riga: $\frac{n_{i,j}}{n_{i,\bullet}}$; altri di dividerlo per il totale di colonna: $\frac{n_{i,j}}{n_{\bullet,j}}$. L'insegnante chiede di esprimersi sul significato dei valori ottenuti trasformando i dati iniziali, eventualmente, in percentuale.

Il primo gruppo afferma che il rapporto rappresenta la quota di studenti con voto finale *xi* che hanno conseguito l'esito *yj* allo scrutinio di giugno e l'insegnante la definisce come **frequenza relativa del carattere *Y* condizionata ad una modalità del carattere *X***. Gli altri studenti dicono che il loro rapporto rappresenta la quota di studenti con esito *yj* allo scrutinio di giugno che hanno conseguito voto finale *xi*. e, in questo caso, l'insegnante lo definisce come **frequenza relativa del carattere *X* condizionata ad una modalità del carattere *Y***.

Dopo una articolata discussione, l'insegnante fa notare che, l'uso delle frequenze relative condizionate, permette il confronto tra i diversi sottogruppi in cui il collettivo totale può essere suddiviso e può aiutare ad evidenziare l'esistenza di un legame tra "Voto finale della scuola secondaria di 1° grado" e "Esito dello scrutinio di giugno".

L'insegnante propone di confrontare il numero degli studenti promossi fra quelli che hanno avuto 7 come voto finale con quelli che hanno avuto 9.

L'insegnante chiede se le due frequenze assolute sono commensurabili. Ne dovrebbe nascere una discussione matura dalla quale dovrebbe emergere chiaramente che il voto finale 7 e il voto finale 9 creano nel collettivo degli studenti due sottoinsiemi diversamente numerosi, perciò occorre eliminare

ciò che è diverso prima di poter effettuare confronti: a questo punto il ricorso alle frequenze relative o percentuali è d'obbligo. Dagli studenti emerge, in conseguenza, la considerazione che la soluzione implica il confronto fra le due frequenze relative: $\frac{19}{53} = 0{,}3585$ e $\frac{23}{24} = 0{,}9583$, e da tali rapporti la classe può dedurre che un "buon" voto finale dalla scuola secondaria di 1° grado "aiuta" nella promozione alla fine del primo anno della secondaria superiore. Questo risultato fa parte della loro esperienza o li sorprende?

L'insegnante, assicuratosi che il concetto sia ben compreso, formalizza tali rapporti, frequenze relative condizionate, con la seguente simbologia:

$f(Y = \text{promosso}/X = 7) = f(y_1/x_2) = \frac{n_{2,1}}{n_{2\bullet}} = \frac{19}{53} = 0{,}3585$ e $f(Y = \text{promosso}/X = 9) = f(y_1/x_4) = \frac{n_{4,1}}{n_{4\bullet}} = \frac{23}{24} = 0{,}9583$. Invita la classe a costruire la tabella delle distribuzioni condizionate delle frequenze relative di Y rispetto a ciascuna modalità di X.

| **Studenti delle classi prime dell'a.s. 2010/11 per voto finale della scuola secondaria di 1° grado ed esito dello scrutino di giugno 2011 (frequenza condizionata relativa di *Y/X* )** | | | | | | |
|---|---|---|---|---|---|---|
| **Voto finale $x_i$** | **Esito scrutino di giugno $y_j$** | | | | | **Totale** |
| | **Promosso** | **Sosp_1** | **Sosp_2** | **Sosp_3** | **Respinto** | |
| 6 | 0,1429 | 0,0952 | 0,0794 | 0,0794 | 0,6032 | 1,0000 |
| 7 | 0,3585 | 0,0755 | 0,2075 | 0,1509 | 0,2075 | 1,0000 |
| 8 | 0,8571 | 0,0476 | 0,0476 | | 0,0476 | 1,0000 |
| 9 | 0,9583 | 0,0417 | | | | 1,0000 |
| 10 | 1,0000 | | | | | 1,0000 |
| **Totale** | 0,4947 | 0,0691 | 0,0957 | 0,0692 | 0,2713 | 1,0000 |

Tabella 6

L'insegnante fa osservare la riga indicata con "Totale" ed osserva che si tratta della distribuzione marginale delle frequenze relative degli esiti allo scrutinio di

giugno. Ciascuna frequenza relativa è compresa fra il minimo e il massimo della colonna sovrastante. Ad esempio 0,1429<0,4947<1e così via.

**Nota per l'insegnante**

*Se la classe lo consente l'insegnante può dimostrare che ciascun elemento dell'ultima riga della tabella 6 è la media ponderata dei corrispondenti valori della colonna sovrastante, ad esempio, per il primo valore si ha:*

$$\frac{f(y_1/x_1)\cdot n_{1\bullet} + f(y_1/x_2)\cdot n_{2\bullet} + \ldots\ldots + f(y_1/x_h)\cdot n_{h\bullet}}{n} = \frac{\frac{n_{11}}{n_{1\bullet}}\cdot n_{1\bullet} + \frac{n_{21}}{n_{2\bullet}}\cdot n_{2\bullet} + \ldots. + \frac{n_{h1}}{n_{h\bullet}} n_{h\bullet}}{n} = \frac{n_{11} + n_{21} + \ldots.. + n_{h1}}{n} = \frac{n_{\bullet 1}}{n}$$

*che rappresenta in questo caso la frequenza relativa della modalità promosso del carattere Esito dello scrutinio di giugno 2011.*

La Tabella delle frequenze relative condizionate consente una nuova importante (e necessariamente diversa) lettura dei dati e porta ad affermare che "gli studenti con voto finale 10 sono stati tutti promossi nello scrutinio di giugno", mentre la percentuale dei promossi fra gli studenti con voto finale 8 risulta inferiore rispetto a quella di coloro che avevano conseguito voto finale pari a 9. L'informazione fornita dalla lettura superficiale delle frequenze assolute, viene così invertita; come l'esperienza insegna è più verosimile essere promosso partendo da un voto finale di 9, che non da un voto finale di 8. Queste diversità avvalorano l'ipotesi che il voto finale "influisce" sull'esito dello scrutinio di giugno.

**Nota per l'insegnante**

*Qui il termine "influire" non va inteso in senso deterministico: se a ….allora b. Non esiste infatti un legame matematico - una funzione che lega il voto finale e l'esito allo scrutinio - ma un legame di tipo statistico-probabilistico: se il suo voto finale è 8 sono propenso a ritenere molto verosimile che lo studente sia stato promosso.*

## Fase 4 - Rappresentazione delle distribuzioni condizionate

Il legame fra il voto finale e l'Esito dello scrutinio di giugno, può essere letto anche attraverso un grafico. L'insegnante invita la classe a costruire la rappresentazione grafica più idonea ad interpretare il legame fra voto finale della scuola secondaria di 1° grado ed esito dello scrutinio di giugno. Un gruppo di studenti potrebbe proporre le rappresentazioni di Figura 1 e di Figura 2 che evidenziano, per ogni esito dello scrutinio di giugno (yi) la distribuzione dei voti finali riportati dagli studenti al termine della scuola secondaria di 1° grado (X).

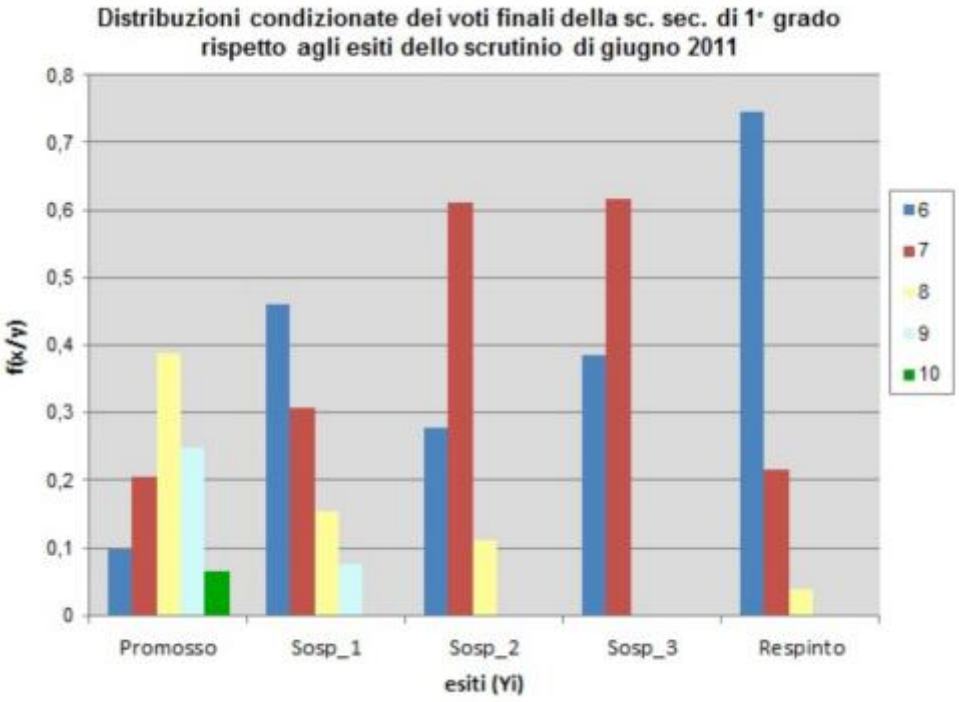



Figura 1

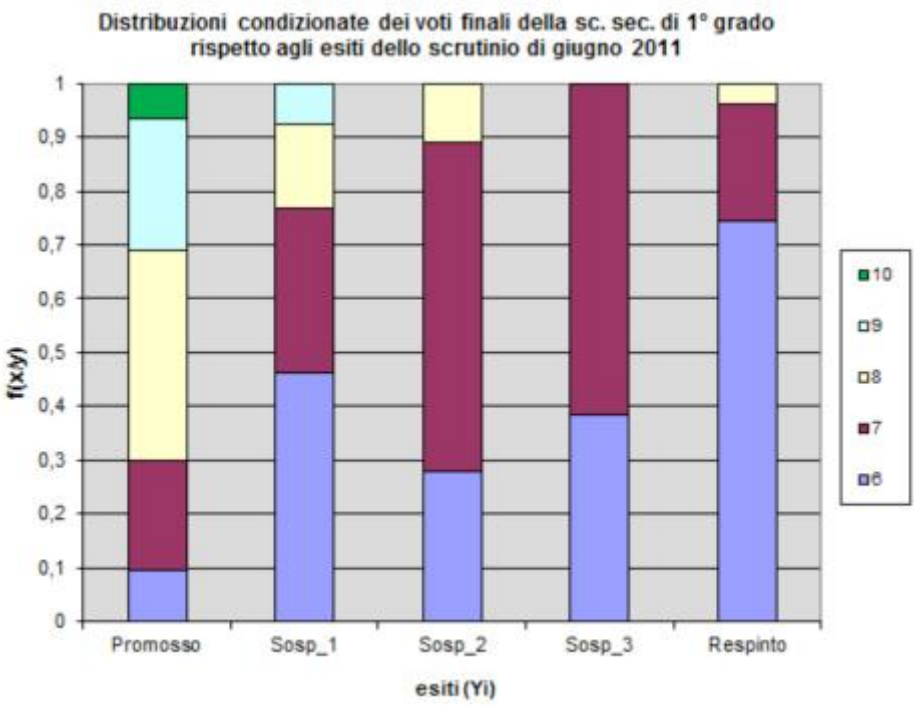



Figura 2

L'insegnante invita gli studenti del gruppo a descrivere le due rappresentazioni. Dalla discussione emerge che entrambe le rappresentazioni mostrano che aumentando il numero delle materie con sospensione del giudizio, fino alla non ammissione alla classe successiva, si "perdono" alcune modalità del carattere X. Ad esempio nei sospesi con 3 materie sono presenti solo studenti che hanno avuto voto finale pari a 6 o a 7. È interessante osservare che tra i respinti vi è qualche studente che aveva ottenuto 8 come voto finale dalla scuola secondaria di 1° grado. L'insegnante fa osservare la differenza fra le due tipologie di grafico ed evidenzia che nella Figura 2 il fatto che le colonne abbiano la stessa altezza (1) mostra in maniera più immediata la diversa ripartizione, in termini relativi, degli studenti rispetto all'esito dello scrutinio di giugno. Altri gruppi di studenti lavorano sulla rappresentazione grafica della distribuzione degli esiti dello scrutinio di giugno condizionato al voto finale ottenuto alla scuola secondaria di 1° grado, proponendo per ogni voto finale la

corrispondente rappresentazione grafica dell'esito di giugno. In seguito se ne riportano ed analizzano due.

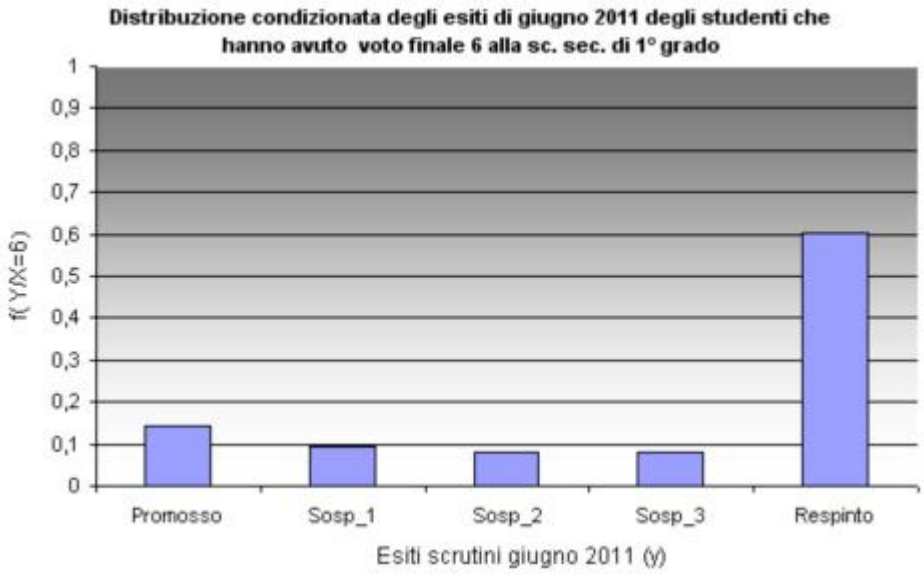



Figura 3

Nel grafico di Figura 3 gli studenti notano che il 60% degli studenti con voto finale 6 sono stati respinti allo scrutinio di giugno. L'informazione era recepibile anche nel grafico di Figura 2 dove nella colonna dei respinti la quota più alta era riferita agli studenti con voto finale 6. L'insegnante fa notare che la parte della colonna dei respinti nella Figura 2 è di altezza diversa da quella della Figura 3 e chiede come ciò sia possibile. La risposta è che sono due frequenze condizionate diverse e precisamente quella della Figura 2 indica

$$f(X = 6/Y = \text{respinto}) = \frac{\text{numero respinti con 6}}{\text{numero totale dei respinti}} = \frac{n_{1,5}}{n_{\cdot 5}} = \frac{38}{51} = 0{,}745$$

mentre quella del grafico 4 è

$$f(Y = \text{respinto} / X = 6) = \frac{\text{numero respinti con 6}}{\text{numero totale studenti con voto finale 6}} = \frac{n_{1,5}}{n_{1\cdot}} = \frac{38}{63} = 0{,}603$$

Il grafico che segue mostra che pure studenti con voto finale 8 sono stati respinti come era già stato evidenziato leggendo il grafico 2.

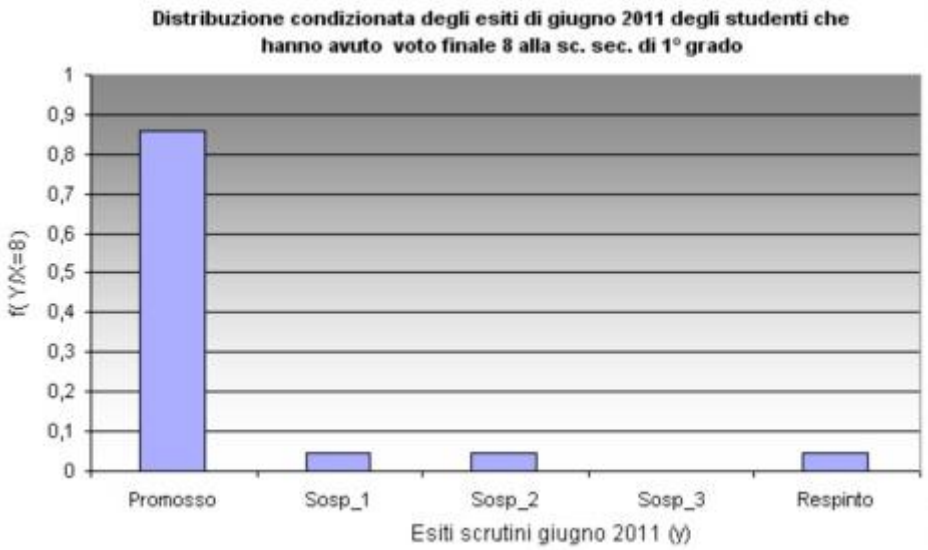



Figura 4

Come sintesi l'insegnante mostra il grafico che segue

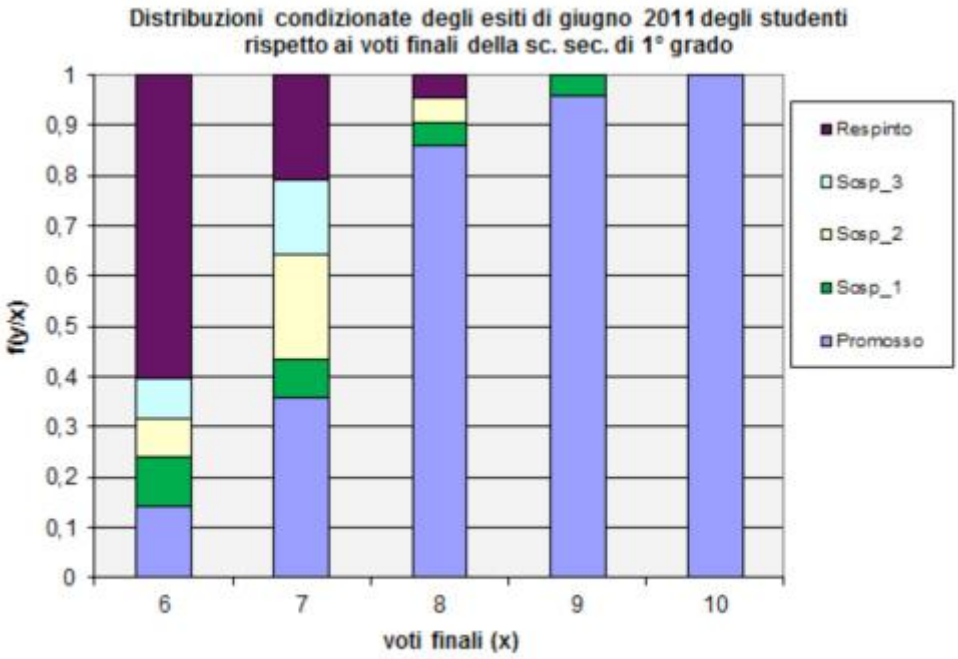



Figura 5

e fa osservare come la diversa ripartizione degli esiti dello scrutinio in ciascuna colonna, evidenzi l'ipotesi di una relazione con il voto finale dalla scuola secondaria di 1° grado.

L'insegnante può anche proporre alla classe di visualizzare, per ogni esito dello scrutinio di giugno, la distribuzione dei voti finali attraverso il box plot ricordando che tale rappresentazione grafica mette in luce la diversa distribuzione dei dati e in particolare consente di visualizzare immediatamente la variabilità dei dati osservati e la sua simmetria rispetto alla mediana.

**Nota per l'insegnante**

*Nel caso in cui la classe è già a conoscenza delle principali nozioni di statistica descrittiva si suggerisce comunque di ripassare le nozioni di base sui percentili e sulla costruzione del box plot oppure, se la classe non possiede tali conoscenze, è possibile reperire le informazioni sui percentili, sulla costruzione del box plot e sua interpretazione sull'unità: Siamo" vincoli o sparpagliati"*

L'insegnante divide la classe in cinque gruppi e ad ognuno assegna il compito di costruire il box plot relativo alla cinque distribuzioni condizionate di ($X/Y = y_i$). Riporta in seguito in uno stesso cartellone i cinque lavori in modo da poterli confrontare e analizzarne le differenze.

Fa osservare che, per esigenze di lettura:

- il valore minimo della distribuzione è posizionato sotto al corrispondente simbolo;
- il valore del 25-esimo percentile (primo quartile) indicato con q1 è posizionato a sinistra del corrispondente simbolo;
- il valore della mediana è posizionato sopra il corrispondente simbolo;

- il valore del 75-esimo percentile (terzo quartile) indicato con q3 è posizionato a sinistra del corrispondente simbolo;
- il massimo della distribuzione è posizionato a destra del corrispondente simbolo.

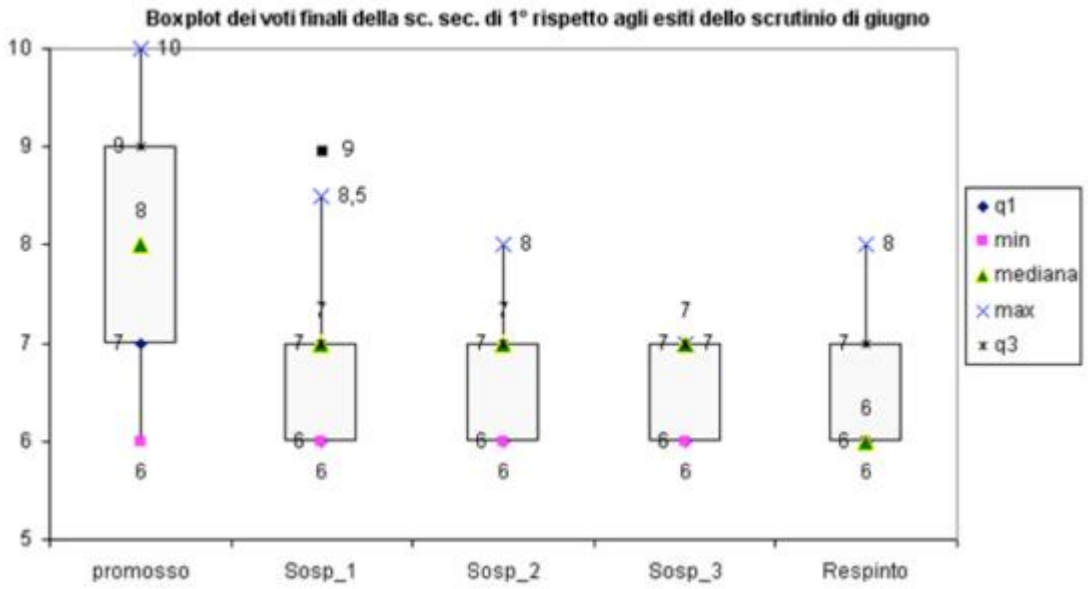



Figura 6

Dall'osservazione dei grafici in Figura 6 emerge una diversa lettura della distribuzione doppia di frequenze, e che ancora una volta ribadisce l'esistenza di una relazione fra le variabili coinvolte in quanto i box plot sono completamente diversi. L'insegnante aiuta gli studenti a commentare il box plot associato ai "sospesi con 1 materia" e quello associato ai "sospesi con 3 materie". Il primo box plot analizzato mostra che il minimo coincide con q1 e quindi almeno il 25% degli studenti era stato promosso con voto finale X=6, che la mediana coincide con q3 e quindi che almeno il 75% degli studenti è stato promosso con voto finale X≤7. Il valore 8,5 non rappresenta il massimo valore assunto dal voto, pertanto il grafico contiene un "punto isolato" in corrispondenza al valore massimo 9.

Vedi l'attività Siamo "vincoli o sparpagliati"?

Il "punto isolato" evidenzia che "pochi studenti" erano stati promossi con voto 9. Nello specifico un solo studente su 13.

Vedi la tabella 2 (fase 1).

[Nel secondo box plot preso in considerazione il 100% degli studenti sospesi con tre materie erano stati promossi con voto compreso tra 6 e 7 e si nota che 7 rappresenta sia la mediana, che q3 che il massimo e quindi è un valore con alta frequenza (8 studenti su 13).

## Fase 5 - Dipendenza o indipendenza?

L'insegnante, prendendo spunto dal grafico di Figura 5 chiede: come dovrebbero essere le colonne se il carattere X non influisse su Y? Se la X non influisce su Y a livello di distribuzione, il variare delle modalità della X dovrebbe lasciare immodificate le distribuzioni condizionate di Y. Di conseguenza le colonne dovrebbero mostrare una uguale suddivisione rispetto alle modalità di Y. Nella tabella 6 le distribuzioni condizionate di Y/xi dovrebbero essere tutte uguali fra loro al variare di xi ed inoltre dovrebbero coincidere con la corrispondente distribuzione marginale del carattere Y.

Usando il linguaggio formale, l'insegnante, fa notare che, ad esempio, se il carattere "Voto finale della scuola secondaria di 1° grado" non avesse influenza sull'"Esito dello scrutinio di giugno" si avrebbero, per i "Promossi", le seguenti situazioni:

$$\frac{n_{1,1}}{n_{1\bullet}} = \frac{n_{2,1}}{n_{2\bullet}} = \frac{n_{3,1}}{n_{3\bullet}} = \frac{n_{4,1}}{n_{4\bullet}} = \frac{n_{5,1}}{n_{5\bullet}}$$

L'insegnante chiede: quanti studenti con voto finale pari ad 8 e promossi nello scrutinio di giugno (n $_{3,1}$) ci sarebbero stati se il voto non influisse sull'esito dello scrutinio?

Suggerisce alla classe di procedere, applicando la proprietà del comporre nelle proporzioni precedentemente scritte, mettendo in evidenza l'elemento che contiene $n_{3,1}$:

$$\frac{n_{1,1}+n_{2,1}+n_{3,1}+n_{4,1}+n_{5,1}}{n_{1,*}+n_{2,*}+n_{3,*}+n_{4,*}+n_{5,*}} = \frac{n_{3,1}}{n_{3,*}}$$

Chiede di specificare il significato del numeratore e del denominatore del primo membro dell'uguaglianza. La classe osserva che il numeratore è il numero dei "Promossi" mentre il denominatore è l'insieme degli studenti scrutinati, cioè:

$$\frac{n_{*1}}{n_{**}} = \frac{n_{3,1}}{n_{3,*}}$$

Dall'uguaglianza si ricava il numero di studenti richiesto, cioè il numero degli studenti con voto finale pari a 8 e promossi a giugno se non ci fosse influenza fra voto finale e esito allo scrutinio:

$$n_{3,1} = \frac{n_{*1} \times n_{3,*}}{n_{**}} = \frac{93 \times 42}{188} = 20,78$$

L'insegnante afferma che questa è la frequenza teorica associata alla coppia di modalità (Voto finale 8; Promosso) nell'ipotesi di **indipendenza in distribuzione** o frequenza teorica in condizione di indipendenza o indifferenza fra i caratteri.

Fa anche notare che la condizione di indipendenza in distribuzione è una relazione simmetrica cioè se *X* è indipendente da *Y* anche *Y* è indipendente da *X*.

**Nota per l'insegnante**

*Può essere opportuno suggerire agli studenti di rivedere la regola per il calcolo della probabilità composta per eventi indipendenti e di confrontare il valore ottenuto in termini di frequenza assoluta con quello trovato in termini di probabilità.*

L'insegnante fa notare che questo ragionamento vale per ogni coppia di modalità e quindi è possibile costruire la tabella teorica di frequenze in condizione di indipendenza, in cui ogni cella soddisfa alla condizione precedente. In tale tabella la frequenza teorica di ogni cella è indicata con $n^*_{i,j}$.

**Tabella di connessione "nulla" fra Voto finale della scuola secondaria di 1° grado ed Esito scrutinio di giugno ($n^*_{i,j}$)**

| Voto finale | Esito scrutino di giugno | | | | | Totale |
|---|---|---|---|---|---|---|
| | Promosso | Sosp_1 | Sosp_2 | Sosp_3 | Respinto | |
| 6 | 31,16 | 4,36 | 6,03 | 4,36 | 17,09 | 63 |
| 7 | 26,22 | 3,66 | 5,07 | 3,66 | 14,38 | 53 |
| 8 | 20,78 | 2,90 | 4,02 | 2,90 | 11,39 | 42 |
| 9 | 11,87 | 1,66 | 2,30 | 1,66 | 6,51 | 24 |
| 10 | 2,97 | 0,41 | 0,57 | 0,41 | 1,63 | 6 |
| **Totale** | 93 | 13 | 18 | 13 | 51 | 188 |

Tabella 7

L'insegnante pone le domande: come sono le distribuzioni marginali della Tabella 1 (tabella osservata) e quelle della Tabella 7 (tabella teorica)? Come mai la Tabella 1 contiene solo numeri interi mentre la Tabella 7 contiene numeri decimali? Come mai nella Tabella 7 tutte le caselle sono piene? È possibile valutare la "distanza" fra le frequenze osservate e quelle teoriche in ipotesi di indipendenza? Se sì, quali proposte si possono fare?

**Nota per l'insegnante**

*Nella frase precedente è stato scritto distanza fra virgolette perché quella definita non è una vera distanza in senso matematico. Ad esempio nel seguito si troveranno anche "distanze" negative, proprietà che una distanza di tipo matematico, non possiede mai.*

## Fase 6 - Indice di dipendenza distributiva (connessione)

Dopo aver lasciato dibattere gli studenti sulle questioni poste, l'insegnante propone una possibile quantificazione di tale "distanza" costruendo, per ogni coppia di modalità ($x_i$, $y_j$) la differenza tra la frequenza osservata e quella calcolata in condizione di indipendenza ($n_{iJ}$ - $n^*_{i,j}$) . Chiama tale differenza Contingenza e la indica con $c_{i,j}$. La classe mostra la seguente tabella di **contingenza** in cui l'insegnante fa evidenziare alcuni valori:

| **Tabella delle differenze (contingenze) tra i dati della Tabella 1 e quelli della Tabella 4 ($c^2_{i,j}$)** | | | | | |
|---|---|---|---|---|---|
| **Voto finale** | **Esito scrutino di giugno** | | | | |
| | **Promosso** | **Sosp_1** | **Sosp_2** | **Sosp_3** | **Respinto** |
| 6 | **-22,16** | 1,64 | -1,03 | 0,64 | **20,91** |
| 7 | -7,22 | 0,34 | 5,93 | 4,34 | -3,38 |
| 8 | **15,22** | -0,90 | -2,02 | -2,90 | **-9,39** |
| 9 | **11,13** | -0,66 | -2,30 | -1,66 | -6,51 |
| 10 | 3,03 | -0,41 | -0,57 | -0,41 | -1,63 |

Tabella 8

I dati della Tabella 8 mettono in luce la presenza di alcune “distanze” (in grassetto) piuttosto elevate rispetto alle altre. Esse segnalano un

allontanamento “consistente” dalla condizione teorica di indipendenza. L’insegnante chiede se la media delle contingenze può fornire un indice che sintetizza la “distanza” tra la distribuzione osservata e quella teorica di indipendenza.

Gli studenti, a conti fatti, osservano che la media risulta pari a zero in quanto la somma delle contingenze è zero. Questo significa che non c’è differenza tra le due distribuzioni?

L’insegnante fa notare che questa conclusione non concorda con l’osservazione fatta nella Fase 3 dove si affermava che le distribuzioni condizionate erano diverse; in effetti, la media risulta pari a zero per costruzione Volendo si può dimostrare che

$$\sum_{i=1}^{h} c_{i,j} = \sum_{j=1}^{k} c_{i,j} = \sum_{i=1}^{h}\sum_{j=1}^{k} c_{i,j} = 0$$

Per evitare di fare intervenire i segni delle contingenze, si potrebbero usare le contingenze in modulo, ottenendo sicuramente un valore medio aritmetico positivo. Ma questo non è il solo modo per arrivare ad un risultato non nullo. L'insegnante propone di utilizzare l'informazione fornita dalle contingenze con un indice basato sul quadrato delle stesse, e fa costruire alla classe la tabella che riporta il quadrato delle contingenze.

| **Quadrati delle contingenze ($c^2_{i,j}$)** | | | | | |
|---|---|---|---|---|---|
| **Voto finale** | **Esito scrutino di giugno** | | | | |
| | **Promosso** | **Sosp_1** | **Sosp_2** | **Sosp_3** | **Respinto** |
| 6 | **491,28** | 2,70 | 1,06 | 0,41 | **437,21** |
| 7 | 52,10 | 0,11 | 35,11 | 18,79 | 11,41 |
| 8 | **231,75** | 0,82 | 4,09 | 8,43 | **88,24** |
| 9 | **123,82** | 0,44 | 5,28 | 2,75 | 42,39 |
| 10 | 9,19 | 0,17 | 0,33 | 0,17 | 2,65 |

Tabella 9

Fa osservare che l'elevamento al quadrato amplifica le differenze "grandi" (in valore assoluto maggiore di 1) e minimizza quelle "piccole" (in valore assoluto minore di 1) e quindi, per recuperare l'ordine di grandezza da cui si è partiti, propone di dividere tali quadrati per la corrispondente frequenza teorica in caso di indipendenza. Ovviamente si sarebbe potuto anche dividere $c_{i,j}$ per ni,J , ma come la Tabella 1 (fase 1) mostra ci sarebbero state celle vuote dove il calcolo del rapporto sarebbe stato impossibile, da qui la proposta della frazione $\frac{c_{i,j}^2}{n_{i,J}^*}$

| $\frac{c_{i,j}^2}{n_{i,J}^*}$ | | | | | |
|---|---|---|---|---|---|
| **Voto finale** | **Esito scrutino di giugno** | | | | |
| | **Promosso** | **Sosp_1** | **Sosp_2** | **Sosp_3** | **Respinto** |
| 6 | **15,76** | 0,62 | 0,18 | 0,10 | **25,58** |
| 7 | 1,99 | 0,03 | 6,92 | 5,13 | 0,79 |
| 8 | **11,15** | 0,28 | 1,02 | 2,90 | **7,74** |
| 9 | **10,43** | 0,26 | 2,30 | 1,66 | 6,51 |
| 10 | 3,10 | 0,41 | 0,57 | 0,41 | 1,63 |

Tabella 10

Ogni rapporto evidenzia uno scostamento relativo del dato osservato rispetto a quello teorico. La somma di tutti i valori riportati in tabella è un indice di connessione noto come **indice di dipendenza chi-quadro di Pearson** la cui espressione generale è data dalla formula:

$$\chi^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{c_{i,j}^2}{n_{i,j}^*}$$

Nel caso analizzato gli studenti trovano per l'indice il valore 107,49.

L'insegnante pone alcune domande stimolo:

1. cosa significa tale valore?
2. la distribuzione osservata si può ritenere "significativamente" diversa da quella teorica? Ossia si può ritenere che la diversità fra le due distribuzioni, quella osservata e quella teorica in ipotesi di indipendenza, non si produca per il solo effetto del caso, ma che sia l'indizio di una effettiva connessione fra i due caratteri che allontana dati osservati da dati teorici?

**Nota per l'insegnante**

*Il punto b) apre al problema dell'inferenza statistica induttiva, che va al di là del livello di scolarità al quale questa unità si riferisce. Pare tuttavia opportuno che il discorso venga avviato, in modo che gli studenti sappiano affrontare concettualmente il problema quando avranno gli strumenti per farlo.*

Per guidare alle risposte fa notare che Chi quadro:

- vale 0 se e solo se vi è indipendenza distributiva fra i caratteri;
- ha un massimo pari al minimo tra i seguenti due prodotti: *n*(h-1);n*(k-1)*, dove n è il numero di unità statistiche osservate, h è il numero delle modalità del carattere X e k è il numero delle modalità del carattere Y.

Dunque Chi quadro è condizionato da *n* e dalle dimensioni della tabella. È chiaro che per un n di media dimensione Chi quadro può assumere valori grandi e lontani da zero perché è calcolato su molte unità. Nel caso studiato Chi-quadro è compreso tra 0 (connessione nulla) e 752 (connessione massima) estremi inclusi.

L'insegnante fa notare che il valore ottenuto (107,49) mostra una connessione debole in quanto è circa 1/7 del valore massimo. A questo punto si può dire che

la connessione fra il voto finale della scuola secondaria di 1° grado e l'esito dello scrutinio di giugno della classe prima è pari al 14% (107,49X100/188x4) circa del valore massimo che l'indice può raggiungere.

L'ultimo valore è esprimibile come rapporto indicato da: $\frac{\chi^2}{\chi^2_{Max}}$ ottenendo in tal modo un numero compreso tra 0 e 1 estremi inclusi. Il rapporto prende il nome di indice **normalizzato di connessione**.

**Nota per l'insegnante**

*Può essere interessante porre alla classe il problema di scrivere, a partire da quella data, una possibile distribuzione doppia con valore di chi quadro sicuramente massimo. Nel nostro caso dovrà essere chi-quadro 752. Il fatto che la tabella sia quadrata agevola la risposta. Infatti in una tabella quadrata la connessione è massima se per ogni riga ed ogni colonna vi è una ed una sola frequenza. Tenendo conto che n=188 e h=k=5 la tabella che segue ha connessione massima.*

| **Voto finale** | **Esito scrutino di giugno** | | | | | **Totale** |
|---|---|---|---|---|---|---|
| | **Promosso** | **Sosp_1** | **Sosp_2** | **Sosp**_3 | **Respinto** | |
| 6 | | | 70 | | | 70 |
| 7 | | | | 85 | | 85 |
| 8 | | | | | 14 | 14 |
| 9 | | 13 | | | | 13 |
| 10 | 6 | | | | | 6 |
| Totale | 6 | 13 | 70 | 85 | 14 | 188 |

## Indicazioni metodologiche

Questa attività può essere introdotta in una classe quarta della scuola secondaria di secondo grado dopo che l'insegnante ha verificato le conoscenze di base sulle nozioni di statistica descrittiva.

È opportuno che l'insegnante aiuti gli studenti a capire la necessità di considerare la classificazione simultanea di un collettivo secondo due caratteristiche per individuare l'eventuale dipendenza distributiva o la dipendenza in media dei caratteri nel collettivo studiato. A tale scopo li può invitare anche a reperire i dati su tematiche a loro vicine o di loro interesse specifico magari in collaborazione con le altre discipline. Nell'analisi delle distribuzioni doppie, marginali e condizionate gli studenti possono lavorare in piccoli gruppi, mettendo a confronto le proprie conoscenze statistiche e matematiche per arrivare a una soluzione comune. Seguirà una discussione e un confronto collettivo per arrivare ad una formalizzazione, da parte dell'insegnante, dei concetti emersi dall'attività compresa la necessità di arrivare al concetto di connessione tra di i due caratteri della distribuzione doppia analizzata.

È in ogni caso opportuno che le situazioni più complicate, che richiedono di trovare gli indici di sintesi utili all'analisi della connessione tra variabili, vengano risolte attraverso l'uso del computer ponendo attenzione agli arrotondamenti e al numero di cifre decimali che interessano. In particolare nelle operazioni che coinvolgono le tabelle doppie, come il calcolo delle frequenze assolute o delle frequenze percentuali, occorre prestare attenzione perché le cifre decimali scelte per l'approssimazione dei risultati possono non far quadrare le somme rispettivamente a 1 o a 100. L'uso del calcolatore, accelerando le attività di calcolo, consente di poter dedicare più tempo all'analisi e alla discussione.

## Spunti per approfondire

**Spunti per un approfondimento disciplinare**

**Si propone la seguente attività:**

Analisi della dipendenza in media tra due variabili

**Fase 1**

**Una diversa analisi delle distribuzioni condizionate**

L'insegnante fa notare alla classe che la distribuzione doppia analizzata nell'attività precedente e riportata in tabella 1

| **Studenti delle classi prime dell'a.s. 2010/11 per Voto finale della scuola secondaria di 1° grado ed esito dello scrutino di giugno 2011 (frequenze assolute)** | | | | | | |
|---|---|---|---|---|---|---|
| **Voto finale $x_i$** | **Esito scrutino di giugno $y_j$** | | | | | **Totale** |
| | **Promosso** | **Sosp_1** | **Sosp_2** | **Sosp_3** | **Respinto** | |
| 6 | 9 | 6 | 5 | 5 | 38 | 63 |
| 7 | 19 | 4 | 11 | 8 | 11 | 53 |
| 8 | 36 | 2 | 2 | | 2 | 42 |
| 9 | 23 | 1 | | | | 24 |
| 10 | 6 | | | | | 6 |
| **Totale** | 93 | 13 | 18 | 13 | 51 | 188 |

Tabella 1

può consentire anche una lettura diversa da quella fatta per lo studio della connessione distributiva (Attività, Fase 5 e Fase 6 ).

Per semplificare lo studio, propone di raggruppare alcune modalità degli esiti finali in modo da distinguere solo: promossi, sospesi e respinti. Si ottiene così la tabella che segue:

| **Studenti delle classi prime dell'a.s. 2010/11 per Voto finale della scuola secondaria di 1° grado ed esito dello scrutino di giugno 2011 (frequenze assolute)** | | | | |
|---|---|---|---|---|
| **Voto finale $x_i$** | **Esito scrutino di giugno $y_j$** | | | **Totale** |
| | **Promosso** | **Sospeso** | **Respinto** | |
| 6 | 9 | 16 | 38 | 63 |
| 7 | 19 | 23 | 11 | 53 |
| 8 | 36 | 4 | 2 | 42 |
| 9 | 23 | 1 | | 24 |
| 10 | 6 | | | 6 |
| **Totale** | 93 | 44 | 51 | 188 |

Tabella 2

L'insegnante ricorda alla classe che, per stabilire se nella Tabella 1 fosse proponibile ipotizzare l'esistenza di indipendenza distributiva fra i due caratteri, era stata costruita la Figura 6 dove i box plot, contraddicendo l'ipotesi proposta, mostrano che ogni modalità di esito dello scrutinio presenta una diversa distribuzione condizionata dei voti finali. La Tabella 2 ammette una lettura dello stesso tipo che fa corrispondere la distribuzione condizionata dei voti finali ad ogni esito allo scrutinio. Se si trattano i voti finali come caratteri quantitativi, oltre che mediana e quartili, sulle distribuzioni condizionate si possono calcolare media aritmetica, varianza e scostamento quadratico medio. L'insegnante, dopo aver fatto scrivere le tre distribuzioni dei voti finali condizionati agli esiti, chiede agli studenti, riuniti in piccoli gruppi, di calcolare media aritmetica e varianza di ciascuna distribuzione e di raccogliere i risultati ottenuti in una tabella doppia.

Calcolate le sintesi, l'insegnante fa riflettere gli studenti sulla necessità di costruire nuovi simboli per identificarli. In particolare è necessario un simbolo che leghi la media aritmetica alla distribuzione condizionata sulla quale è stata calcolata. Così suggerisce che la media aritmetica della distribuzione $X/Y = y_i$. venga indicata con: $\frac{\chi^2}{\chi^2_{Max}}$. Rispetto alla Tabella 2 si ottengono in tal modo i tre valori medi:

$\bar{x}_1 = \bar{x}_{Promosso} = 7,978$

$\bar{x}_2 = \bar{x}_{Sospeso} = 6,773$

$\bar{x}_3 = \bar{x}_{Respinto} = 6,294$

Dalla discussione, gli studenti evidenziano che nel collettivo in esame il voto finale medio diminuisce man a mano che peggiora l'esito dello scrutinio e dunque l'esito dello scrutinio può essere proposto come indicatore del voto dell'esito finale.

In modo analogo è possibile indicare la varianza di ogni distribuzione condizionata con un simbolo che la colleghi alla stessa: la varianza calcolata sulla distribuzione condizionata $X/Y = y_i$. si indica con $\alpha^2_{x/y1}$. Rispetto alla Tabella 2 si ottengono le tre varianze:

$\sigma^2_{X/y_1} = \sigma^2_{X/Promosso} = 1,096$

$\sigma^2_{X/y_2} = \sigma^2_{X/Sospeso} = 0,494$

$\sigma^2_{X/y_3} = \sigma^2_{X/Respinto} = 0,286$

Le distribuzioni condizionate del voto finale non solo presentano medie diverse, ma anche variabilità diversa, che si riduce, man a mano che l'esito dello scrutinio peggiora.

Raccogliendo le sintesi calcolate in una tabella doppia si ha:

| | Promosso | Sospeso | Respinto |
|---|---|---|---|
| $\overline{x_j}$ | 7,978 | 6,773 | 6,294 |
| $\alpha^2_{x/y1}$ | 1,096 | 0,494 | 0,286 |
| $\alpha_{x/y1}$ | 1,047 | 0,703 | 0,535 |
| $n._j$ | 93 | 44 | 51 |

Tabella 3

Per una migliore comprensione dei risultati finora ottenuti, l'insegnante mostra il grafico di Figura 1 che riporta, per ogni Esito dello scrutinio di giugno il range dei voti finali ottenuti alla scuola secondaria di 1° grado, il loro valore medio $\overline{x_j}$ e l'intervallo $\overline{x_j} \pm \alpha_{x/y1}$ evidenziato dall'altezza del rettangolo, costruito su una base arbitraria ma identica per ogni modalità condizionante.

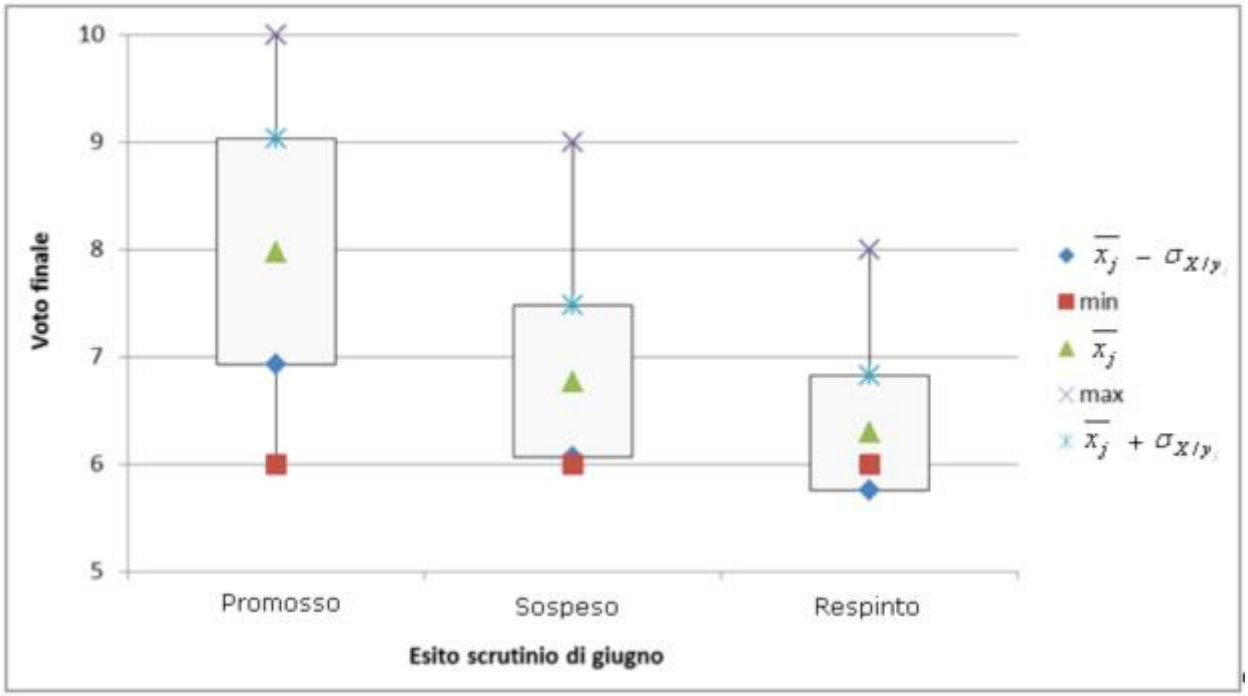



Figura 1

Fa notare come gli studenti promossi avevano mediamente voti finali migliori rispetto agli studenti sospesi e respinti; che le tre distribuzioni condizionate hanno variabilità diversa evidenziata dalla diversa altezza dei rettangoli e dalle diverse lunghezze dei segmenti che uniscono il rettangolo agli estremi della distribuzione.
Gli studenti possono ricavare, osservando il grafico ed utilizzando i dati della Tabella 2, alcune informazioni:

- che tutti i respinti avevano ottenuto voti da 6 a 8;
- che dei 44 sospesi, 43 avevano ottenuto voti tra il 6 e l'8;
- che dei 93 promossi, 64 avevano ottenuti voti finali nella stessa fascia (6|-|8);

per quanto riguarda la parte di studenti con voto compreso nell'intervallo $\overline{x_j} \pm \alpha_{x/y1}$ gli studenti osservano che:

- 38 respinti su 51 avevano ottenuto voto finale compreso tra 5,76 e 6,83. L'insegnante fa notare che il valore 5,76 è il risultato della differenza tra la media e la deviazione standard mentre nella realtà il loro voto minimo è stato pari a 6;
- che 23 dei 44 sospesi avevano ottenuto voto finale compreso tra 6,07 e 7,48;
- che 78 dei 93 promossi avevano ottenuto voto finale compreso tra 6,93 e 9,03.

Dalla discussione gli studenti devono arrivare ad evidenziare che le medie condizionate cambiano al variare della modalità di *Y* e ciò li porta ad affermare che la variabile *X* dipende in media da *Y*. Anche in termini di variabilità le distribuzioni del voto finale variano al variare dell'Esito dello scrutinio di giugno evidenziando che i promossi sono quelli che hanno ottenuto esito medio più alto, manifestando anche una variabilità più elevata.

**Nota per l'insegnante**

*Per descrivere in modo puntale la diversità della variabilità delle distruzioni condizionate conviene utilizzare il coefficiente di variazione dato dal rapporto tra la deviazione standard e il valor medio corrispondente viste anche le diverse numerosità dei sottogruppi. (Vedi l'attività Siamo vincoli o sparpagliati ).*
*Se la classe lo consente l'insegnante può anche proporre la trattazione formale delle sintesi delle distribuzioni condizionate utilizzando la seguente simbologia:*

- $\overline{x_j}$ *, la media della distribuzione condizionata di X/Y=$y_j$, si calcola mediante la formula:*

$$\overline{x_j} = \frac{\sum_{i=1}^{h} x_i \cdot n_{ij}}{n_{\bullet j}} = \sum_{i=1}^{h} x_i \cdot f_{i/y_j}$$

- $\sigma^2_{x/y1}$*,la varianza della distribuzione condizionata di X/Y=$y_j$ si calcola mediante la formula:*

$$\sigma^2_{X/y_j} = \frac{\sum_{i=1}^{h} \left(x_i - \overline{x_j}\right)^2 \cdot n_{ij}}{n_{\bullet j}} = \sum_{i=1}^{h} \left(x_i - \overline{x_j}\right)^2 \cdot f_{i/y_j}$$

**Fase 2**

**Dipendenza in media e calcolo del rapporto di correlazione**

Dopo le considerazioni precedenti l'insegnante chiede alla classe: è possibile calcolare il valor medio delle medie condizionate? Che relazione c'è fra quest'ultimo e la media di *X* ($\bar{x}$)?

Fa quindi soffermare la classe sul significato delle medie condizionate calcolate:

- 7,978 è il voto medio finale dei 93 studenti che sono stati promossi a giugno;
- 6,773 è il voto medio finale dei 44 studenti che sono stati sospesi a giugno;
- 6,254 è il voto medio finale dei 51 studenti che sono stati respinti a giugno.

Ciò da luogo ad una distribuzione di frequenze dove le modalità sono rappresentate dalle medie condizionate stesse e le frequenze ad esse associate sono le frequenze delle corrispondenti modalità del carattere condizionante, qui l'Esito dello scrutinio di giugno. La distribuzione delle medie condizionate è presentata in Tabella 4.

| **Distribuzione di frequenza delle medie condizionate di *X* rispetto a *Y*** | |
|---|---|
| $\overline{x_j}$ | **n.** j |
| 7,978 | 93 |
| 6,773 | 44 |
| 6,294 | 51 |
| n | 188 |

Tabella 4

L'insegnante, utilizzando i dati della Tabella 4, fa calcolare la media della distribuzione delle medie condizionate:

$$M(\overline{x_j}) = \frac{7,978*93+6,773*44+6,294*51}{188} = 7,239.$$

**Nota per l'insegnante**

*Soffermarsi sul significato della distribuzione delle medie condizionate è fondamentale per evitare un misconcetto presente non solo negli studenti. In effetti, per il calcolo del valore medio delle medie condizionate qualche studente potrebbe pensare di sommare i 3 valori riportati nella prima riga della tabella 3 e dividere la somma per 3 commettendo un grave errore di concetto, poiché l'importanza di ogni valore medio condizionato dipende dal numero delle unità statistiche rispetto al quale è calcolato.*
*L'insegnante chiede poi agli studenti di calcolare il voto medio finale dei 188 studenti che formano il collettivo. Gli studenti, utilizzando i dati di Tabella 2 (fase 1) calcolano:*

$$\bar{x} = \frac{6*63 + 7*53 + 8*42 + 10*6}{188} = 7{,}239$$

Ossia si ha $\bar{x}$ = M($\bar{x}_j$), in altri termini la media della distribuzione del carattere *X* è uguale alla media della distribuzione delle medie condizionate.

***Nota per l'insegnante***

*La relazione $\bar{x}$ = M($\bar{x}_j$) si può dimostrare: partendo dalla formula della media della marginale X e ricordando che: $\sum_{j=1}^{k} n_{ij} = n_{i*}$ si ha:*

$$\bar{x} = \frac{\sum_{i=1}^{h} x_i \cdot n_{i*}}{n} = \frac{\sum_{i=1}^{h} x_i \cdot \sum_{j=1}^{k} n_{ij}}{n} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{h} x_i \cdot n_{ij}}{n}$$

*Dalla formula della media condizionata* $\overline{x_j} = \frac{\sum_{i=1}^{h} x_i \cdot n_{ij}}{n_{\cdot j}}$ *si ottiene:* $\bar{x} = \frac{\sum_{j=1}^{k} n_{\cdot j} \cdot \overline{x_j}}{n}$ *dove il secondo membro rappresenta la media della distribuzione delle medie condizionate.*

La relazione scritta è concettualmente molto importante, perché chiarisce che $\bar{x}$ è la media aritmetica di due diverse distribuzioni: la distribuzione dei voti finali e la distribuzione dei voti medi finali condizionatamente all' (ossia tenuto conto dell') esito di giugno. Ha allora senso confrontare ciascun $\overline{x_j}$ con la media aritmetica della distribuzione a cui appartiene, che è a sua volta uguale al voto medio della distribuzione dei voti finali $\bar{x}$. Quanto più $|\overline{x_j} - \bar{x}|$ è grande, tanto più in media le modalità della distribuzione condizionata *X/Y=yj* sono lontane dalla *media della X* e si può sostenere che *Y* influisce in media su *X*.

Per meglio comprendere queste osservazioni l'insegnante mostra il grafico di figura 2 dove sono riportate le medie delle distribuzioni condizionate *X/Y=yj* e la media della distribuzione marginale *X*. Nel grafico è anche evidenziato lo scostamento in modulo $|\overline{x_j} - \bar{x}|$.

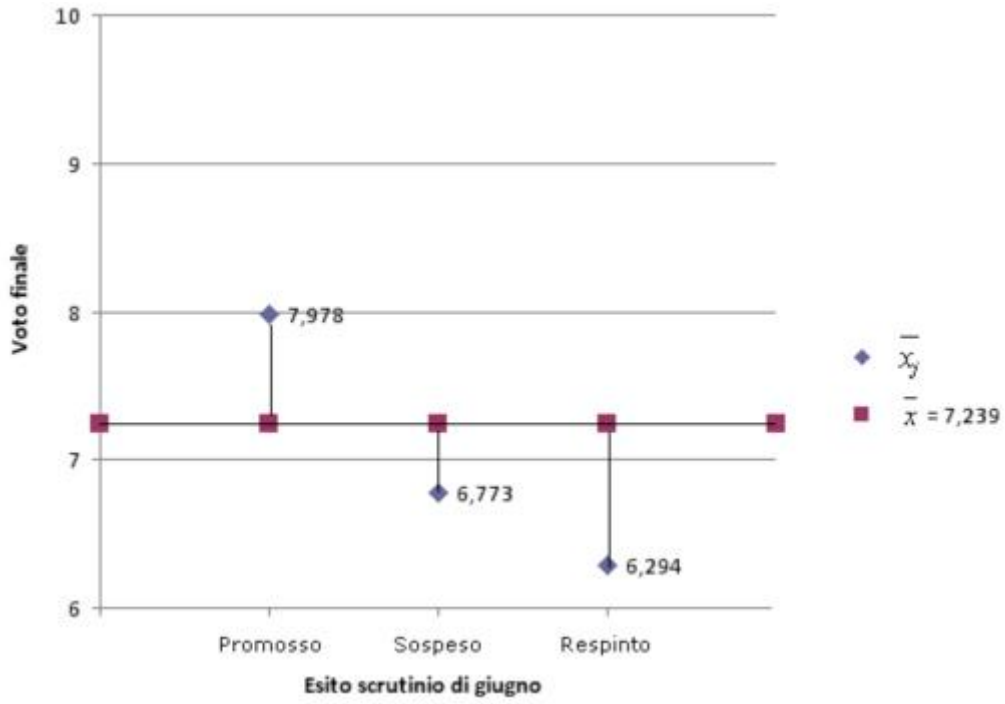



Figura 2

L'insegnante propone agli studenti di misurare la connessione in media nel modo seguente:

$$\frac{(7{,}978-7{,}239)^2\cdot 93+(6{,}773-7{,}239)^2\cdot 44+(6{,}294-7{,}239)^2\cdot 51}{188}=0{,}563$$

e chiede a quale sintesi a loro nota corrisponde la misura di connessione proposta. Gli studenti dovrebbero essere in grado di comprendere che rispetto alla distribuzione di Tabella 4 si tratta della varianza della distribuzione delle medie condizionate, per la quale si usa il simbolo $\sigma^2_{\bar{X}}$.

La varianza $\sigma^2_{\bar{X}}$ è una misura assoluta della connessione in media, infatti varia fra:

il minimo 0, quando $\overline{x_j} = \bar{x}$ per ogni *j* (il che significa che al variare di $y_j$ il valore di $\overline{x_j}$ rimane costante) e il suo massimo che coincide con $\sigma^2_{\bar{X}}$ (la varianza della distribuzione marginale X, qui la varianza della distribuzione dei voti finali). $\sigma^2_{\bar{X}} = \alpha^2_x$ si realizza se e solo se nella distribuzione doppia ad ogni modalità in riga si associa una ed una sola modalità in colonna (il che significa, ad esempio. che tutti i promossi avrebbero voto finale uguale fra loro ma diverso da quelli sia dei sospesi sia dei respinti, e così via).

Per quantificare quanto "è forte" il legame in media di *X* rispetto ad *Y* (cioè quanto "l'Esito dello scrutinio di giugno" sia un indicatore del voto finale) l'insegnante fa calcolare il seguente rapporto tra varianze:

$$\eta^2_{X/Y} = \frac{\sigma^2_{\bar{X}}}{\sigma^2_X}$$

Esso si chiama **rapporto di correlazione**. Per la Tabella 2 (fase 1), l'insegnante fa calcolare agli studenti $\sigma^2_x$, ossia la varianza della distribuzione dei voti, che è uguale a 1,299, maggiore di .0,563. Per il collettivo in esame si ha allora:

$$\eta^2_{X/Y} = \frac{\sigma^2_{\bar{X}}}{\sigma^2_X} = 0{,}563/1{,}299 = 0{,}433$$

Gli studenti osservano che il numeratore è minore del denominatore e quindi il rapporto è un numero adimensionale compreso tra 0 ed 1(estremi inclusi). L'insegnante aiuta nell'interpretazione del valore ottenuto affermando che esso esprime la parte di varianza di *X* dovuta al legame in media con *Y*.

**Nota per l'insegnante**

*La giustificazione che* $\sigma^2_{\bar{X}} \leq \alpha^2_X$ *è fornita dalla scomposizione della varianza di X tenendo conto delle distribuzioni condizionate di Y/X.*

$$\sigma^2_{\bar{X}} = \frac{\sum_{i=1}^{h}\left(\bar{x}_i - \bar{\bar{x}}\right)^2 \cdot n_{i\bullet}}{n} = \frac{\sum_{i=1}^{h}\left(\bar{x}_i - \bar{\bar{x}}\right)^2 \cdot \sum_{j=1}^{k} n_{ij}}{n}$$

*Nelle parentesi tonda si aggiunge e si toglie* $\bar{x}_j$ *e si raggruppa nel modo seguente:*

$$= \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}\left[\left(x_i - \bar{x}_j\right) + \left(\bar{x}_j - \bar{\bar{x}}\right)\right]^2 \cdot n_{ij}}{n}$$

*Sviluppando il quadrato del binomio si ha:*

$$= \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}\left[\left(x_i - \bar{x}_j\right)^2 + 2\left(x_i - \bar{x}_j\right)\cdot\left(\bar{x}_j - \bar{\bar{x}}\right) + \left(\bar{x}_j - \bar{\bar{x}}\right)^2\right] \cdot n_{ij}}{n}$$

*Applicando la proprietà distributiva:*

$$\frac{\sum_{i=1}^{h}\sum_{j=1}^{k}\left(x_i - \bar{x}_j\right)^2 \cdot n_{ij}}{n} + \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}\left(\bar{x}_j - \bar{\bar{x}}\right)^2 \cdot n_{ij}}{n} + \frac{2\sum_{i=1}^{h}\sum_{j=1}^{k}\left(x_i - \bar{x}_j\right)\left(\bar{x}_j - \bar{\bar{x}}\right)\cdot n_{ij}}{n}$$

*Analizzando il doppio prodotto si ottiene:*

$$2\sum_{i=1}^{h}\sum_{j=1}^{k}\left(x_i - \bar{x}_j\right)\left(\bar{x}_j - \bar{\bar{x}}\right)\cdot n_{ij} = 2\sum_{j=1}^{k}\left(\bar{x}_j - \bar{\bar{x}}\right)\cdot\left(\sum_{i=1}^{h}\left(x_i - \bar{x}_j\right)\cdot n_{ij}\right)$$

*La sommatoria tra parentesi rappresenta, fissato j, la somma di tutti gli scarti dei valori dalla propria media che, come noto, è zero.*

*Quindi rimane:*

$$\frac{\sum_{i=1}^{h}\sum_{j=1}^{k}\left(x_i-\bar{x}_j\right)^2\cdot n_{ij}}{n}+\frac{\sum_{j=1}^{k}\left(\bar{x}_j-\bar{x}\right)^2\sum_{i=1}^{h}\cdot n_{ij}}{n}$$

*Osservando che:*

$= \sum_{i=1}^{h}\left(x_i-\bar{x}_j\right)^2\cdot n_{ij}=\sigma^2_{X/y_j}\cdot n_{\bullet j}$ *e che* $\sum_{i=1}^{h} n_{ij}=n_{\bullet j}$ si ottiene:

*Od anche*

$$\sigma^2_X=\sigma^2_{\bar{X}}+\overline{\sigma^2}_{X/y_j}$$

*Ciò giustifica fra l'altro la scrittura:*

$= \eta^2_{X/Y}=\frac{\sigma^2_{\bar{X}}}{\sigma^2_X}=\frac{\sigma^2_{\bar{X}}}{\sigma^2_{\bar{X}}+\overline{\sigma^2}_{X/y_j}}$

*che mostra che il rapporto vale 1 se non c'è variabilità all'interno delle distribuzioni condizionate. Inoltre, poiché il numeratore è parte del denominatore il rapporto di correlazione esprime la parte di varianza di X dovuta al legame in media con Y.*

*Ancora, se la distribuzione doppia fa riferimento a due variabili quantitative è opportuno, nell'analisi della dipendenza in media, individuare quale fra le due variabili "spiega" l'altra (una è conseguenza dell'altra). Ad esempio: consumo di latte ed età (l'età "spiega" il consumo di latte); spesa per viaggi e reddito delle famiglie, (il reddito influisce sulla spesa per viaggi). Va comunque notato che se entrambe le variabili sono quantitative è possibile calcolare la dipendenza in media sia di X rispetto a Y che di Y rispetto a X; i due indici di dipendenza in media sono rispettivamente:*

*e* $\eta^2_{Y|X} = \frac{\sigma^2_{\bar{y}}}{\sigma^2_y}$ *, e i loro valori sono in generale diversi. Per un approfondimento su tutti i temi trattati nell'unità è possibile consultare le prime 6 pagine dell'articolo tratto dalla rivista Induzioni n. 24, 2002 della prof.ssa M. G. Ottaviani messo tra i materiali dell'unità con il titolo:* Distribuzioni statistiche doppie .

**Fase 3**

**Rapporti di correlazione "particolari"**

Per rafforzare i concetti studiati legati al calcolo e all'interpretazione del rapporto di correlazione , l'insegnante propone alla classe di analizzare la dipendenza in media del voto finale rispetto all'esito dello scrutinio nei casi che seguono.

**Caso 1**

Calcolare il rapporto di correlazione in media della Tabella 1

| **Studenti delle classi prime dell'a.s. 2010/11 per Voto finale della scuola secondaria di 1° grado ed esito dello scrutino di giugno 2011 (frequenze assolute)** | | | | |
|---|---|---|---|---|
| **Voto finale $x_i$** | **Esito scrutino di giugno $y_j$** | | | **Totale** |
| | **Promosso** | **Sospeso** | **Respinto** | |
| 6 | | | 51 | 51 |
| 7 | | 44 | | 44 |
| 8 | 93 | | | 93 |
| 9 | | | | 0 |
| 10 | | | | 0 |
| **Totale** | 93 | 44 | 51 | 188 |

Tabella 1

Gli studenti ricavano la distribuzione delle medie condizionate sotto riportata e di questa trovano la media e varianza e calcolano la varianza della marginale di colonna

| | **Promosso** | **Sospeso** | **Respinto** |
|---|---|---|---|
| $\overline{x_j}$ | 8 | 7 | 6 |
| **n.**$_j$ | 93 | 44 | 51 |

$\bar{x} = 7{,}2234 \quad \sigma_x^2 = 0{,}71605 \quad \sigma_{\bar{x}}^2 = 0{,}71605$

Quindi il rapporto di correlazione risulta essere: $\eta_{X/Y}^2 = \frac{\sigma_{\bar{X}}^2}{\sigma_X^2} = 1$

Gli studenti osservano che le distribuzioni condizionate hanno variabilità nulla e quindi la Tabella 1 presenta un caso di massima connessione in media di X rispetto ad Y.

**Caso 2**

Calcolare il rapporto di correlazione in media della Tabella 2.

| **Studenti delle classi prime dell'a.s. 2010/11 per Voto finale della scuola secondaria di 1° grado ed esito dello scrutino di giugno 2011 (frequenze assolute)** | | | | |
|---|---|---|---|---|
| **Voto finale $x_i$** | **Esito scrutino di giugno $y_j$** | | | **Totale** |
| | **Promosso** | **Sospeso** | **Respinto** | |
| 6 | 46 | 22 | | 68 |
| 7 | | | 51 | 51 |
| 8 | 46 | 22 | | 68 |
| 9 | | | | 0 |
| 10 | | | | 0 |
| **Totale** | 92 | 44 | 51 | 187 |

Tabella 2

Anche per la Tabella 2 gli studenti ricavano la distribuzione delle medie condizionate sotto riportata e di questa trovano media e varianza e calcolano inoltre la varianza della marginale di colonna

| | **Promosso** | **Sospeso** | **Respinto** |
|---|---|---|---|
| $\overline{x_j}$ | 7 | 7 | 7 |
| **n.**$_j$ | 93 | 44 | 51 |

$\bar{x} = 7 \qquad \sigma_x^2 = 0 \qquad \sigma_x^2 = 0{,}727273$

Quindi il rapporto di correlazione risulta essere: $\eta^2_{X|Y} = \frac{\sigma^2_{\bar{X}}}{\sigma^2_X} = 0.$

Gli studenti attenti osservano che l'aver trovato medie condizionate uguali implica immediatamente che il voto finale X non dipende in media dall'esito dello scrutinio di giugno.

**Spunti per altre attività con gli studenti**

Si propone l'attività:

**Costruzione di una variabile casuale doppia**

Fase 1

L'insegnante propone la seguente attività:

si lancino tre monete uguali e si indichi con X il numero delle teste ottenute e con Y il numero di variazioni nella sequenza ottenuta. Chiede di costruire la variabile doppia (X, Y) e di analizzare se esiste una relazione tra le due variabili osservate.
Gli studenti forniscono lo spazio fondamentale Ω e lo riportano in Tabella 1 assieme alla probabilità associata ad ogni esito elementare.

| **Eventi elementari** | | | **Probabilità dell'evento** |
|---|---|---|---|
| C | C | C | 1/8 |
| C | C | T | 1/8 |
| C | T | C | 1/8 |
| T | C | C | 1/8 |
| T | T | C | 1/8 |
| T | C | T | 1/8 |
| C | T | T | 1/8 |
| T | T | T | 1/8 |

Tabella 1

Gli studenti evidenziano che in questo esperimento casuale i valori assunti dalla variabile casuale *X* sono: 0, 1, 2,3 e quelli assunti dalla variabile casuale *Y* sono: 0, 1, 2. Per ogni evento elementare, l'insegnante fa individuare i valori delle coppie (*X, Y*) e li fa riportare in Tabella 2

| **Eventi elementari** | | | **X** | **Y** |
|---|---|---|---|---|
| C | C | C | 0 | 0 |
| C | C | T | 1 | 1 |
| C | T | C | 1 | 2 |
| T | C | C | 1 | 1 |
| T | T | C | 2 | 1 |
| T | C | T | 2 | 2 |
| C | T | T | 2 | 1 |
| T | T | T | 3 | 0 |

Tabella 2

Per rispondere al quesito posto invita la classe a costruire la distruzione di probabilità della variabile casuale doppia (*X, Y*) riportando i valori delle probabilità congiunte in una tabella e costruendo anche le distribuzioni marginali di probabilità.

| ***X/Y*** | **0** | **1** | **2** | **$P_1(x)$** |
|---|---|---|---|---|
| 0 | 1/8 | 0 | 0 | 1/8 |
| 1 | 0 | 2/8 | 1/8 | 3/8 |
| 2 | 0 | 2/8 | 1/8 | 3/8 |
| 3 | 1/8 | 0 | 0 | 1/8 |
| $P_2(y)$ | 2/8 | 4/8 | 2/8 | 1 |

Tabella 3

**Fase 2**

Gli studenti osservano che, ad esempio p($X$=0, $Y$= 1) =0 ma $P_1(X=0) \cdot P_2(Y=1)$ è diverso da zero. Ciò permette di affermare che X e Y sono **dipendenti in legge**.

L'insegnante invita la classe a quantificare il livello di dipendenza in legge usando le conoscenze acquisite.

Dai calcoli dell'indice chi- quadro di Pearson la classe ottiene un valore pari a 1 che normalizzato diventa ½. L'insegnante fa osservare che la dipendenza in legge tra $X$ e $Y$ è il valore centrale tra i limiti dell'indice normalizzato.

L'insegnante chiede se la dipendenza in media è pure significativa e quale delle due variabili spiega l'altra?

Dalla discussione emerge che è indifferente scegliere la variabile che spiega la relazione. L'insegnante, dopo aver diviso la classe in due gruppi, fa calcolare ad un gruppo le medie condizionate di $Y/X$ e all'altro le media condizionate di $X/Y$.

Il primo gruppo presenta la seguente tabella che riporta, per ogni $x_i$, la corrispondente media condizionata e la probabilità associata osservando che c'è dipendenza in media di $Y$ rispetto ad $X$.

| ***x*** | $\bar{y}_i$ | **$P_1(x)$** |
|---|---|---|
| 0 | 0,00 | 1/8 |
| 1 | 1,33 | 3/8 |
| 2 | 1,33 | 3/8 |
| 3 | 0,00 | 1/8 |

Il secondo gruppo presenta la successiva tabelle che riporta, per ogni $y_j$, la corrispondente media condizionata e la probabilità associata osservando che,

essendo i valori tutti uguali ed uguali alla media di *X*, *X* non dipende in media da *Y*.

| **y** | $\overline{x_j}$ | **$P_2(y)$** |
|---|---|---|
| 0 | 1,5 | 1/4 |
| 1 | 1,5 | 1/2 |
| 2 | 1,5 | 1/4 |

L'insegnante, da questi risultati fa osservare che la dipendenza in media non è una relazione simmetrica come invece è l'indipendenza in distribuzione.

## Elementi per prove di verifica

### Esercizio 1

Enrico ha rilevato per ogni giorno e nell'arco di un intero anno, le condizioni meteorologiche al momento della sua prima uscita di casa e il mezzo di trasporto che ha scelto di usare. Alla fine dell'anno ha raccolto le osservazioni nella tabella qui sotto.

| **Distribuzione del mezzo di trasporto utilizzato da Enrico nell'arco di un anno a seconda delle condizioni meteo al momento della prima uscita giornaliera** | | | | |
|---|---|---|---|---|
| **Condizioni meteo *X*** | **Mezzo di trasporto *Y*** | | | |
| | **Bicicletta** | **Autobus** | **Automobile** | |
| **Sereno** | 84 | 26 | 11 | |
| **Variabile** | 29 | 98 | 29 | |
| **Pioggia** | 7 | 26 | 55 | |
| | | | | |

Esaminata la tabella si chiede.

a) Sono stati di più i giorni di pioggia o quelli di tempo variabile?

b) Alla fine qual è stato il mezzo che Enrico ha usato di più?.

c) Al variare delle condizioni meteorologiche Enrico ha mantenuto la sua scelta proporzionalmente invariata?

d) Nell'insieme esiste o non esiste relazione fra condizioni meteo e scelta del mezzo di trasporto che Enrico ha utilizzato?

**Esercizio 2**

Data la seguente tabella che riporta l'esito all'esame di stato e il voto all'esame di statistica di 123 studenti della Facoltà di Economia di Ca' Foscari di Venezia nell'a.a. 2004/05.

| **Esito esame di stato *X*** | **Voto all'esame di statistica *Y*** | | | **Totale** |
|---|---|---|---|---|
| | **18 -\| 23** | **23 -\| 27** | **27 -\| 30** | |
| **60 -\| 73** | 12 | 9 | 0 | 21 |
| **73 -\| 87** | 6 | 16 | 10 | 32 |
| **87 -\| 100** | 18 | 33 | 19 | 70 |
| **Totale** | 36 | 58 | 29 | 123 |

a) L'esito all'esame di stato ha influito sull'esito dell'esame di statistica? Da cosa lo hai dedotto? Fornisci una breve motivazione.

b) In condizione di indipendenza fra X e Y, quanti studenti con esito all'esame di stato nella classe 73 -|87 avrebbero ottenuto un voto all'esame di statistica nella classe 27 -| 30?

c) Valuta il grado di connessione fra le variabili X e Y attraverso l'indice di Chi-quadro normalizzato di Pearson.

**Esercizio 3**

La seguente tabella riporta la distribuzione di un gruppo di automobilisti che hanno avuto un sinistro (incidente) nel corso di una settimana in una media città di provincia rispetto all'età media *X* degli assicurati e alla entità del sinistro *Y*

| **Età media dell'assicurato** | **Entità del sinistro** | | |
|---|---|---|---|
| | **Lievi** | **Medi** | **Gravi** |
| 25 | 350 | 10 | 15 |
| 45 | 200 | 180 | 5 |
| 72 | 150 | 10 | 80 |

a) rappresenta graficamente la distribuzione data scegliendo il grafico più opportuno per descrivere la connessione fra i caratteri osservati;

b) calcola:

b1) la percentuale degli automobilisti di età media 45 che hanno avuto sinistri lievi;

b2) la percentuale degli automobilisti con sinistri lievi;

b3) sapendo che l'automobilista ha 45 anni calcola la percentuale di sinistri lievi;

c) costruisci la distribuzione, espressa in termini relativi, dei sinistri degli automobilisti di età media 25;

d) verifica se all'aumentare della gravità del sinistro si riscontra anche un aumento dell'età media dell'automobilista.

**Esercizio 4**

É data la seguente distribuzione doppia (*X, Y*) osservata su un campione di n = 90 unità statistiche.

| ***X*** | ***Y*** | |
|---|---|---|
| | **1** | **2** |
| 1 | c | 2c |
| 2 | 2c | 4c |
| 3 | 3c | 6c |

a) Calcola la costante c

b) Scrivi la tabella teorica di indipendenza e commenta il risultato

c) X dipende in media da Y?

**Esercizio 5**

Su un totale di 150 medici che svolgono assistenza mutualistica in una piccola città di provincia, 50 sono le femmine e 30 sono pediatri. Delle 50 femmine, 8 sono pediatri.

L'essere contemporaneamente "femmina" e "pediatra" è un evento indipendente in quella piccola città? Spiega la conclusione alla quale sei pervenuto rispondendo alla domanda precedente.

**Esercizio 6**

In un canale televisivo, in un certo giorno, si è pubblicizzato un nuovo profumo. Il giorno seguente, è stato fatto un sondaggio per verificare alcuni risultati ottenuti con la pubblicità del nuovo profumo. Dopo l'analisi delle risposte, si è concluso che:

- Il 75% degli individui intervistati hanno visto la pubblicità;
- Il 45% degli individui intervistati hanno comperato il nuovo profumo;
- Il 20% degli individui intervistati non hanno visto la pubblicità, né hanno
- comprato il nuovo profumo[1].

a) Con i dati in tuo possesso, scelto a caso un intervistato, completa la tabella doppia di probabilità sotto riportata:

| **Pubblicità vista** | **Acquisto del nuovo profumo** | | |
|---|---|---|---|
| | Si | No | |
| **Si** | | | |
| **No** | | | |
| | | | |

b) Si può ritenere che l'aver visto la pubblicità influisca sull'acquisto del nuovo profumo?

c) Quantifica la connessione fra i due caratteri nel collettivo studiato attraverso l'indice chi-quadro di Pearson.

[1] Testo tratto dall'articolo di G. Baruzzo e P. Ranzani INDUZIONI n.43.

## Soluzioni

### Esercizio 1

a) È sufficiente calcolare $n_{i\bullet} = \sum_{j=1}^{k} n_{ij}$ per ciascuna modalità del carattere *X* e si deduce che sono stati di più i giorni con tempo variabile.

b) È sufficiente calcolare $n_{\bullet j} = \sum_{i=1}^{h} n_{ij}$ per ciascuna modalità del carattere *Y* e si osserva che è stato usato di più l'autobus.

| Condizioni meteo *X* | Mezzo di trasporto *Y* | | | **ni.** |
|---|---|---|---|---|
| | Bicicletta | Autobus | Automobile | |
| Sereno | 84 | 26 | 11 | **121** |
| Variabile | 29 | 98 | 29 | **156** |
| Pioggia | 7 | 26 | 55 | **88** |
| **n.j** | **120** | **150** | **95** | **365** |

c) Si devono calcolare le distribuzioni condizionate $Y/X=x_i$ e commentare il risultato. Si ricorda che $f(y_j/x_i)= \frac{n_{ij}}{n_{i\bullet}}$

| Y | Bicicletta | Autobus | Automobile | Totale |
|---|---|---|---|---|
| f(*Y*/*X*=Sereno) | 0,694 | 0,215 | 0,091 | 1,000 |
| f(*Y*/*X*=Variabile) | 0,186 | 0,628 | 0,186 | 1,000 |
| f(*Y*/*X*=Pioggia) | 0,080 | 0,295 | 0,625 | 1,000 |

Dalla tabella delle distribuzioni condizionate si osserva che Enrico preferisce andare in bicicletta solo se il tempo è sereno.

d) La riposta è immediata basta osservare che le distribuzioni condizionate del punto. c) sono diverse al variare delle modalità di *X*. Si ricorda che due variabili sono indipendenti se e solo se $f(y_j/x_i)= \frac{n_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{n}$ per ogni (*i*, *j*).

**Esercizio 2**

a) La risposta può essere immediata osservando che nessuno tra quelli che hanno avuto esito tra 60 e 73 all'esame di stato ha meritato un voto tra 27 e 30 all'esame di statistica mentre tutti gli altri con esito superiore a 73 si sono distribuiti in tutte le classi di voto all'esame di statistica.

La tabella delle distribuzioni condizionate di $Y/X=x_i$ è la seguente:

| *Y* | 18 -\| 23 | 23 -\| 27 | 27 -\| 30 | Totale |
|---|---|---|---|---|
| f(*Y*/*X* = 60 -\| 73) | 0,571 | 0,429 | 0,000 | 1,000 |
| f(*Y*/*X*=73 -\| 87) | 0,188 | 0,500 | 0,313 | 1,000 |
| f(*Y*/*X*=87 -\| 100) | 0,257 | 0,471 | 0,271 | 1,000 |

Da cui emerge che al variare delle modalità di *X* le distribuzioni condizionate cambiano.

c) Se *X* ed *Y* fossero indipendenti in distribuzione le frequenze congiunte sarebbero: $n_{ij}^* = \frac{n_{\bullet j} \times n_{i\bullet}}{n}$ per ogni coppia. Nello specifico si chiede di calcolare $n_{23}^* = \frac{29 \times 32}{123} = 7{,}545$ studenti.

c) Si devono costruire:

- la tabella teorica di indipendenza $n^*_{i,j} = \frac{n_{\bullet j} \times n_{i\bullet}}{n}$

| Esito esame di stato *X* | Voto all'esame di statistica *Y* | | | Totale |
|---|---|---|---|---|
| | 18 -\| 23 | 23 -\| 27 | 27 -\| 30 | |
| 60 -\| 73 | 6,146 | 9,902 | 4,951 | 21 |
| 73 -\| 87 | 9,366 | 15,089 | 7,545 | 32 |
| 87 -\| 100 | 20,488 | 33,008 | 16,504 | 70 |
| Totale | 36 | 58 | 29 | 123 |

- la tabella delle contingenze $c_{i,j} = (n_{i\,J} - n^*_{i,j})$

| Esito esame di stato *X* | Voto all'esame di statistica *Y* | | |
|---|---|---|---|
| | 18 -\| 23 | 23 -\| 27 | 27 -\| 30 |
| 60 -\| 73 | 5,854 | -0,902 | -4,951 |
| 73 -\| 87 | -3,366 | 0,911 | 2,455 |
| 87 -\| 100 | -2,488 | -0,008 | 2,496 |

- calcolo dell'indice di Pearson: $\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{c^2_{i,j}}{n^*_{i,j}}$

In tabella sono riportati i rapporti: $\frac{c^2_{i,j}}{n^*_{i,j}}$

| Esito esame di stato *X* | Voto all'esame di statistica *Y* | | | |
|---|---|---|---|---|
| | 18 -\| 23 | 23 -\| 27 | 27 -\| 30 | Totali |
| 60 -\| 73 | 5,575 | 0,082 | 4,951 | 10,608 |
| 73 -\| 87 | 1,210 | 0,055 | 0,799 | 2,064 |
| 87 -\| 100 | 0,302 | 0,000 | 0,377 | 0,680 |
| Totali | 7,087 | 0,137 | 6,128 | **13,352** |

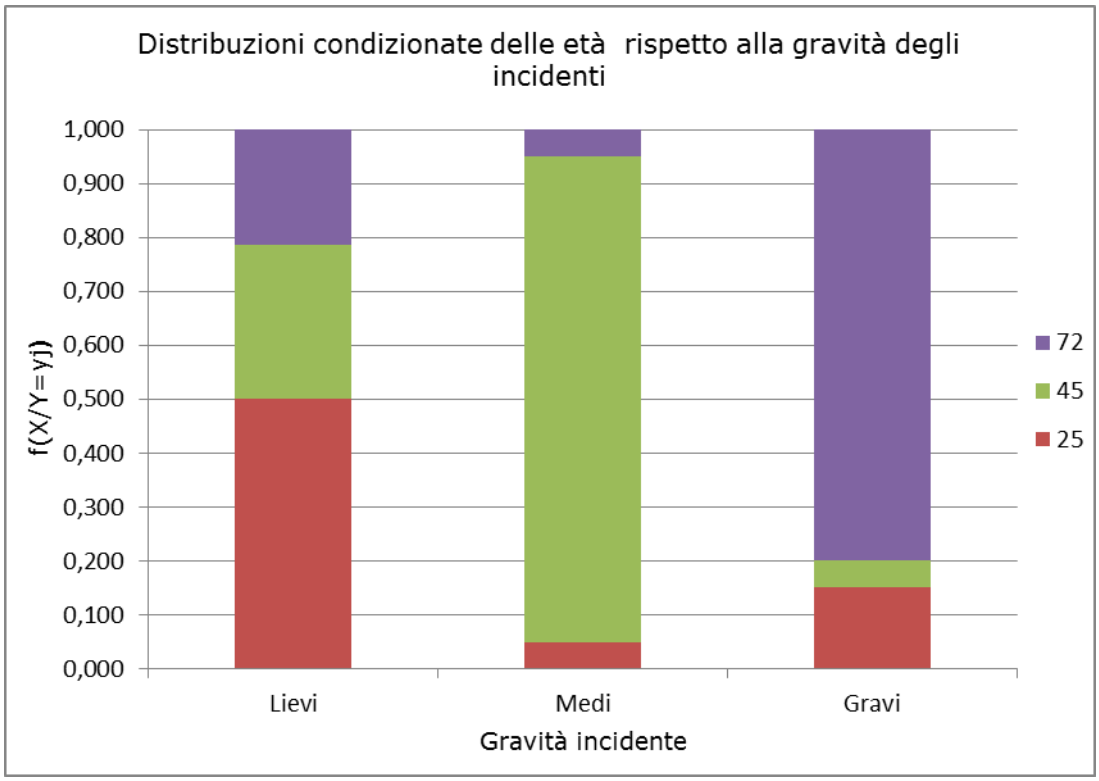
$\chi^2 = 13{,}352$

- per normalizzare occorre calcolare il valore di Chi-quadro in caso di massima dipendenza: minimo tra $n*(h-1)$; $n*(k-1)$ dove $h$ = 3 (numero di righe) e $k$ = 3 (numero di colonne). Il valore è uguale a 246 = 123*2.

  $\frac{\chi^2}{\chi^2_{Max}}$ = 0,0543 che evidenzia una bassa relazione tra *X* e *Y*. Ciò significa che vi è poca influenza tra voto all'esame di statistica ed esito all'esame di stato.

**Esercizio 3**

a) Il grafico idoneo è:



In tale grafico è immediato cogliere la diversa distribuzione delle età rispetto alla gravità dell'incidente. La metà degli incidenti lievi è causata da automobnilisti di età media pari a 25 anni.

b1) Si chiede: $\frac{n_{21}}{n} * 100 = \frac{200}{1142} * 100 = 17{,}51\%$,

b2) Si chiede: $\frac{n_{\bullet 1}}{n} * 100 = \frac{700}{1142} * 100 = 61{,}30\%$

b3) Si chiede: $\frac{n_{21}}{n_{2\bullet}} * 100 = \frac{200}{385} * 100 = 51{,}95\%$

c) La tabella che segue riporta la distribuzione cercata.

| *Y* | Lievi | Medi | Gravi | Totale |
|---|---|---|---|---|
| f(*Y*/*X* = 25) | 0,875 | 0,025 | 0,0375 | 1 |

d) La tabella di sintesi con le diverse medie condizionate alle entità dei sinistri è la seguente:

| *Y* | |
|---|---|
| Lievi | 40,78 |
| Medi | 45,35 |
| Gravi | 63,60 |

La tabella sopra evidenzia un aumento dell'età media quindi la risposta alla domanda è affermativa.

Si riporta, come esempio, il calcolo della media condizionata $\bar{x}_{lievi}$:

| Età (*X*) | f(*X*/*Y* = Lievi) | Xi * f(*xi*/*Y*=Lievi) |
|---|---|---|
| 25 | 0,500 | 12,5 |
| 45 | 0,286 | 12,87 |
| 72 | 0,214 | 15,408 |
| Totale | 1,000 | 40,78 |

| | |
|---|---|
| $\overline{x}_{lievi}$ | **40,778** |

**Esercizio 4**

a) Ricordando che $\sum_{i=1}^{3}\sum_{j=1}^{2} n_{ij} = 90$ si ha: $\sum_{i=1}^{3}\sum_{j=1}^{2} n_{ij}$ =18*c = 90. Da cui c = 5.

Pertanto la tabella iniziale diventa:

| X | Y | | |
|---|---|---|---|
| | 1 | 2 | Totale |
| 1 | 5 | 10 | 15 |
| 2 | 10 | 20 | 30 |
| 3 | 15 | 30 | 45 |
| Totale | 30 | 60 | 90 |

b) La tabella teorica di indipendenza $n_{ij}^{*} = \frac{n_{\bullet j} \times n_{i\bullet}}{n}$ è:

| Tabella di indipendenza $n_{i,j}^{*}$ | | | |
|---|---|---|---|
| X | Y | | ni. |
| | 1 | 2 | |
| 1 | 5 | 10 | 15 |
| 2 | 10 | 20 | 30 |
| 3 | 15 | 30 | 45 |
| n.j | 30 | 60 | 90 |

Dalla tabella si evidenzia che è uguale alla tabella data cioè vale $n_{i,j}^{*} = n_{i,j}$

per ogni coppia. Pertanto X e Y sono indipendenti in distribuzione.

d) Essendo le variabili indipendenti in distribuzione *X* **non** dipende in media da *Y*.

**Esercizio 5**

Per rispondere alla domanda posta si costruisce la seguente tabella doppia di frequenza:

| Distribuzione di 150 medici di una piccola città di provincia per genere e tipo di assistenza mutualistica | | | |
|---|---|---|---|
| genere | tipo di assistenza | | totale |
| | pediatra | altro | |
| Maschio | **30-8=22** | **100-22=78** | **150-50=100** |
| femmina | 8 | **50-8=42** | 50 |
| totale | 30 | **150-30=120** | 150 |

e si calcola la frequenza teorica di indipendenza

$n_{21}^{*} = \frac{n_{2\bullet} \cdot n_{\bullet 1}}{n} = \frac{50 \cdot 30}{150} = 10 \neq n_{21} = 8$ da cui si afferma che l'essere femmina e pediatra in quella città non sono eventi indipendenti.

## Esercizio 6

a)

| Pubblicità vista | Acquisto del nuovo profumo | | $P_1(x_i)$ |
|---|---|---|---|
| | Si | No | |
| Si | 0.75-0.35=**0.4** | 0.55-0.2=**0.35** | 0.75 |
| No | 0.25-0.2=**0.05** | 0.2 | 1-0.75=**0.25** |
| $P_2(y_j)$ | 0.45 | 1-0.45=**0.55** | 1 |

b) E' sufficiente verificare se per qualche coppia $(x,y)$ risulta: $P(x,y) \neq P_1(x)*P_2(y)$; ad esempio per la coppia ($X$=sì, $Y$=sì) si ha: $P_1(X=\text{sì})*P_2(Y=\text{sì}) = 0.75*0.45 = 0.3375$ che è diverso da 0,4. Si può quindi affermare che l'aver visto la pubblicità influisce sull'acquisto.

c) Per rispondere si devono calcolare le probabilità teoriche in caso di non connessione date da $P^*(x_i, y_j) = P_1(x_i) \cdot P_2(y_j)$

| | tabella teorica di non connessione | | |
|---|---|---|---|
| Pubblicità vista | Acquisto del nuovo profumo | | $P_1(x_i)$ |
| | Sì | No | |
| Sì | 0,3375 | 0,4125 | 0,75 |
| No | 0,1125 | 0,1375 | 0,25 |
| $P_2(y_j)$ | 0,45 | 0,55 | 1 |

Le contingenze $c_{i,j} = P(x_i,y_j) - P^*(x_i,y_j)$ e di seguito le quantità sotto riportate per il calcolo dell'indice chi-quadro di connessione di Pearson.

| | tabella di $cij^2/p^*_{ij}$ | | |
|---|---|---|---|
| Pubblicità vista | Acquisto del nuovo profumo | | totale |
| | Sì | No | |
| Sì | 0,012 | 0,009 | 0,021 |
| No | 0,035 | 0,100 | 0,135 |
| totale | 0,046 | 0,110 | 0,156 |

Si perviene al valore $\chi^2=\sum_{i=1}^{n}\frac{c_{i,j}^2}{p_{i,j}^*}=0.156$ Essendo tale indice compreso fra 0 ed 1 (massimo valore assunto in questo caso) si ritiene che l'influenza della pubblicità sull'acquisto del nuovo profumo
sia bassa.

## Risorse

### Documentazione e materiali

Tabella 1

Tabella 2

Tabella 3

Tabella 4

Tabella 5

Tabella 6

Tabella 7

Tabella 8

Tabella 9

Tabella 10

### Approfondimenti disciplinari

Tabella 1

Tabella 2

Tabella 3

Tabella 4

Tabella 1 – caso 1

Tabella 2 – caso 1

**Altre attività con gli studenti**

Tabella 1

Tabella 2

Tabella 3

Tabella 4

**Bibliografia**

Leti, G., Cerbara, L. *Elementi di statistica descrittiva*. Collana “manuali”, Il Mulino, Bologna 2009.

Ferrari, P., Nicolini, G., Tommasi, C. *Introduzione all'inferenza statistica*, G. Giappichelli Editore, Torino 2006.

Gnedenko, B. *Teoria della probabilità*, Editori Riuniti *Univ. Press,* Roma 2011.

Baldi, P. *Calcolo delle probabilità e statistica*. McGraw-Hill, Milano 2003.

Cicchitelli, G. *Probabilità e Statistica*. Maggioli editore ed.2, 2004.

Boggio, A., Borello, G. *Statistica*. Vol.1 e 2, Petrini editore, 2008.

Matematica 2003, *La matematica del cittadino* (A proposito di valutazione scolastica).

Quinn, R.J., Wiest, L.R. *Una linea costruttiva nell'insegnare permutazioni e combinazioni*. In: *Induzioni,* 41, 2010, pp.137-146.

**Sitografia**

Distribuzioni di frequenza

(Visitato nel luglio 2013)

Distribuzioni doppie di frequenze (tabelle a doppia entrata)

(Visitato nel luglio 2013)

Statistica descrittiva bivariata

(Visitato nel luglio 2013)

*Questo prodotto multimediale è stato realizzato nel 2013 da INDIRE con i fondi stanziati dal MIUR – Uff. VI nell'ambito del progetto m@t.abel – Apprendimenti di Base.* 