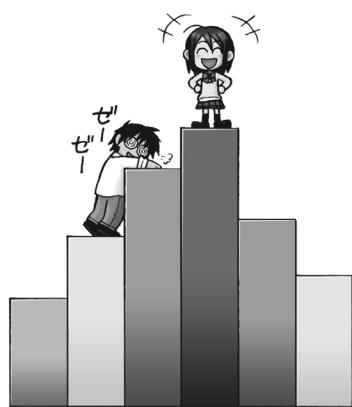


I MANGA DELLE SCIENZE

STATISTICA

SHIN TAKAHASHI
TREND-PRO, CO., LTD.





I MANGA DELLE SCIENZE

STATISTICA

SHIN TAKAHASHI
TREND-PRO CO., LTD.



SOMMARIO

PREFAZIONE	VII
PROLOGO: QUANDO LA STATISTICA ♥ FA BATTERE IL CUORE ♥	1
1	
I TIPI DI DATI	13
1. Dati categorici e dati numerici	14
2. Un esempio di dati categorici complicati	20
3. Come vengono gestite le risposte multiple	28
Esercizio con soluzione	29
Riassumendo	29
2	
UNO SGUARDO D'INSIEME: COME CAPIRE I DATI NUMERICI	31
1. Tabelle di distribuzione di frequenza e istogrammi	32
2. La media	40
3. La mediana	44
4. Deviazione standard	48
5. L'ampiezza di classe di una tabella delle frequenze	54
6. Teoria della stima e statistica descrittiva	57
Esercizio con soluzione	57
Riassumendo	58
3	
UNO SGUARDO D'INSIEME: COME CAPIRE I DATI CATEGORICI	59
1. Tabelle di contingenza	60
Esercizio con soluzione	64
Riassumendo	64
4	
VALORI STANDARD E DI DEVIAZIONE	65
1. Normalizzazione e valore standard	66
2. Caratteristiche dei dati normalizzati	73
3. Valore di deviazione	74
4. Interpretazione del valore di deviazione	76
Esercizio con soluzione	78
Riassumendo	80
5	
TROVIAMO LE PROBABILITÀ	81
1. Funzione di densità di probabilità	82
2. Distribuzione normale	86
3. Distribuzione normale standard	89

Esempio I	95
Esempio II	97
4. Distribuzione chi-quadro	99
5. Distribuzione t	106
6. Distribuzione F	106
7. Distribuzioni e fogli elettronici	107
Esercizio con soluzione	108
Riassumendo	109
6	
CHE RELAZIONE C'È TRA DUE VARIABILI?	111
1. Coefficiente di correlazione	116
2. Rapporto di correlazione	121
3. Il Coefficiente di Cramer	127
Esercizio con soluzione	138
Riassumendo	142
7	
APPROFONDIAMO UN PO' I TEST D'IPOTESI STATISTICA	143
1. Test d'ipotesi statistica	144
2. Il test d'indipendenza chi-quadro	151
Spiegazione	152
Esercizio	157
Pensiamoci un po' su	158
Soluzione	160
3. L'ipotesi nulla e quella alternativa	170
4. P-value e procedure per i test d'ipotesi	175
5. Test d'indipendenza e test di omogeneità	184
Esempio	184
Procedimento	185
6. Conclusioni	187
Esercizio con soluzione	188
Riassumendo	189
APPENDICE	
FACCIAMO UN PO' DI CALCOLI CON UN FOGLIO ELETTRONICO	191
1. Compilare una tabella delle frequenze	192
2. Calcolare media aritmetica, mediana e deviazione standard	195
3. Compilare una tabella di contingenza	197
4. Calcolare il valore standard e quello di deviazione	199
5. Calcolare le probabilità della distribuzione normale standard	204
6. Calcolare il valore dell'ascissa della distribuzione chi-quadro	205
7. Calcolare il coefficiente di correlazione	206
8. Test d'indipendenza	208
INDICE	213

PREFAZIONE

Questo volume è un'introduzione alla statistica che si rivolge a chi:

- ha necessità di condurre analisi dei dati a scopi di ricerca o per lavoro;
- non ha necessità lavorative immediate, ma è interessato ad avere un'idea di cosa siano l'analisi dei dati e il mondo della statistica;
- ha già una certa conoscenza generale dell'argomento e desidera approfondirla.

La statistica è una delle aree della matematica più strettamente legate alla vita e al lavoro di tutti i giorni. Averne una qualche famigliarità può tornare utile in situazioni come:

- stimare quante porzioni di spaghetti fritti riuscirete a vendere nello stand che avete intenzione di allestire per la festa del liceo;
- valutare le probabilità di passare uno scritto;
- confrontare le probabilità di guarigione di una persona malata nel caso in cui assuma una certa medicina e nel caso in cui non l'assuma.

Il libro è composto da sette capitoli e ognuno è più o meno organizzato in queste sezioni:

- fumetto;
- schede di spiegazione degli argomenti esposti nel fumetto;
- esercizi e soluzioni;
- riepilogo.

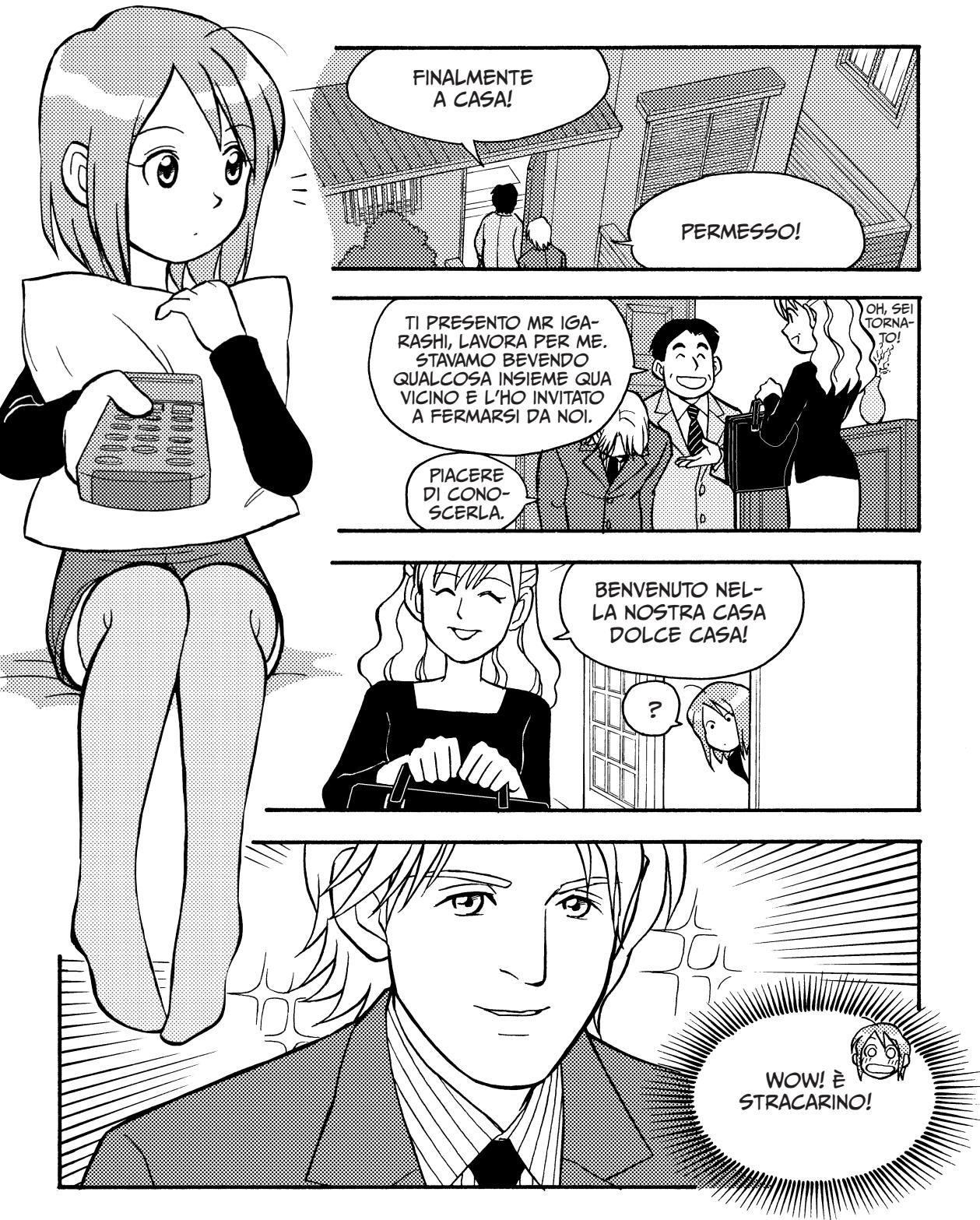
Anche solo col fumetto imparerete un bel po' di cose, ma leggendo anche le altre sezioni approfondirete molto la vostra conoscenza della materia.

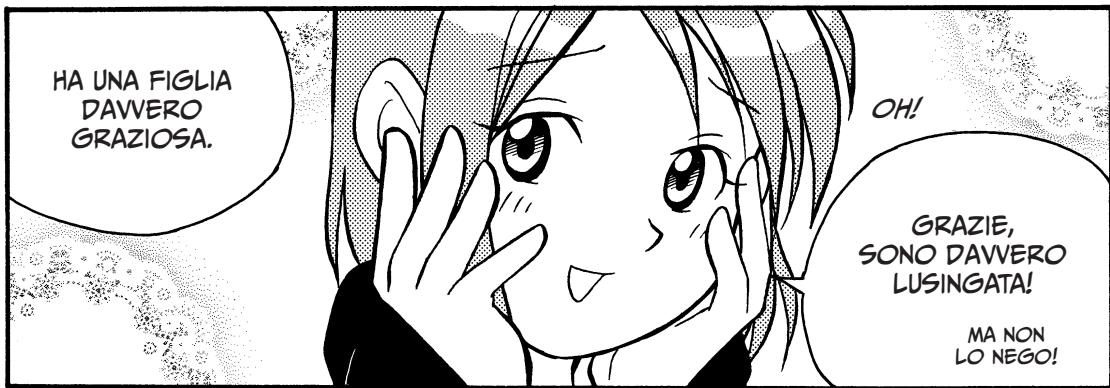
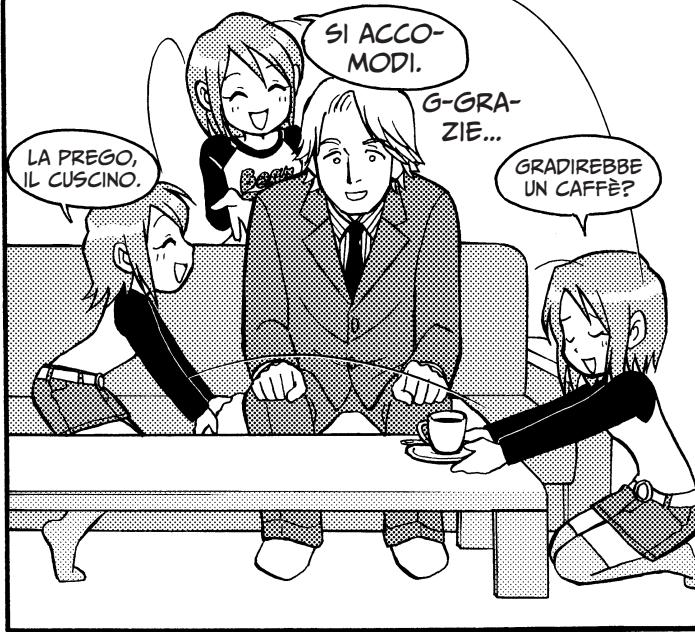
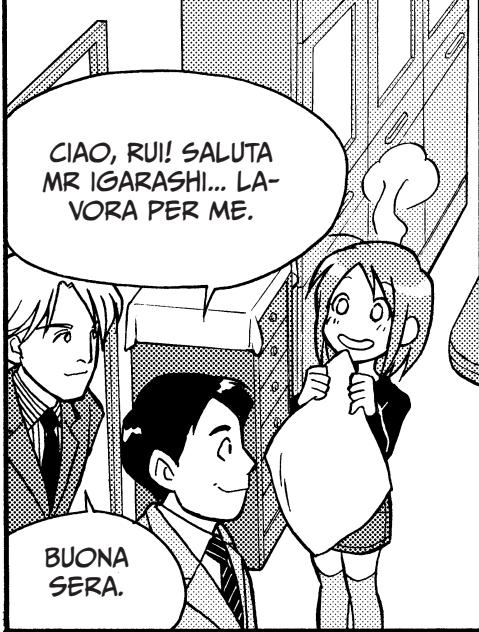
Sarei felice se questo libro vi facesse capire quanto la statistica possa essere utile e divertente.

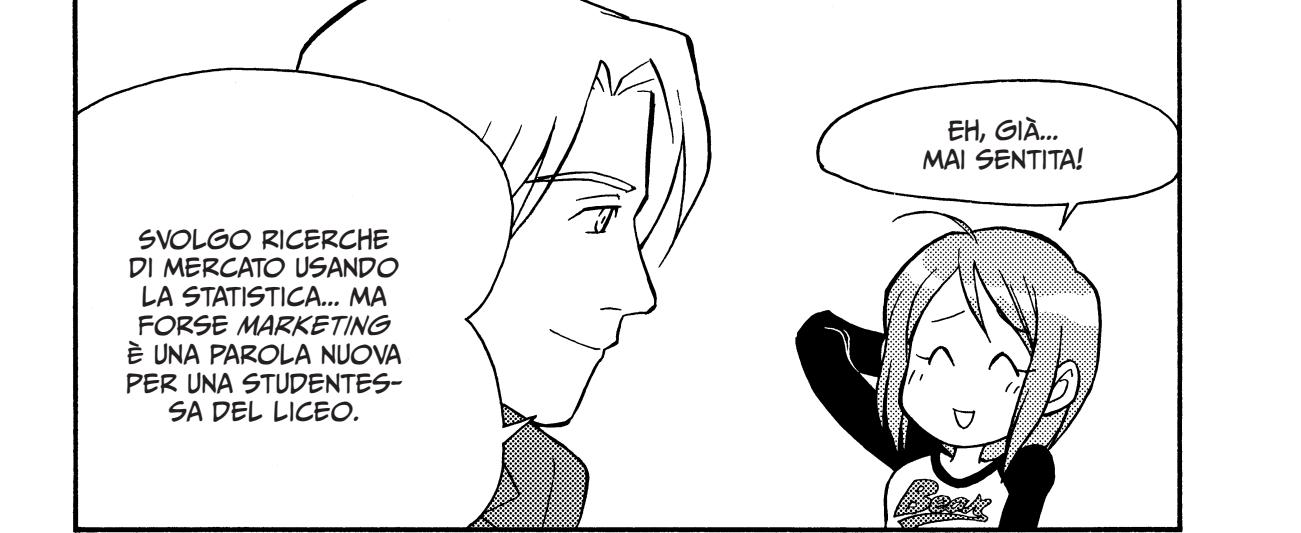
Vorrei ringraziare lo staff di Ohmsha, Ltd., che mi ha offerto la possibilità di scrivere il libro, e TREND PRO, Co., Ltd., per avere adattato il mio testo a fumetti, insieme allo sceneggiatore re_akino e al disegnatore Iroha Inoue. Infine, desidero ringraziare il Dottor Sakaori Fumitake dell'università della Rikkyo University, per i suoi preziosissimi consigli nel corso della stesura del volume.

SHIN TAKAHASHI

PROLOGO:
QUANDO LA STATISTICA
♥ FA BATTERE IL CUORE ♥







SVOLGO RICERCHE DI MERCATO USANDO LA STATISTICA... MA FORSE MARKETING È UNA PAROLA NUOVA PER UNA STUDENTESSA DEL LICEO.

EH, GIÀ...
MAI SENTITA!



SII SINCERA! SAI CHE COS'È LA STATISTICA?

UUUHHH...



DIREI CHE TI SUONA SCONSCIUTA ANCHE QUESTA. IN PAROLE POVERE, LA STATISTICA CERCA DI VALUTARE LE CARATTERISTICHE DI UNA POPOLAZIONE UTILIZZANDO INFORMAZIONI RICAVATE DA DEI CAMPIONI.



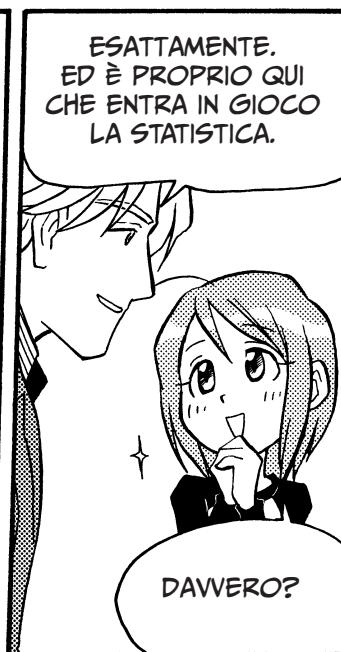
HO DETTO COSE TROPPO DIFFICILI?

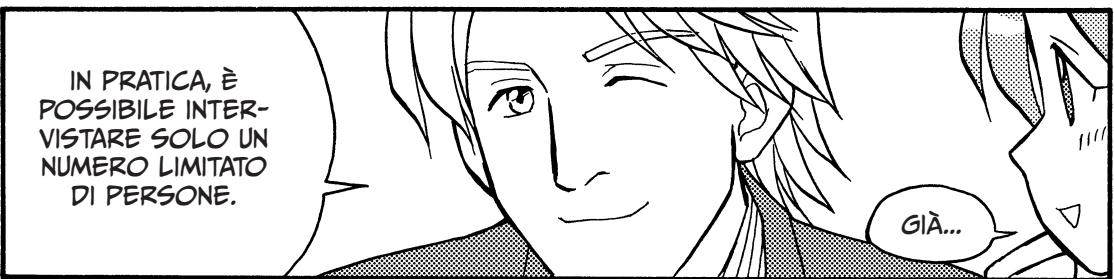
OH, ECCO
QUA UN
ESEMPIO.

E INCOMPENSIBILE
TUTTO
BENE,
RUI?



PRENDIAMO IL
GIORNALE DI
OGGI.





MA NO... STA SOLO DICENDO CHE NEL CASO DELLA POPOLARITÀ DEL GOVERNO, LA POPOLAZIONE È COSTITUITA DALL'INSIEME DI TUTTI GLI ELETTORI.

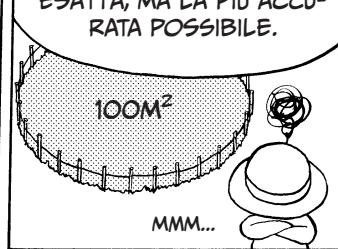
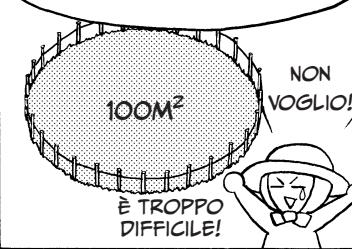
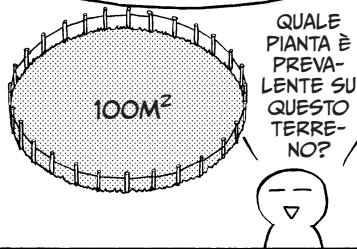
MA SE IL SONDAGGIO È STATO EFFETTUATO INTERVISTANDO 2.000 PERSONE, SONO QUESTE A COSTITUIRE IL CAMPIONE.



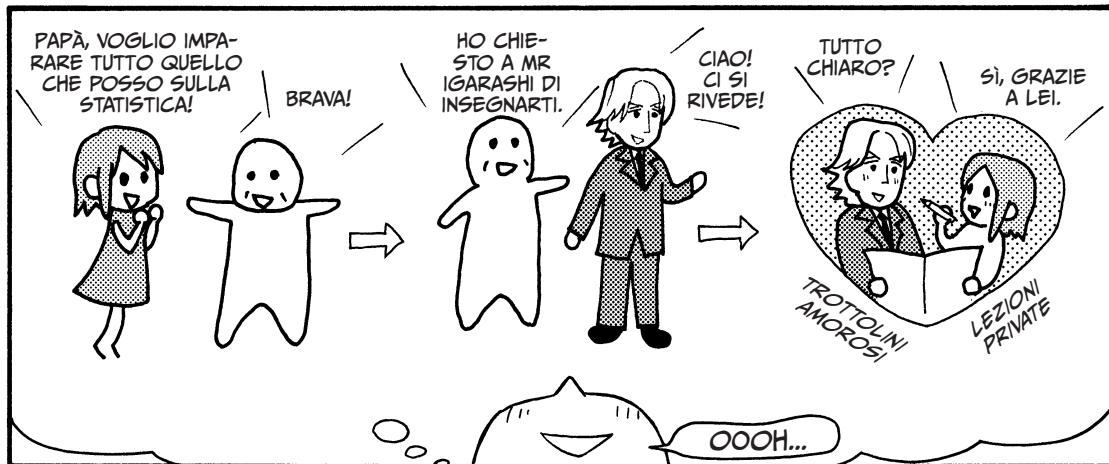
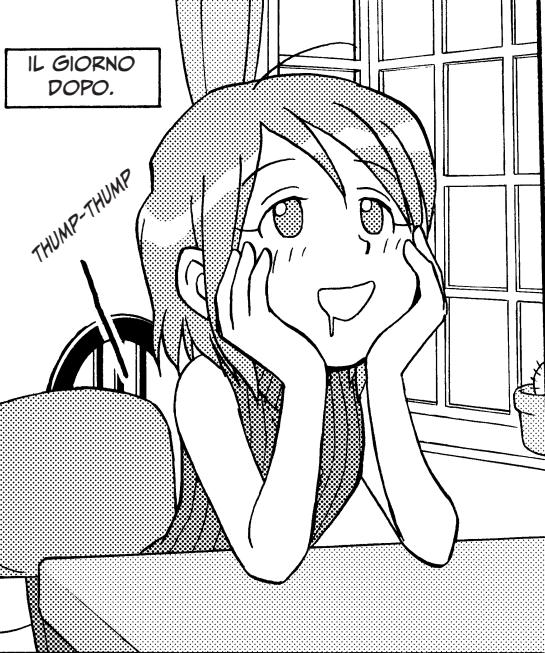
VORREI ANALIZZARE LA POPOLAZIONE...

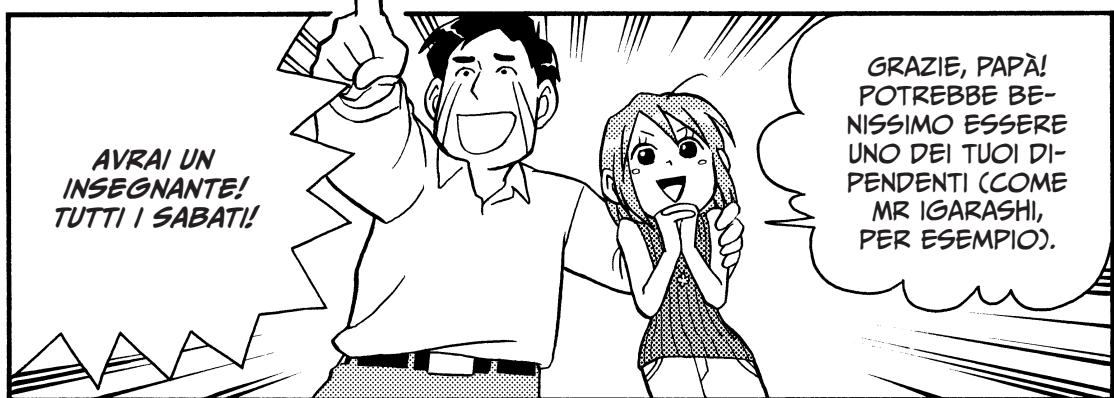
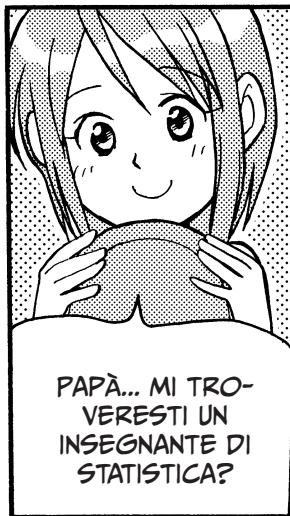
MA SCOPRO CHE È TECNICAMENTE IMPOSSIBILE. CHE COSA POSSO FARE?

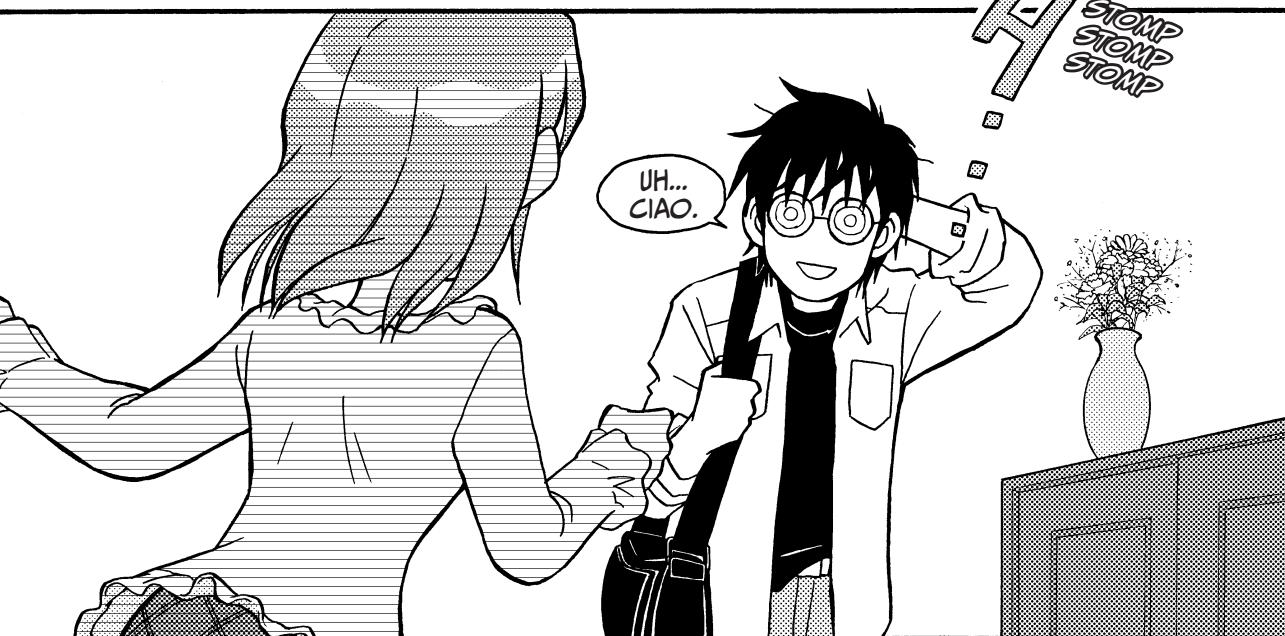
COME POSSO FARMI UN'IDEA DELLE CARATTERISTICHE DELLA POPOLAZIONE? NON DEV'ESSERE ESATTA, MA LA PIÙ ACCURATA POSSIBILE.

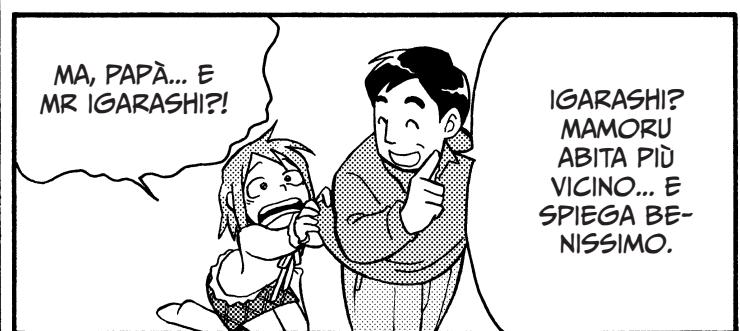


IL GIORNO
DOPO.

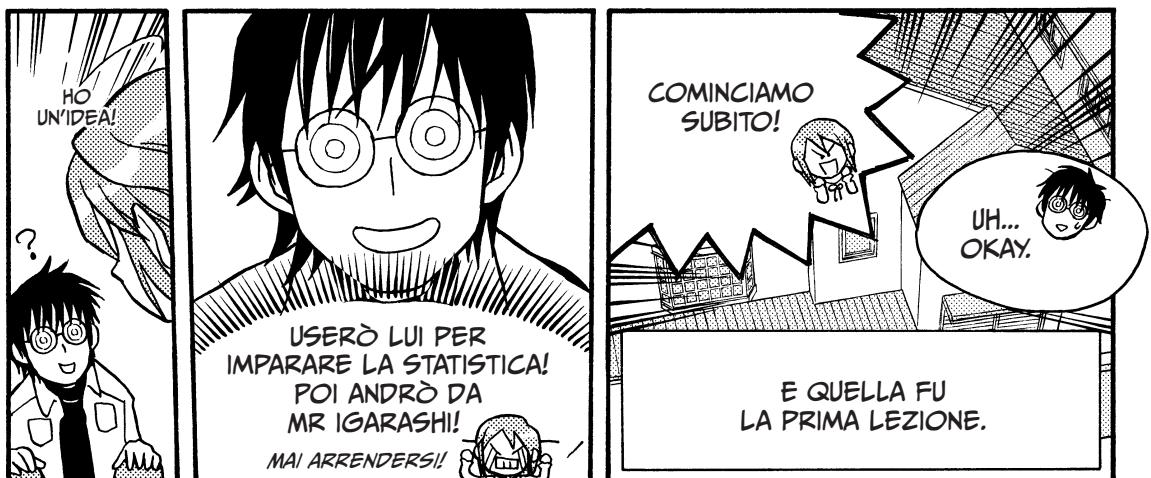
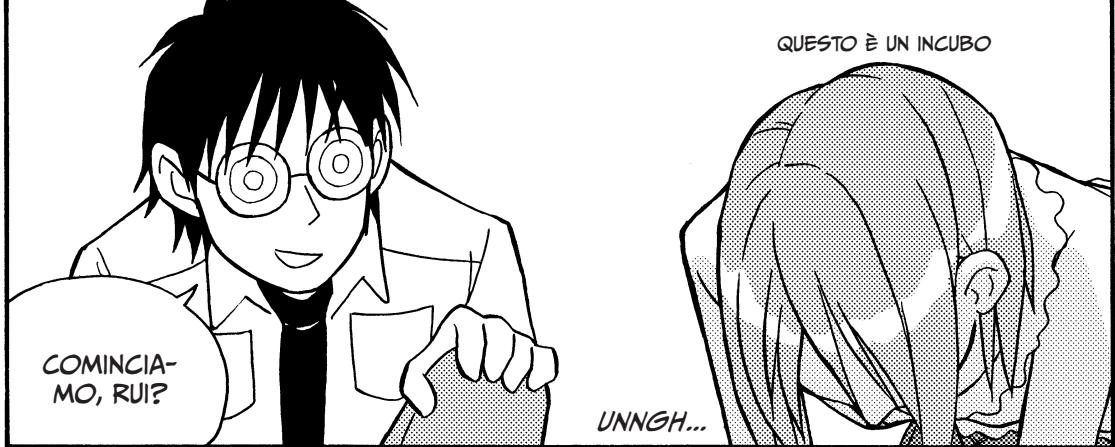








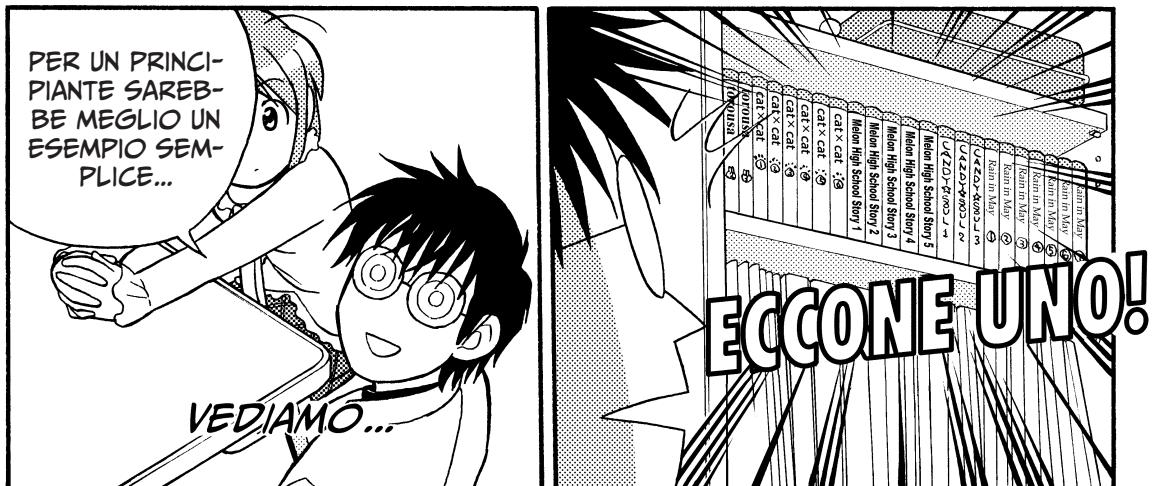
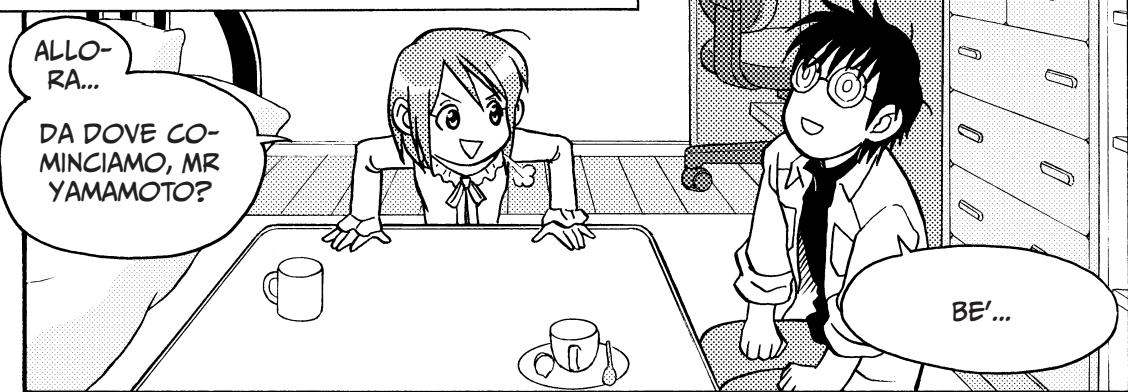
QUESTO È UN INCUBO

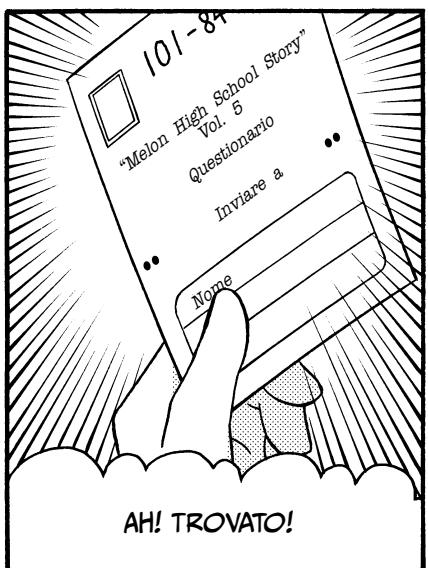
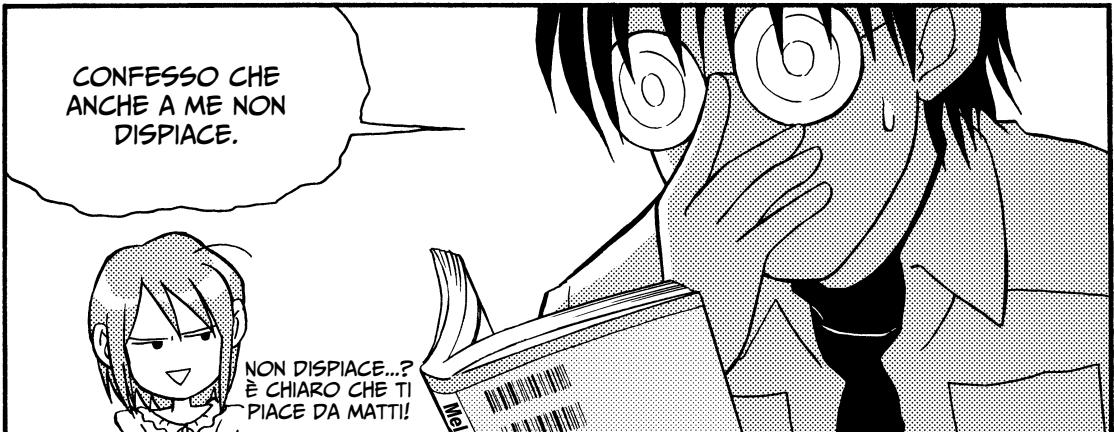


1

I TIPI DI DATI

1. DATI CATEGORICI E DATI NUMERICI





Melon High School Story Vol. 5 Questionario

D1. Cosa ne pensi di Melon High School Story Vol. 5?

1. Molto divertente
2. Abbastanza divertente
3. Normale
4. Piuttosto noioso
5. Molto noioso

D2. Sesso

1. Femmina
2. Maschio

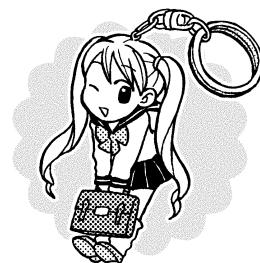
D3. Età

_____ anni

D4. Quanti fumetti acquisti al mese?

_____ serie

A trenta fortunati vincitori estratti tra chi invierà questo questionario sarà inviato un simpatico portachiavi!



GRAZIE PER LA TUA COLLABORAZIONE, LA TUA IMPORTANISSIMA OPINIONE VERRÀ TENUTA IN CONSIDERAZIONE NELLE PROSSIME USCITE

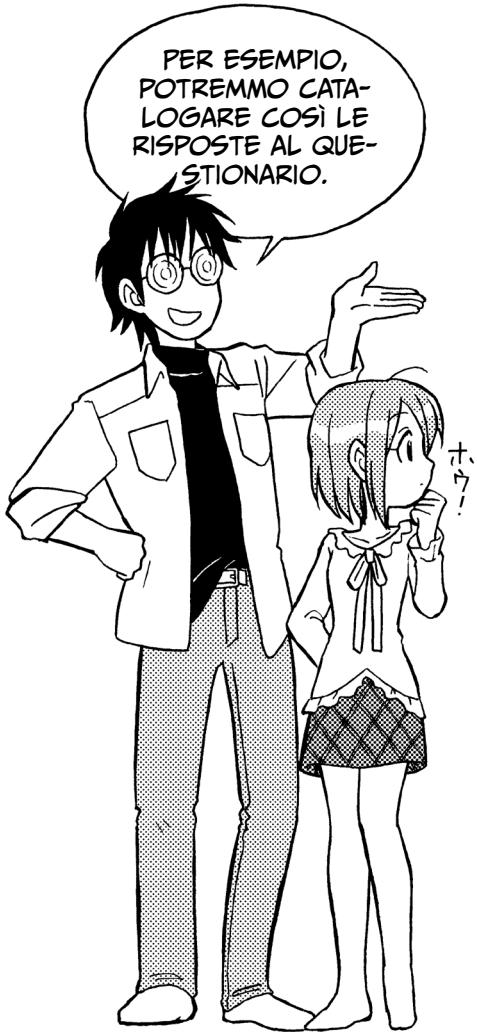


RISULTATI DEL QUESTIONARIO

INTERVISTATO	D1 COSA NE PENSI DI MELON HIGH SCHOOL STORY	D2 SESSO	D3 ETÀ	D4 ACQUISTI MENSILI
RUI	MOLTO DIVERTENTE	FEMMINA	17	2
A	ABbastanza divertente	FEMMINA	17	1
B	NORMALE	MASCHIO	18	5
C	PIUTTOSTO NOIOSO	MASCHIO	22	7
D	ABbastanza divertente	FEMMINA	25	4
E	MOLTO NOIOSO	MASCHIO	20	3
F	MOLTO DIVERTENTE	FEMMINA	16	1
G	ABbastanza divertente	FEMMINA	17	2
H	NORMALE	MASCHIO	18	0
I	NORMALE	FEMMINA	21	3
...







Melon High School Story Vol. 5 Questionario

D1. Cosa ne pensi di Melon High School Story Vol. 5?

1. Molto divertente
2. Abbastanza divertente
3. Non è
4. Piuttosto noioso
5. Molto noioso

NON MISURABILE

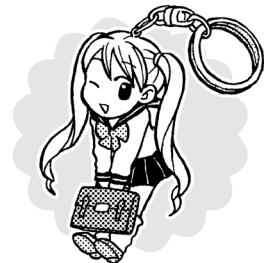
D2. Sesso

1. Femmina
2. Maschio

D3. Età _____ anni

D4. Quanti fumetti **MISURABILE** _____ serie acquisti al mese?

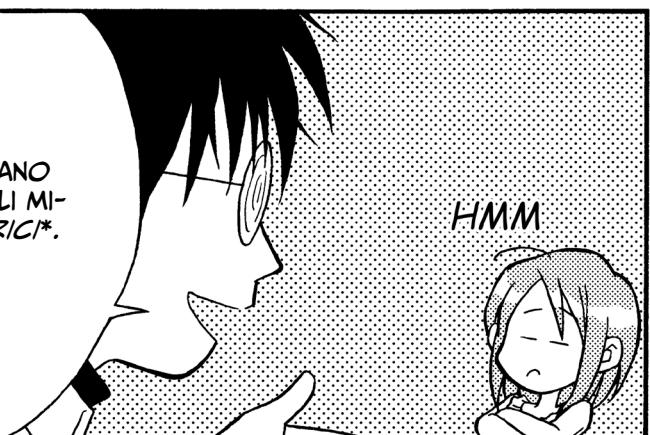
A trenta fortunati vincitori estratti tra chi invierà questo questionario sarà inviato un simpatico portachiavi!



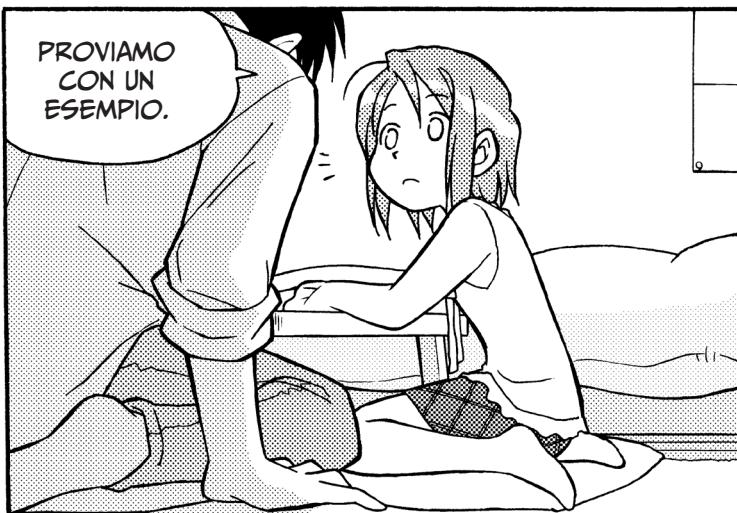
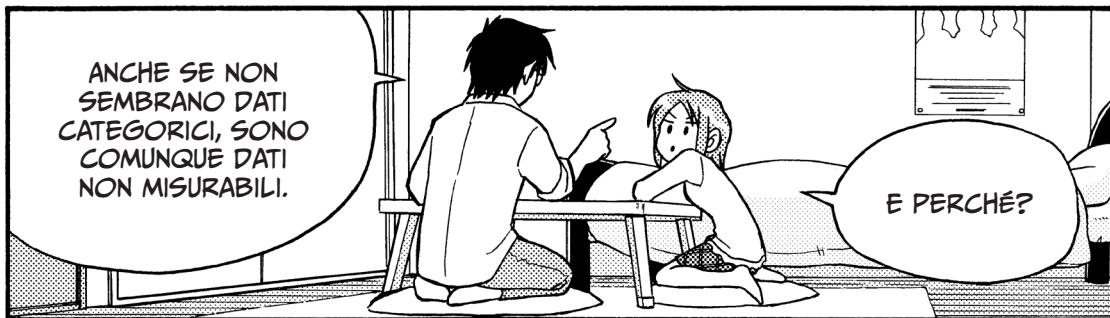
GRAZIE PER LA TUA COLLABORAZIONE, LA TUA IMPORTANISSIMA OPINIONE VERRÀ TENUTA IN CONSIDERAZIONE NELLE PROSSIME USCITE

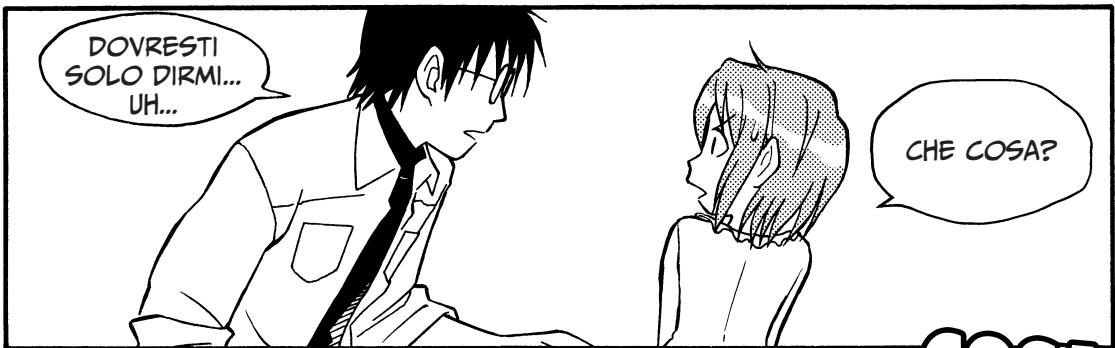
I DATI NON MISURABILI SI CHIAMANO DATI CATEGORICI MENTRE QUELLI MISURABILI SI DICONO DATI NUMERICI*.

*I DATI CATEGORICI A VOLTE VENGONO ANCHE DETTI QUALITATIVI E QUELLI NUMERICI QUANTITATIVI.

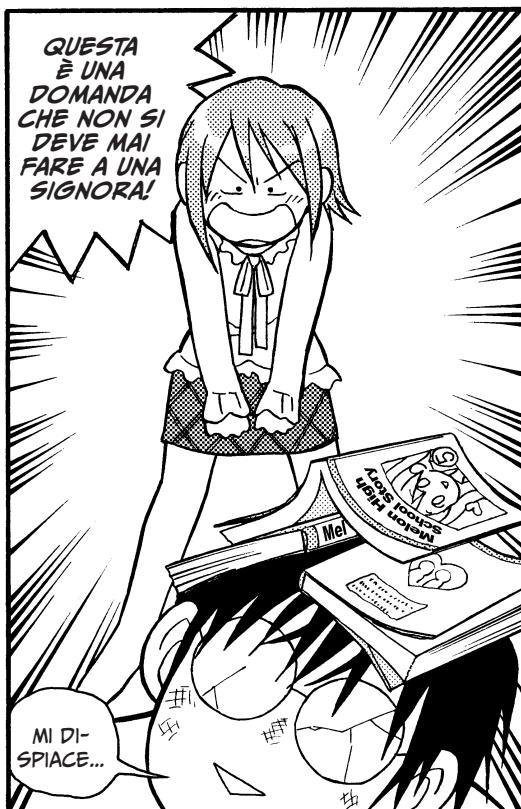


2. UN ESEMPIO DI DATI CATEGORICI COMPLICATI





COSA?



POM!

ESATTAMENTE
151 CM.

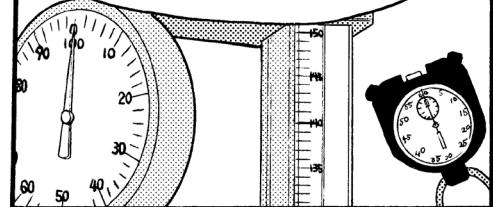
È SUDDIVISA IN
CENTIMETRI.

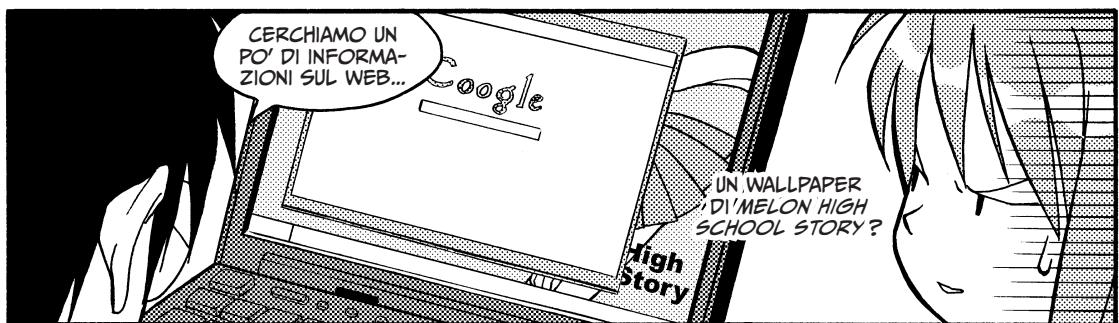


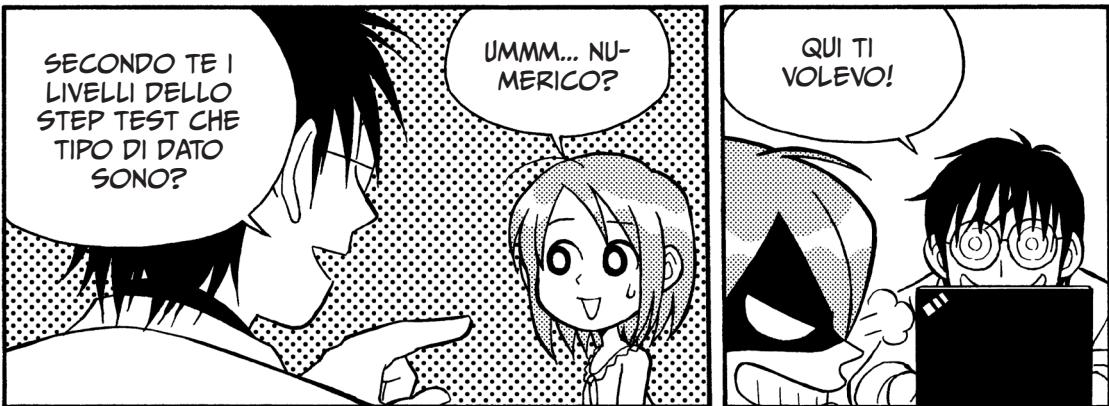
LA TACCA SOPRA 151 È 152 E
COSÌ VIA, CON 153, 154, SECON-
DO INTERVALLI UGUALI.



SE GLI INTERVALLI TRA
CIASCUNA TACCA SONO
UGUALI...



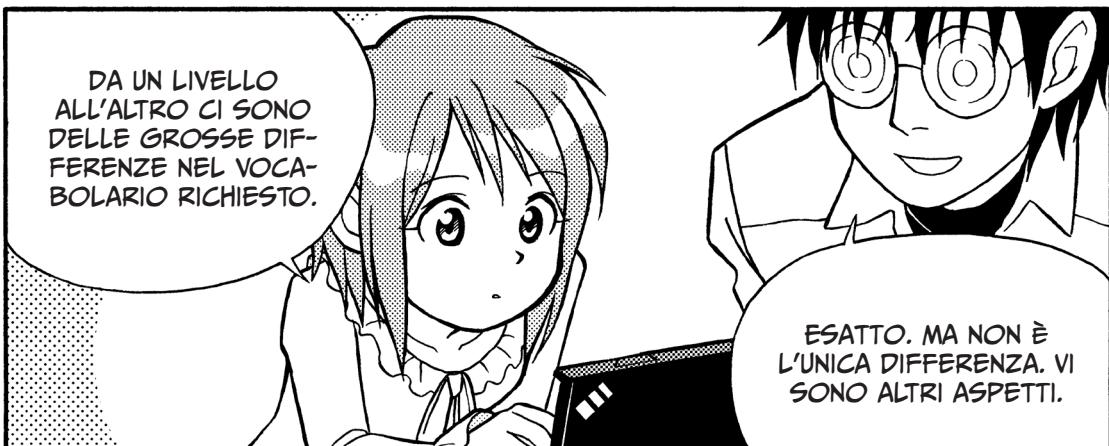




RISULTATI STEP-TEST

Voto	Requisito
Livello 1	Avanzato/Laurea; vocabolario di 10.000-15.000 parole
Livello 2	Diplomato; vocabolario di 5.100 parole
Livello 3	Terzo anno di Liceo; vocabolario di 2.100 parole
Livello 4	Secondo anno di Liceo; vocabolario di 1.300 parole
Livello 5	Matricola liceale; vocabolario di 600 parole

(fonte: Society for Testing English Proficiency <http://www.eiken.or.jp/>)





ORA DOVRESTI RIUSCIRE A RISONDERE A QUESTA DOMANDA...

Melon High School Story Vol. 5 Questionario

D1. Cosa ne pensi di Melon High School Story Vol. 5?

1. Molto divertente
2. Abbastanza divertente
3. Normale
4. Piuttosto noioso
5. Molto noioso

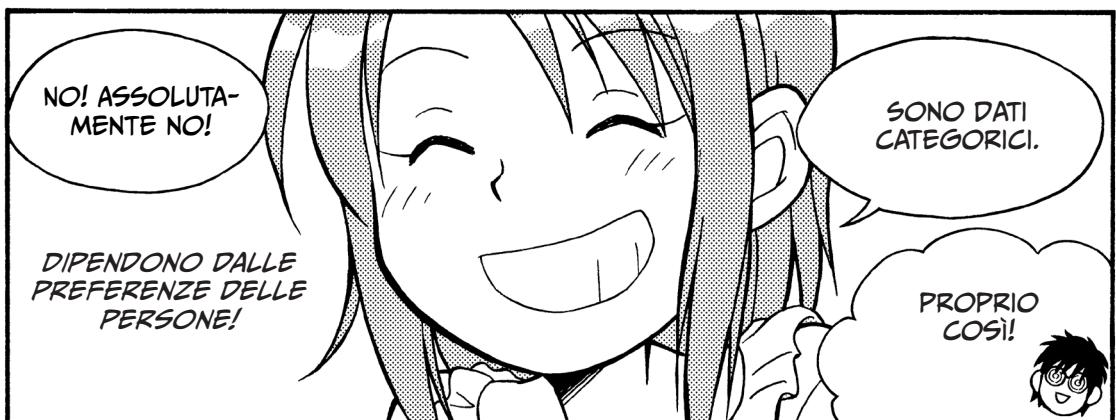
D2. Sesso

1. Femmina
2. Maschio

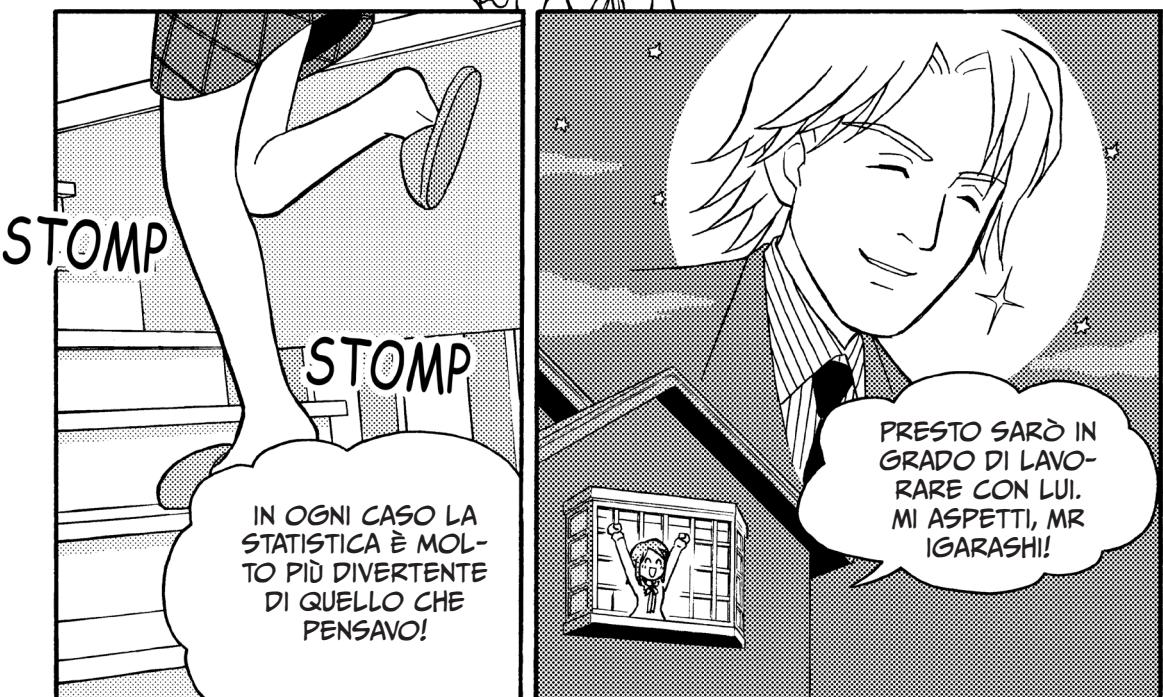
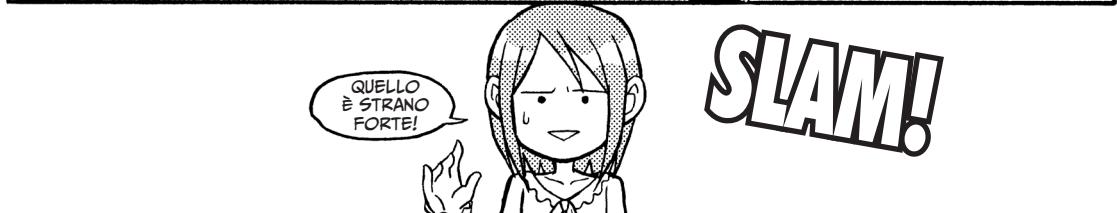
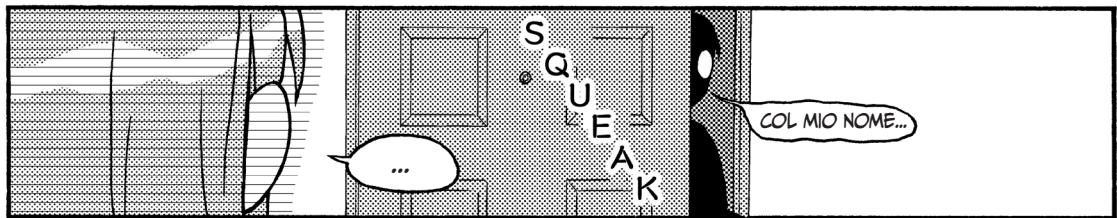
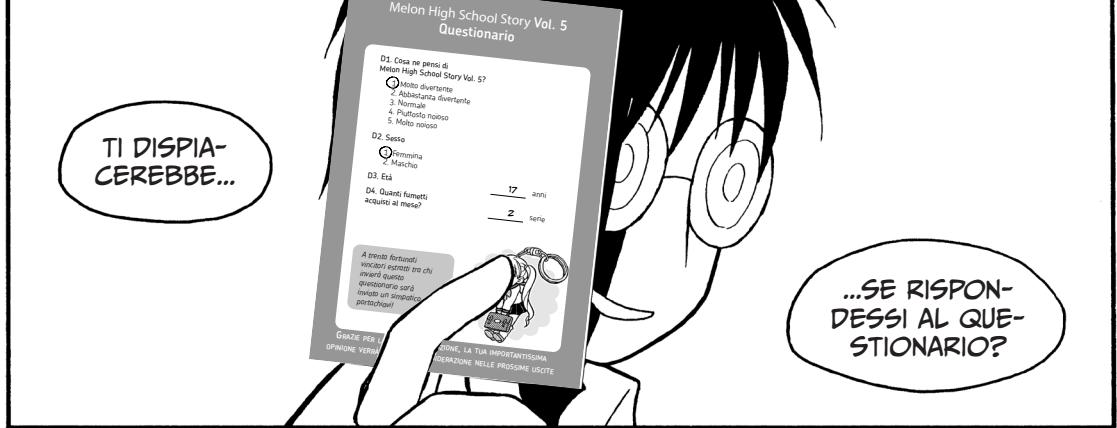
D3. Età _____ anni

D4. Quanti fumetti acquisti al mese? _____ serie

GLI INTERVALLI TRA LE OPPORTUNITÀ PROPOSTE IN D1 SONO UGUALI?







3. COME VENGONO GESTITE LE RISPOSTE MULTIPLE



Come detto a pagina 25, le risposte multiple come quelle alla prima domanda del sondaggio tra i lettori sono dati categorici. In pratica però, quando si elaborano i questionari dei consumatori, è possibile gestirle come se fossero dati numerici. Ecco alcuni esempi:

Molto divertente	⇒	5 punti
Abbastanza divertente	⇒	4 punti
Normale	⇒	3 punti
Abbastanza noioso	⇒	2 punti
Molto noioso	⇒	1 punto

Molto divertente	⇒	2 punti
Abbastanza divertente	⇒	1 punto
Normale	⇒	0 punti
Abbastanza noioso	⇒	-1 punto
Molto noioso	⇒	-2 punti

Gli stessi dati possono essere gestiti in maniera diversa nella teoria e nella pratica. E non dimenticate che in situazioni diverse i dati potrebbero essere categorizzati in maniera diversa.

ESERCIZIO CON SOLUZIONE

ESERCIZIO

Stabilire se i dati nella seguente tabella sono numerici o categorici

Intervistato	Gruppo sanguigno	Valutazione Bevanda energetica X	Temperatura ideale aria condizionata	tempo sui 100 m piani (in secondi)
Mr./Ms. A	B	Non buona	25	14.1
Mr./Ms. B	A	Buona	24	12.2
Mr./Ms. C	AB	Buona	25	17.0
Mr./Ms. D	O	Normale	27	15.6
Mr./Ms. E	A	Non buona	24	18.4
...

SOLUZIONE

Il gruppo sanguigno e la valutazione sulla bevanda X sono esempi di dati categorici. Una temperatura ideale per l'aria condizionata e il risultato sui 100 m piani sono esempi di dati numerici.

RIASSUMENDO

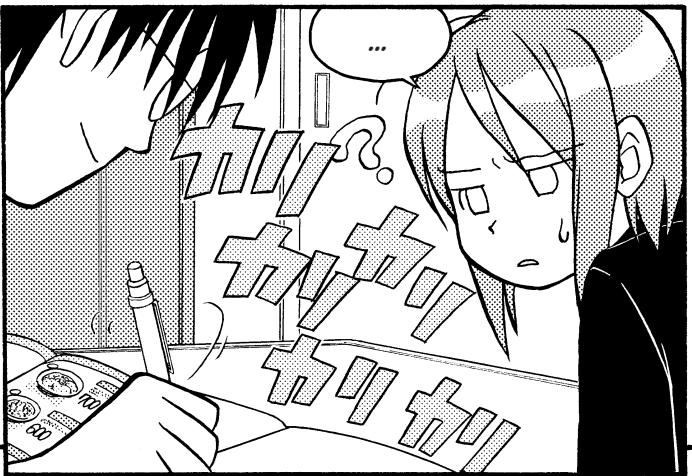
- I dati si classificano come categorici oppure numerici.
- Alcuni dati che in teoria sarebbero categorici, come "molto divertente" o "molto noioso", in pratica possono essere gestiti come numerici.

Z

UNO SGUARDO D'INSIEME:
COME CAPIRE I DATI NUMERICI

1. TABELLE DI DISTRIBUZIONE DI FREQUENZA E ISTOGRAMMI





PREZZI NEI RISTORANTI DI RAMEN (DA I 50 MIGLIORI RISTORANTI DI RAMEN)

RISTORANTE	PREZZO	RISTORANTE	PREZZO
1	700	26	780
2	850	27	590
3	600	28	650
4	650	29	580
5	980	30	750
6	750	31	800
7	500	32	550
8	890	33	750
9	880	34	700
10	700	35	600
11	890	36	800
12	720	37	800
13	680	38	880
14	650	39	790
15	790	40	790
16	670	41	780
17	680	42	600
18	900	43	670
19	880	44	680
20	720	45	650
21	850	46	890
22	700	47	930
23	780	48	650
24	850	49	777
25	750	50	700

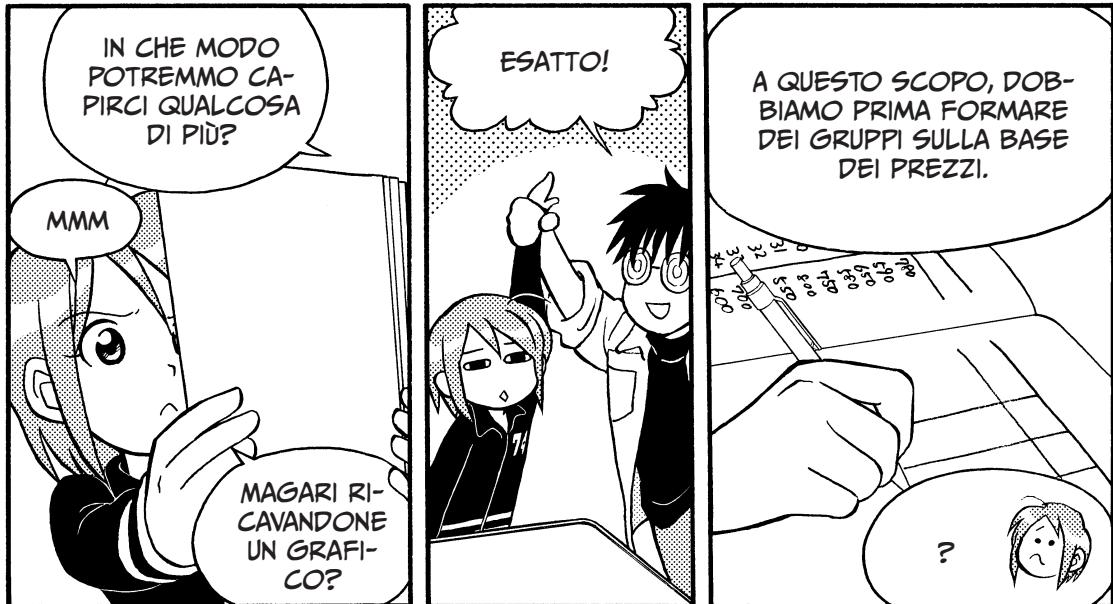
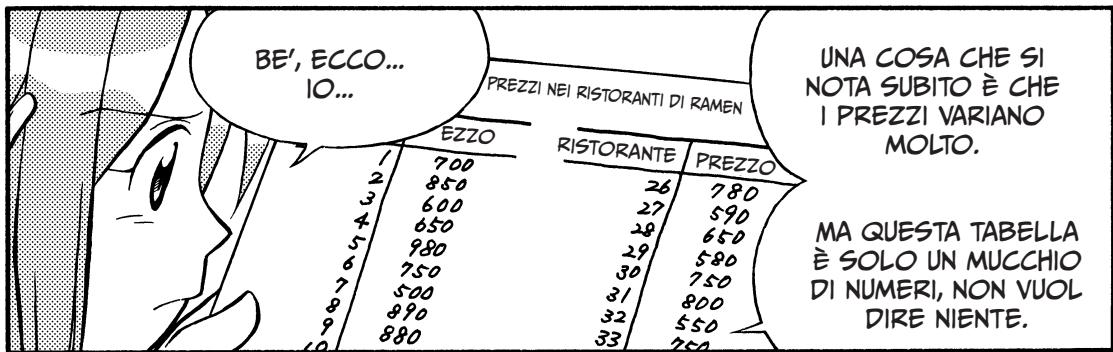
ECCO UNA
TABELLA DEI
PREZZI

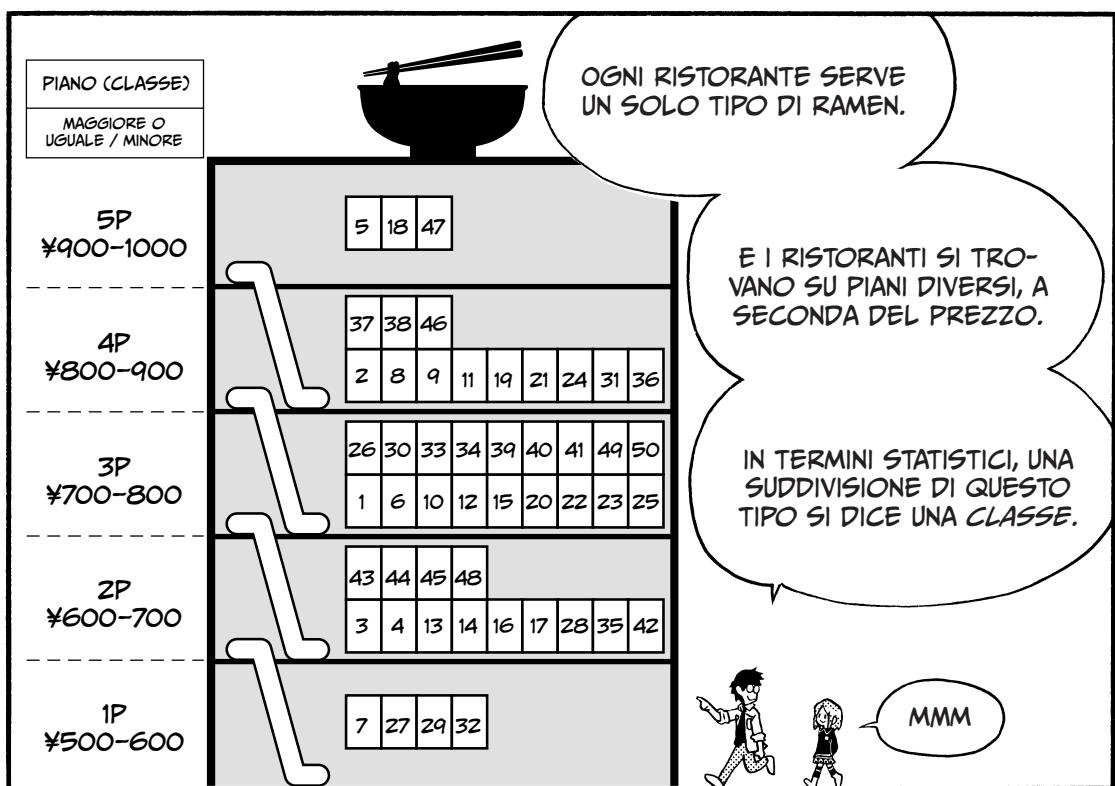


HA COMINCIATO
LA LEZIONE
ALL'IMPROVISO!

NON SEI
NORMALE!







A OGNI PIANO
C'È UN'INSEGNA
CHE INDICA IL
PREZZO MEDIO
DI CIASCUNA
CLASSE.

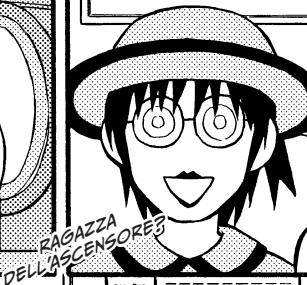
2F
650

IL SECONDO
PIANO È NELLA CLAS-
SE DEI 600-700 YEN,
QUINDI L'INSEGNA
INDICA 650!

I RISTORANTI VENGONO
COLLOCATI SUI VARI PIANI A
SECONDA DEI PREZZI CHE
PRATICANO, QUINDI IL LORO
NUMERO CAMBIA DA
PIANO A PIANO.

PIANTA DEI PIANI

PIANO	NOME DEL RISTORANTE	PREZZO MEDIO
2P ¥900-1000	■■■■	950
4P ¥800-900	■■■■■■■■	850
3P ¥700-800	■■■■■■■■■■	750
2P ¥600-700	■■■■■■■■■■■■	650
1P ¥500-600	■■■■■■■■■■■■■■	550



IL PREZZO
MEDIO TRA
MASSIMO E MI-
NIMO SI CHIAMA
VALORE CEN-
TRALE.

EH, EH,
EH!

800-900	=====	850
700-800	=====	750
600-700	====	650
500-600	====	

È VERO.

4 AL PRIMO
PIANO, 13 AL
SECONDO...

IL NUMERO DI
RISTORANTI A
CIASCUN PIANO
VIENE DETTO
FREQUENZA.

IL PIANO CON PIÙ
RISTORANTI È IL
TERZO... CE NE
SONO 18.

ADDESSO
CERCHIAMO DI CALCO-
LARE LA FREQUENZA
RELATIVA DEI RISTORANTI
DEL TERZO PIANO.



È IL RAPPORTO TRA UNA PARTE E IL TUTTO, CONVENENDO CHE IL TUTTO VALGA 1.

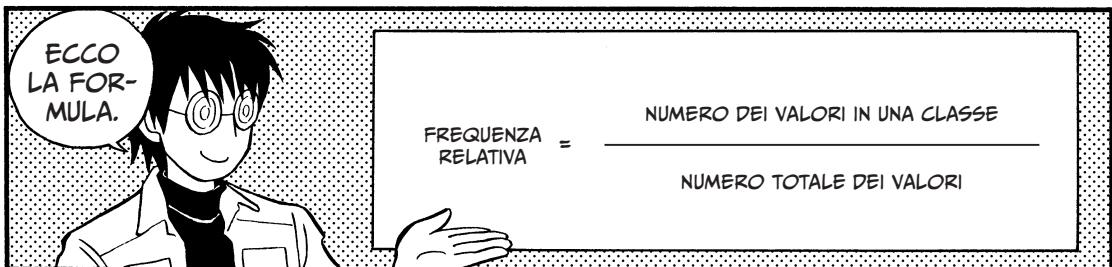
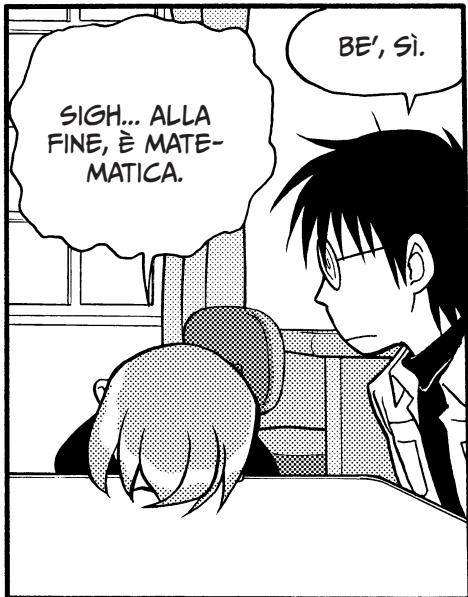




TABELLA DELLE FREQUENZE DEI 50 MIGLIORI RISTORANTI DI RAMEN

CLASSE (MAGGIORE O UGUALE / MINORE)	PREZZO MEDIO	FREQUENZA	FREQUENZA RELATIVA
500-600	550	4	0.08
600-700	650	13	0.26
700-800	750	18	0.36
800-900	850	12	0.24
900-1000	950	3	0.06
SOMMA		50	1.00

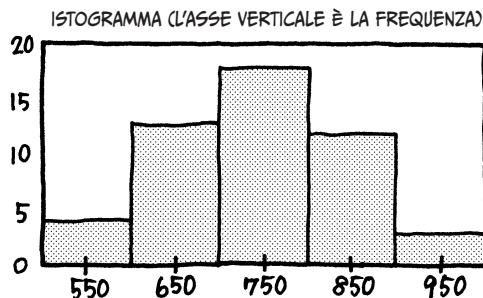


SULL'ASSE ORIZZONTALE ABBIAMO LE VARIABILI... IN QUESTO CASO, IL PREZZO DEL RAMEN.

LA LARGHEZZA DI CIASCUNA BARRA È L'AMPIEZZA DELLA CLASSE.

IL PUNTO MEDIO DELLA BASE DI CIASCUNA BARRA È IL VALORE CENTRALE.

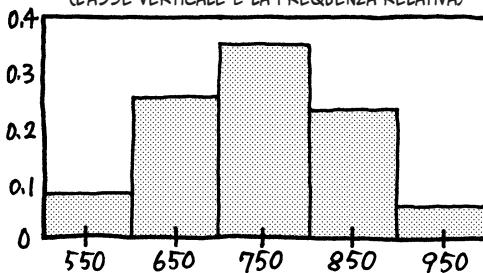
ISTOGRAMMA RICAVATO DALLA TABELLA DELLE FREQUENZE DEI 50 MIGLIORI RISTORANTI DI RAMEN



NEL PRIMO ISTOGRAMMA, SULL'ASSE VERTICALE ABBIAMO LA FREQUENZA.

E NEL SECONDO LA FREQUENZA RELATIVA.

ISTOGRAMMA (L'ASSE VERTICALE È LA FREQUENZA RELATIVA)



COSÌ È PIÙ FACILE?

BE' ...

FORSE...
PIÙ O
MENO...
COMIN-
CIO...

...AD AF-
FERRARE I
PREZZI DEL
RAMEN.

"AVERE L'IMPRES-
SIONE" DI AFFER-
RARE È IMPOR-
TANTE. LO SCOPO
DELLA TABELLA
DELLE FREQUENZE
E DELL'ISTOGRAM-
MA È FORNIRE UNA
MIGLIORE COM-
PRENSIONE DEI
DATI.

DAVVERO?

2. LA MEDIA

L'ALTRÒ GIORNO
SONO ANDATA AL
BOWLING CON LE MIE
COMPAGNE DI CLASSE.

PAUSA
CAFFÈ!

HAI BUTTATO GIÙ
QUALCHE BIRILLO?

COME TI PERMETTI?!
IO BUTTO GIÙ TE!

A BOWLING
SONO UN ASSO!

74

SCHERZAVO!

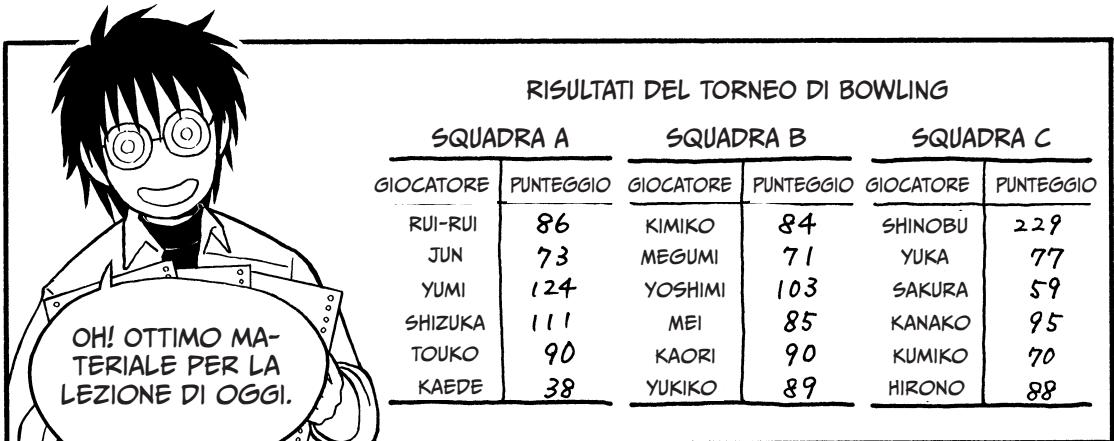
TUTTE LE TUE
COMPAGNE...
DOVEVATE
ESSERE UN
SACCO.

A

C

BE', ERAVAMO 18,
QUINDI ABBIAMO
FORMATO 3 SQUA-
DRE DA 6 E ABBI-
AMO GIOCATO UNA
PARTITA.

ECCO LE
SCHEDE
COI PUN-
TEGGI!



LA PARTITA È STATA GIOCATA A SQUADRE, QUINDI IMMAGINO CHE ABBIATE CONFRONTATO I PUNTEGGI DI CIASCUNA SQUADRA.

ESATTO.

LA MEDIA SI OTTIENE DIVIDENDO LA SOMMA DEI PUNTEGGI PER IL NUMERO DEI GIOCATORI, PERCIÒ...

SQUADRA A

$$\frac{86+73+124+111+90+38}{6} = \frac{522}{6} = 87$$

SQUADRA B

$$\frac{84+71+103+85+90+89}{6} = \frac{522}{6} = 87$$

SQUADRA C

$$\frac{229+77+59+95+70+88}{6} = \frac{618}{6} = 103$$

LA SQUADRA C È TROPPO FORTE!

LA MEDIA DELLA TUA SQUADRA È 87.

E IL PUNTEGGIO DI RUI-RUI È 86.

LO OFFRI TU A ME UN DOLCETTO?

E PERCHÉ?

CHE FASTIDIOSO



LA MEDIA DI CUI PARLAVAMO
PRIMA È QUELLA CHE, PIÙ PRE-
CISAMENTE, VIENE CHIAMATA
MEDIA ARITMETICA.

ESISTONO TANTE MEDIE DIVER-
SE, COME LA MEDIA GEOME-
TRICA E LA MEDIA ARMONICA.
PER ORA NON È NECESSARIO
RICORDARSI LE FORMULE, MA TI
CONSIGLIO DI TENERE A MENTE
QUESTI NOMI.



3. LA MEDIANA

ORA TORNIAMO AI PUNTEGGI.

COSA VUOI
FARE, ADES-
SO?

PER ORA
IGNORIAMO
LE SQUA-
DRE A E B,
E DIAMO
UN'OCCHIA-
TA A C...

RISULTATI DEL TORNEO DI BOWLING

SQUADRA A		SQUADRA B		SQUADRA C	
GIOCATORE	PUNTEGGIO	GIOCATORE	PUNTEGGIO	GIOCATORE	PUNTEGGIO
RUI-RUI	86	KIMIKO	84	SHINOBU	229
JUN	73	MEGUMI	71	YUKA	77
YUMI	124	YOSHIMI	103	SAKURA	59
SHIZUKA	111	MEI	85	KANAKO	95
TOUKO	90	KAORI	90	KUMIKO	70
KAEDE	38	YUKIKO	89	HIRONO	88

QUI È DIF-
FICILE DIRE
CHE LA MEDIA
È "CIRCA IL
PUNTEGGIO DI
CIASCUNA GIO-
CATRICE".

È VERO. LA MEDIA È
SUPERIORE A 100... MA
5 GIOCATORI HANNO
REALIZZATO DI MENO.

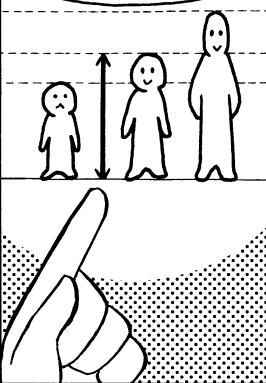
CHE
BRAVA
SHINO-
BU

IN CASI COME QUESTO,
IN PRESENZA DI VALORI
MOLTO GRANDI O MOLTO
PICCOLI...

...INVECE DELLA
MEDIA È PIÙ SIGNI-
FICATIVO USARE LA
MEDIANA.

LA MEDIANA?!

LA MEDIANA È IL VALORE "DI MEZZO" RISULTANTE QUANDO SI ORDINANO I VALORI DISPONIBILI.



PER PRIMA COSA,
METTIAMO I PUNTEGGI
DI CIASCUNA SQUADRA IN ORDINE DI
GRANDEZZA.



SQUADRA A

38 73 86 90 111 124

SQUADRA B

71 84 85 89 90 103

SQUADRA C

59 70 77 88 95 229

NUMERO DISPARI DI VALORI

-1041.6 -39.0 **-5.7** 60.4 77.3

MEDIANA

NUMERO PARI DI VALORI

-0.4 35.2 **37.8** 42.2 46.1 910.3

LA MEDIANA È LA MEDIA DI QUESTI DUE

SE ABBIAMO UN NUMERO DISPARI DI VALORI, LA MEDIANA È IL PUNTEGGIO CENTRALE.

SE IL NUMERO DEI VALORI È PARI, COME NEL CASO DI QUESTA PARTITA DI BOWLING, LA MEDIANA È LA MEDIA ARITMETICA DEI DUE VALORI CENTRALI.

CALCOLIAMO LA MEDIANA DELLA SQUADRA C?



CERTO! È $(77 + 88) \div 2 = 82.5$.

ESATTO!

UN ALTRO SUGGERIMENTO, A PROPOSITO DI MEDIE...



UN DOLCETTO NO, EH...?

CERTO, UN SUGGERIMENTO NON SI MANGIA, MA NON È MENO UTILE DI UN DOLCE...

HAI DA PARTE QUALCHE RISPARMIO?



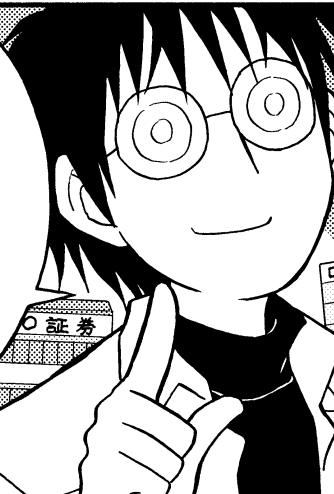
CERTO.

ANCHE SE ALLA FINE HO MENO DI ¥10.000*.

*MENO DI 100 EURO

QUINDI POTRESTI CHIEDERTI COME MAI IL "RISPARMIO MEDIO" RIPORTATO DAI GIORNALI E IN TV È COSÌ ALTO.

銀行
証券



BE', SÌ. IO HO MOLTO MENO, E NEANCHE IL MIO PAPÀ È COSÌ RICCO.



4. DEVIAZIONE STANDARD

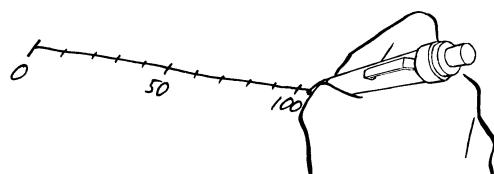
ADESSO DIAMO UN'OCCHIATA AI PUNTEGGI

DELLE SQUADRE A E B.

OKAY.



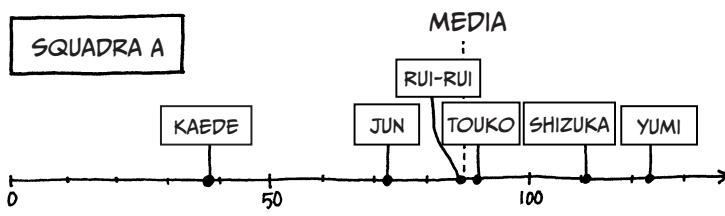
TRACCIAVAMO UN ASSE NUMERATO.



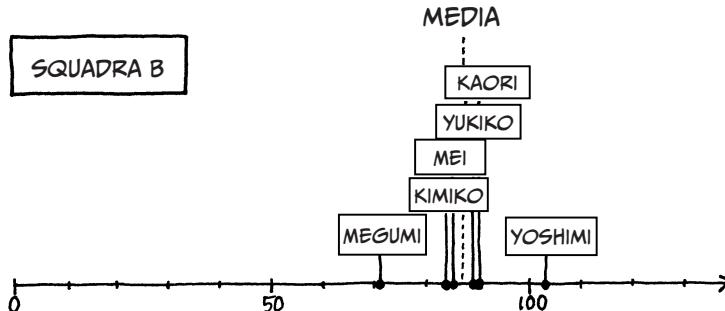
POI, IN CORRISPONDENZA DEI PUNTEGGI RIPORTIAMO I NOMI DEI GIOCATORI.



SQUADRA A



SQUADRA B



ANCHE SE LA MEDIA DI ENTRAMBE LE SQUADRE È 87...

...I DUE ASSI ILLISTRANO ANDAMENTI MOLTO DIVERSI.

POCO MA SICURO.
I PUNTEGGI DELLA
SQUADRA A SONO
SIA ALTI CHE BASSI,
MENTRE QUELLI DEL-
LA SQUADRA B SONO
MOLTO PIÙ SIMILI
TRA LORO.

PER DESCRIVERE
QUESTA DISPER-
SIONE TRA I DATI SI
UTILIZZA LA DEVI-
AZIONE STANDARD.

IN BREVE, LA DEVI-
AZIONE STANDARD È UN
INDICATORE DELLA DI-
STANZA DALLA MEDIA
DI CIASCUN VALORE
DELL'INSIEME.

UN'ALTRA PA-
ROLA NUOVA?

INDICATORE DELLA
DISTANZA...?

LA DEVIAZIONE STANDARD MINIMA
È ZERO E, ALL'AUMENTARE DELLA
"DISPERSIONE DEI DATI", AUMENTA
ANCHE LA DEVIAZIONE.



PROVA A INDOVINARE... HA UNA
DEVIAZIONE STANDARD MAG-
GIORE LA SQUADRA A O LA
SQUADRA B?

MMM...
FORSE LA A?

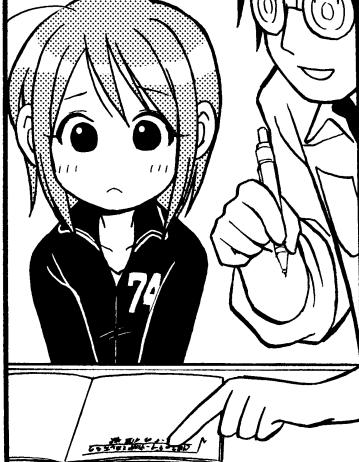
 ESATTO. LA FORMULA
È QUESTA:

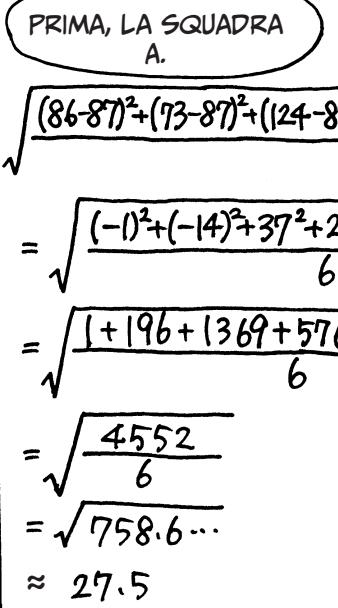
 IMPROVVISAMENTE,
SEMPRA DI NUOVO
MATEMATICA.

$$\frac{\text{SOMMA DI } ((\text{CIASCUN VALORE} - \text{LA MEDIA})^2)}{\text{NUMERO DEI VALORI}}$$

 È FACILE. DEVI
SOLO INSERIRE UN
PO' DI NUMERI NELLA
FORMULA.

PROVIAMOCI INSIEME.

 OKAY,
PROVIA-
MOCI...

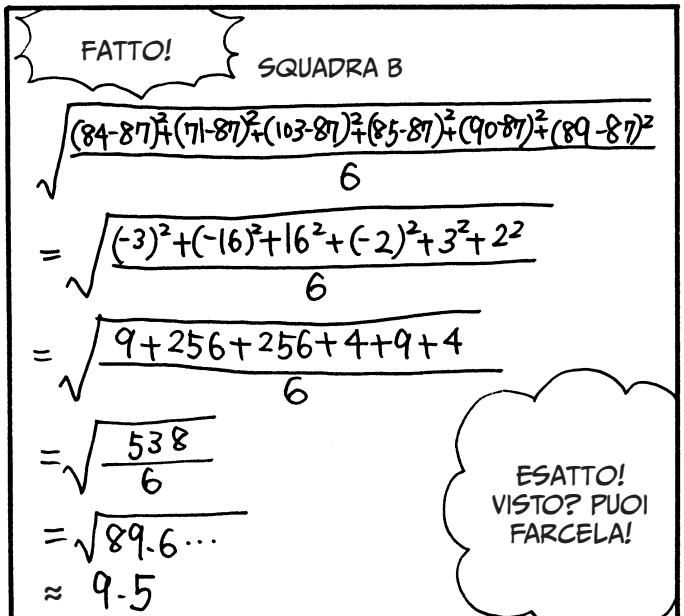
 PRIMA, LA SQUADRA
A.

SQUADRA A

$$\begin{aligned}& \frac{(86-87)^2 + (73-87)^2 + (124-87)^2 + (111-87)^2 + (90-87)^2 + (38-87)^2}{6} \\&= \sqrt{\frac{(-1)^2 + (-14)^2 + 37^2 + 24^2 + 3^2 + (-49)^2}{6}} \\&= \sqrt{\frac{1 + 196 + 1369 + 576 + 9 + 2401}{6}} \\&= \sqrt{\frac{4552}{6}} \\&= \sqrt{758.6\dots} \\&\approx 27.5\end{aligned}$$

 È PIÙ SEMPLI-
CE DI QUAN-
TO PENSASSI.
CREDO DI
POTERCELÀ
FARE.

 ALLORA PROVA DA
SOLA A CALCOLA-
RE LA DEVIAZIONE
STANDARD DELLA
SQUADRA B.



DEVIAZIONE STANDARD

SQUADRA A = 27.5 SQUADRA B = 9.5

I MEMBRI DELLA SQUADRA B HANNO PUNTEGGI PIÙ SIMILI TRA LORO, QUINDI LA DEVIAZIONE STANDARD È INFERIORE A QUELLA DELLA SQUADRA A.

COME DICEVO, LA FORMULA PER
LA DEVIAZIONE STANDARD È

$$\checkmark \frac{\sqrt{\text{SOMMA DI (CIASCUN VALORE - LA MEDIA)}^2}}{\text{NUMERO DEI VALORI}}$$

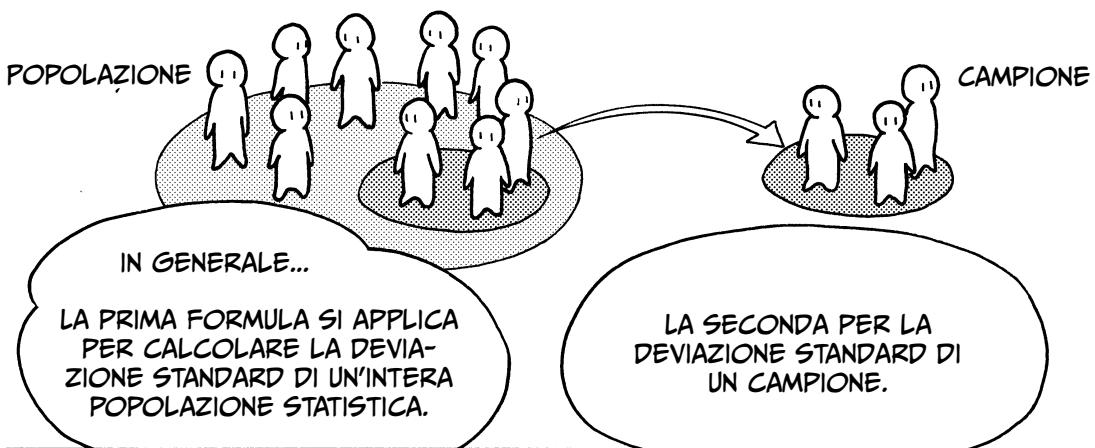
ESISTE ANCHE UN'ALTRA FORMULA:

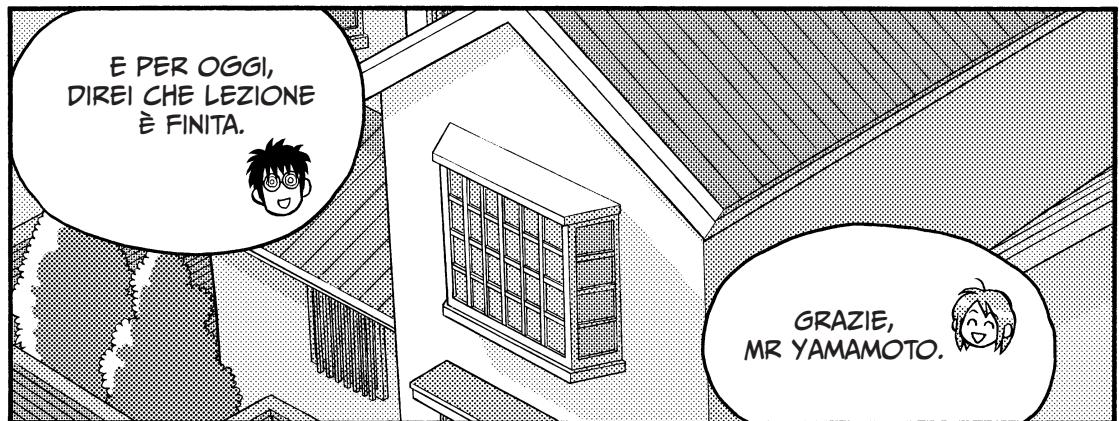
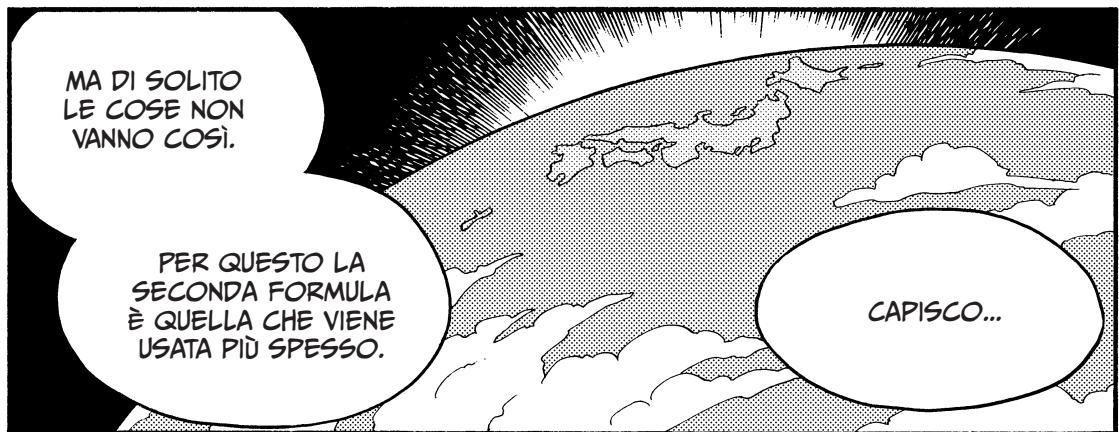
$$\checkmark \frac{\sqrt{\text{SOMMA DI (CIASCUN VALORE - LA MEDIA)}^2}}{\text{NUMERO DEI VALORI} - 1}$$

UN'ALTRA
FORMULA?

HAI TOLTO 1 DAL
NUMERO DEI VA-
LORI?

SÌ.





5. L'AMPIEZZA DI CLASSE DI UNA TABELLA DELLE FREQUENZE



Se le “Tabelle di distribuzione di frequenza e istogrammi” di pagina 32 vi sono sembrate poco chiare, diamo un’altra occhiata a quella introdotta a pagina 38:

TABELLA 2.1 – FREQUENZA DEI 50 MIGLIORI RISTORANTI DI RAMEN

Classe (Maggiore o uguale / minore)	Prezzo medio	Frequenza	Frequenza relativa
500-600	550	4	0.08
600-700	650	13	0.26
700-800	750	18	0.36
800-900	850	12	0.24
900-1000	950	3	0.06
Somma		50	1.00

Come si può vedere, l’ampiezza di classe della tabella è 100: è arbitraria e non è stata stabilita in accordo con un qualche standard matematico: l’ho decisa in maniera soggettiva. In altre parole, la scelta dell’ampiezza di una classe viene effettuata da chi analizza i dati.

Ma non dovrebbe forse esserci un modo di stabilirlo matematicamente? In fondo, una tabella delle frequenze con un’ampiezza soggettiva potrebbe sembrare meno valida.

Un modo del genere in effetti esiste e lo vedremo nelle prossime pagine, insieme a un esempio basato sui dati della tabella 2.1.

Passo 1

Applichiamo la formula di Sturges per calcolare il numero delle classi:

$$1 + \frac{\log_{10}(\text{numero dei valori})}{\log_{10}2}$$

$$1 + \frac{\log_{10}50}{\log_{10}2} = 1 + 5.6438\dots = 6.6438\dots \approx 7$$

Passo 2

Calcoliamo l'ampiezza di classe con la seguente formula:

$$\frac{(\text{valore massimo}) - (\text{valore minimo})}{\text{numero delle classi calcolato con la formula di Sturges}}$$

$$\frac{980 - 500}{7} = \frac{480}{7} = 68.5714\dots \approx 69$$

Questa è una tabella di frequenza riorganizzata usando l'ampiezza di classe calcolata con la formula del Passo 2.

TABELLA 2.2 – TABELLA DI FREQUENZA DEI 50 MIGLIORI RISTORANTI DI RAMEN
(AMPIEZZA DI CLASSE DETERMINATA MATEMATICAMENTE)

Classe (Maggiore o uguale / minore)	Prezzo medio	Frequenza	Frequenza relativa
500-569	534.5	2	0.04
569-638	603.5	5	0.10
638-707	672.5	15	0.30
707-776	741.5	6	0.12
776-845	810.5	10	0.20
845-914	879.5	10	0.20
914-983	948.5	2	0.04
Somma		50	1.00

Che ne pensate? Questa tabella vi sembra meno convincente della 2.1? E perché l'ampiezza è di 69 yen?

Se spiegate a delle persone che “è stata calcolata usando la formula di Sturges” la maggior parte vi risponderà “Chi se ne frega di Sturges... o come diavolo si chiama! Perché hai scelto un'ampiezza assurda come 69 yen?”

In sostanza, alcuni potrebbero avere difficoltà ad accettare un'ampiezza scelta in maniera soggettiva e quindi arbitraria. Tuttavia, come mostra la tabella, formule come quella di Sturges non producono necessariamente una tabella soddisfacente. Insomma, una tabella delle frequenze è solo uno strumento che deve aiutare a visualizzare i dati, e l'analista dovrebbe semplicemente scegliere un'ampiezza che ritiene adeguata.

6. TEORIA DELLA STIMA E STATISTICA DESCRIPTIVA

Nel prologo abbiamo spiegato come la statistica possa fornire stime su caratteristiche di una popolazione sulla base di informazioni raccolte da campioni. A dire la verità, questo non è necessariamente corretto.

Possiamo distinguere approssimativamente due tipi di statistica: la teoria della stima e la statistica descrittiva. Quella introdotta nel prologo è la prima. Cosa sarebbe allora la statistica descrittiva? È un tipo di statistica il cui scopo è descrivere la configurazione di un gruppo in maniera semplice e chiara attraverso l'organizzazione dei dati. In altre parole, la statistica descrittiva considera il gruppo come una popolazione.

Questa definizione potrebbe sembrare astratta, quindi vediamo di chiarire le cose con un esempio. Ricordate quando abbiamo calcolato la media e la deviazione standard della squadra di Rui? Non stavamo cercando di descrivere una popolazione a partire da informazioni riguardanti quella squadra: abbiamo calcolato media e deviazione standard semplicemente perché volevamo dare una descrizione semplice della configurazione della squadra di Rui. Questa è la statistica descrittiva.

ESERCIZIO CON SOLUZIONE



ESERCIZIO

Questa tabella raccoglie i tempi delle ragazze del liceo sui 100 metri piani.

Corridore	Tempo nei 100m piani (in secondi)
Ms. A	16.3
Ms. B	22.4
Ms. C	18.5
Ms. D	18.7
Ms. E	20.1

1. Qual è la media?
2. Qual è la mediana?
3. Qual è la deviazione standard?

SOLUZIONE

1. La media aritmetica è $\frac{16.3 + 22.4 + 18.5 + 18.7 + 20.1}{5} = \frac{96}{5} = 19.2$

2. La mediana è 18.7. 16.3 18.5 **18.7** 20.1 22.4

3. La deviazione standard è

$$\begin{aligned}& \sqrt{\frac{(16.3 - 19.2)^2 + (22.4 - 19.2)^2 + (18.5 - 19.2)^2 + (18.7 - 19.2)^2 + (20.1 - 19.2)^2}{5}} \\&= \sqrt{\frac{(-2.9)^2 + 3.2^2 + (-0.7)^2 + (-0.5)^2 + 0.9^2}{5}} \\&= \sqrt{\frac{8.41 + 10.24 + 0.49 + 0.25 + 0.81}{5}} \\&= \sqrt{\frac{20.2}{5}} \\&= \sqrt{4.04} \\&\approx 2.01\end{aligned}$$

RIASSUMENDO

- Per avere uno sguardo d'insieme intuitivo sui dati, create una tabella delle frequenze o un istogramma.
- Potete fissare l'ampiezza di classe della tabella usando la formula di Sturges.
- Per visualizzare matematicamente i dati, calcolate media aritmetica, mediana e deviazione standard.
- Se l'insieme comprende un valore estremamente grande o estremamente piccolo, è meglio usare la mediana invece che la media.
- La deviazione standard è un indice per descrivere "il grado di dispersione" dei dati.

3

UNO SGUARDO D'INSIEME:
COME CAPIRE I DATI CATEGORICI

1. TABELLE DI CONTINGENZA

I DATI CATEGORICI SONO
QUEI DATI CHE NON SI
POSSENO MISURARE,
RICORDI?

SÌ!
PER ORA,
ALMENO...



OGGI INDOSSI
L'UNIFORME DEL-
LA SCUOLA.

OH,
QUE-
STA?

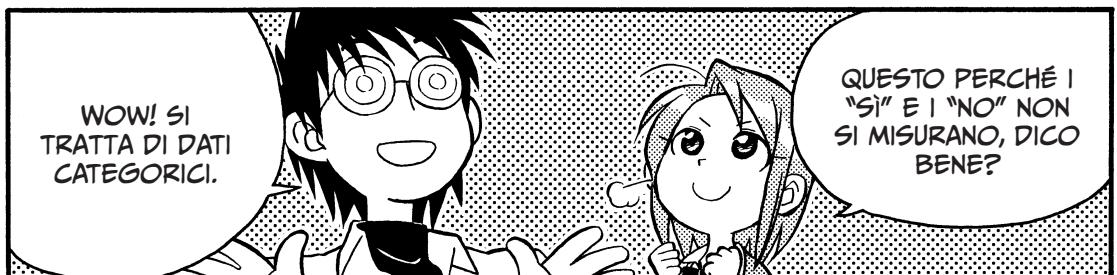
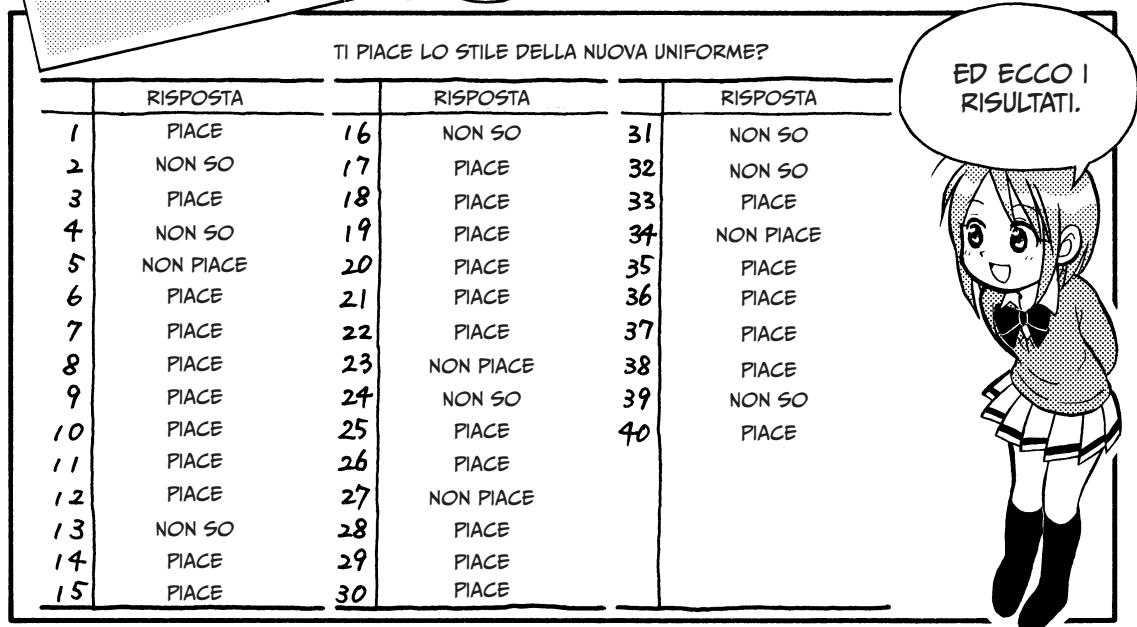
MI MAN-
CHERÀ.

PRESTO
DOVRÒ
DIRLE AD-
DIO.

PERCHÉ
MI FISSI
COSÌ?

COME MAI? È
PRESTO PER IL
DIPLOMA, NO?
NON SEI AL PRI-
MO ANNO?

LA NOSTRA SCUO-
LA ADOTTERÀ
DELLE UNIFORMI
NUOVE!



FACCIAVMO UNA TABELLA PER AVERE UNO SGUARDO D'INSIEME SUI DATI.

QUESTO ME LO RICORDO!

RISPOSTA	FREQUENZA	%
PIACE	28	70
NON SO	8	20
NON PIACE	4	10
SOMMA	40	100

QUESTA SI CHIAMA TABELLA DI CONTINGENZA.

PER LA CRONACA,
TU CHE COSA HAI
RISPOSTO
AL SONDAGGIO?

HO RISPOSTO "SI"....
MI PIACE!

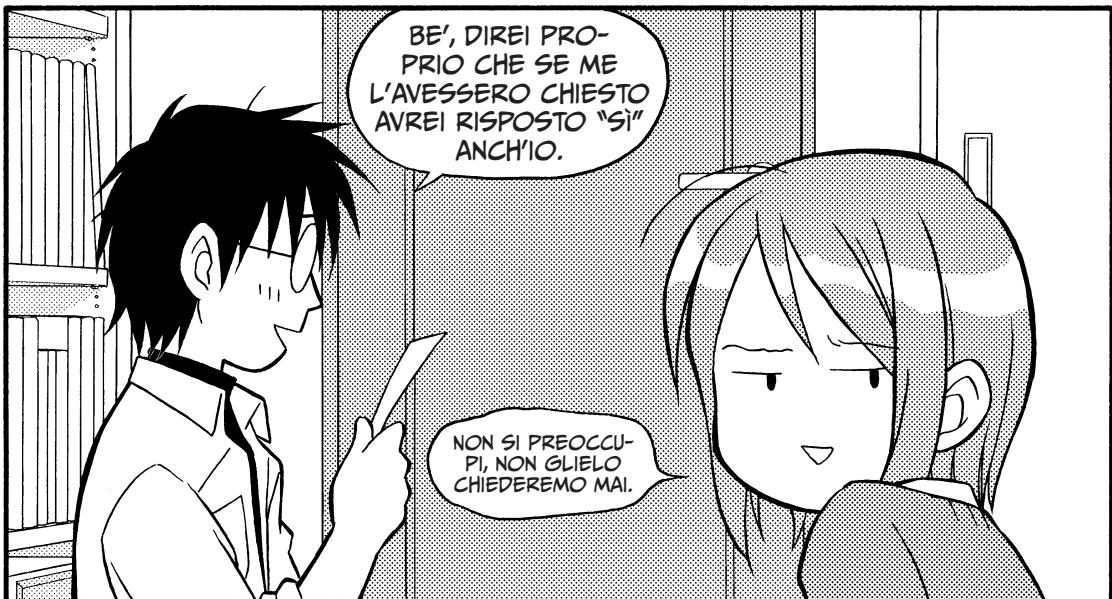
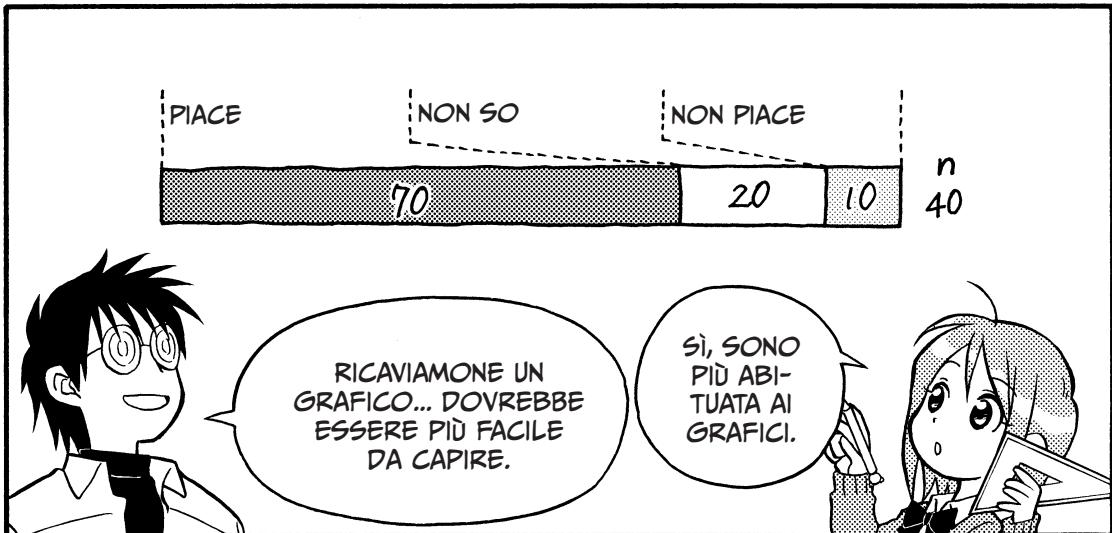
UN PO' DI RIPASSO.
QUAL È LA FREQUENZA DEI "SI"?

CI SONO 28 "SI", QUINDI LA FREQUENZA È QUESTA: 28.

E IN PERCENTUALE
ABBIAMO...

$$\frac{28}{40} \times 100 = \frac{7}{10} \times 100 = 70\%)$$

OK!



ESERCIZIO CON SOLUZIONE



ESERCIZIO

Un quotidiano ha condotto un sondaggio sul partito politico A, che spera di vincere le elezioni. Ecco i risultati:

Intervistato	Il Partito A vincerà contro il Partito B?
--------------	---

1	Perde
2	Perde
3	Perde
4	Non so
5	Vince
6	Perde
7	Vince
8	Non so
9	Perde
10	Perde

Compilare una tabella di contingenza dei dati.

SOLUZIONE

Ecco la tabella di contingenza.

Risposta	Frequenza	%
Vince	2	20
Non so	2	20
Perde	6	60
Somma	10	100

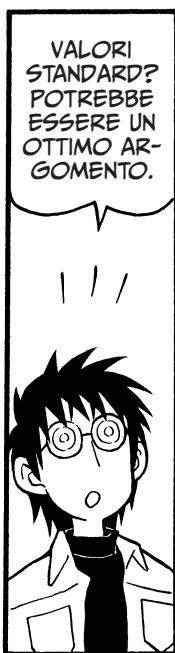
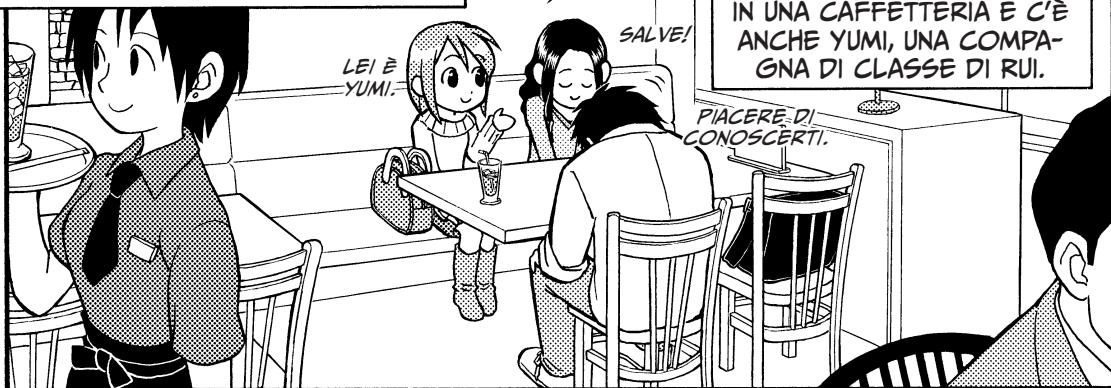
RIASSUMENDO

- La tabella di contingenza è un modo per ottenere uno sguardo d'insieme sui dati.

4

VALORI STANDARD E DI DEVIAZIONE

1. NORMALIZZAZIONE E VALORE STANDARD



BASSO
7A
MA CHISSÀ PERCHÉ IL VOTO DI YUMI IN GIAPPONESE VALE PIÙ DEL MIO IN INGLESE!

NEANCH'IO LO SO...

PERCHÉ?

PERCHÉ IN INGLESE L'INTERVALLO DEI PUNTEGGI E LA LORO MEDIA SONO DIVERSI DA QUELLI IN GIAPPONESE.

COSA?!

SE MI PASSI I VOTI DEI TUOI COMPAGNI DI CLASSE, CERCERÒ DI SPIEGARTELLO...

È IMPOSSIBILE.

YUMI,
COME...

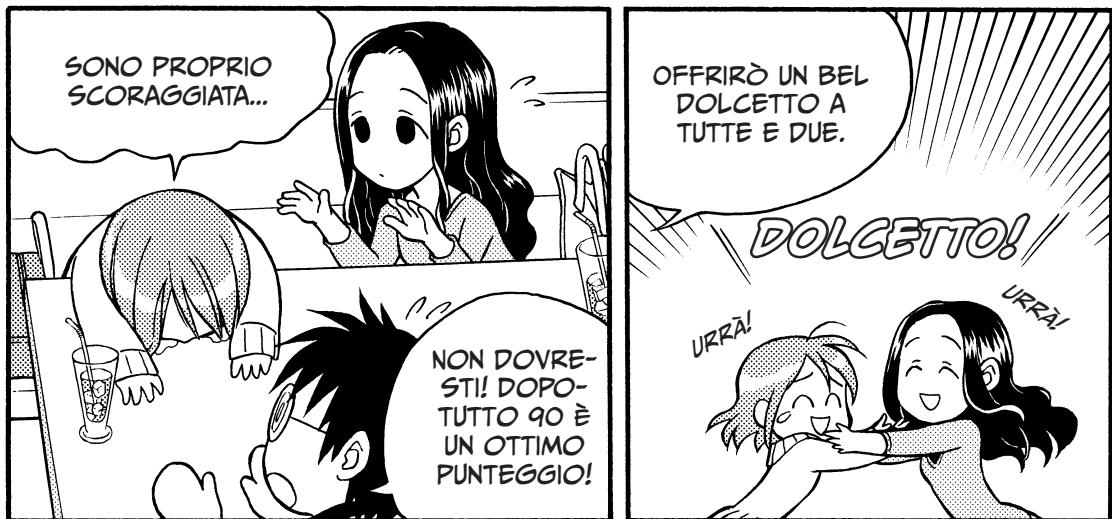
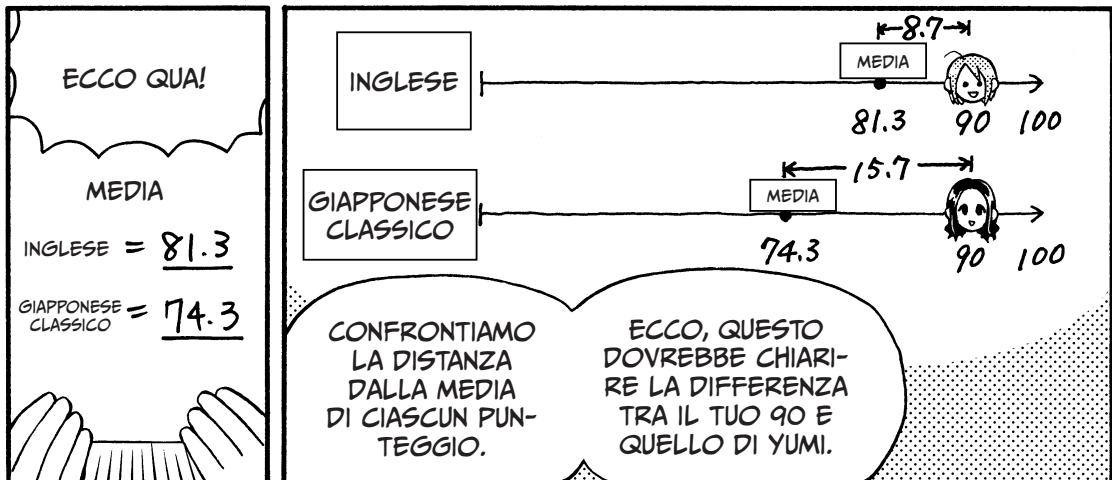
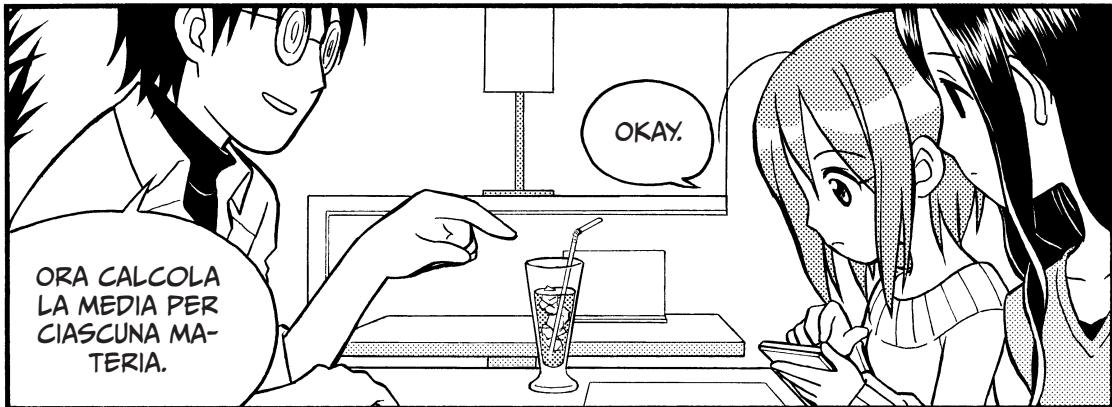
ECCOLI!

GRANDE!

PUNTEGGI ASSOLUTI (IN CENTESIMI)

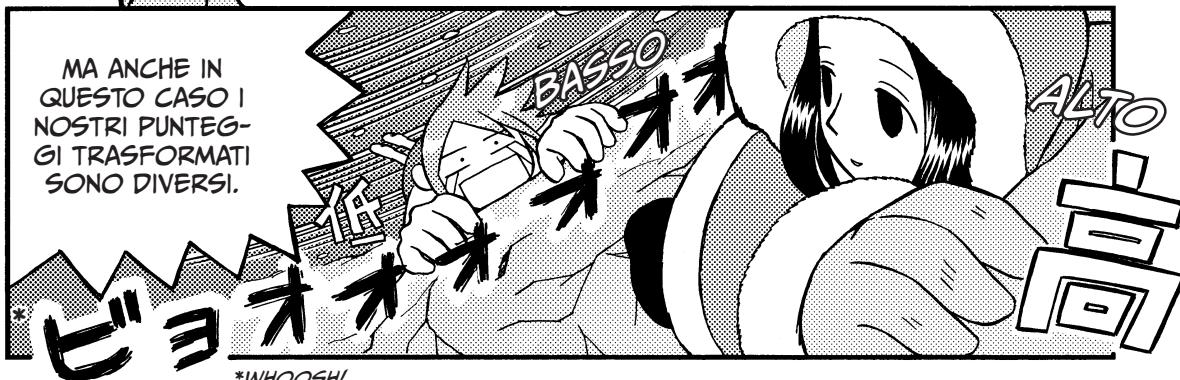


STUDENTE	INGLESE	GIAPPONESE CLASSICO	STUDENTE	INGLESE	GIAPPONESE CLASSICO
RUI	90	71	H	67	85
YUMI	81	90	I	87	93
A	73	79	J	78	89
B	97	70	K	85	78
C	85	67	L	96	74
D	60	66	M	77	65
E	74	60	N	100	78
F	64	83	O	92	53
G	72	57	P	86	80





LA MEDIA DELLE CLASSI DI STORIA E BIOLOGIA ERA 53.



ANCHE SE LA DISTANZA TRA I VOTI E LE MEDIE È LA STESSA!

MMM...

STUDENTE	STORIA	BIOLOGIA	STUDENTE	STORIA	BIOLOGIA
RUI	73	59	H	7	50
YUMI	61	73	I	53	41
A	14	47	J	100	62
B	41	38	K	57	44
C	49	63	L	45	26
D	87	56	M	56	91
E	69	15	N	34	35
F	65	53	O	37	53
G	36	80	P	70	68
MEDIA		53	MEDIA		53

QUAL È LA DEVIAZIONE STANDARD DI QUESTE MATERIE?

BE', LA DEVIAZIONE STANDARD È... UN INDICE CHE DESCRIVE IL "GRADO DI DISPERSIONE"...



SOMMA DI (CIASCUN VALORE - LA MEDIA)²

NUMERO DEI VALORI

...E LA FORMULA È...

DEVIASIONE STANDARD

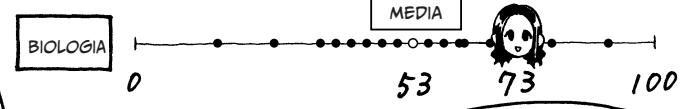
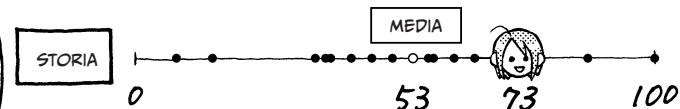
STORIA = 22.7

BIOLOGIA = 18.3

ECCO QUA!

MINORE È LA DEVIASIONE STANDARD E MINORE SARÀ L'“INTERVALLO DI DISPERSIONE DEI DATI”...

QUESTO VUOL DIRE CHE I VOTI IN BIOLOGIA DEI VOSTRI COMPAGNI SONO PIÙ SIMILI DI QUELLI IN STORIA.



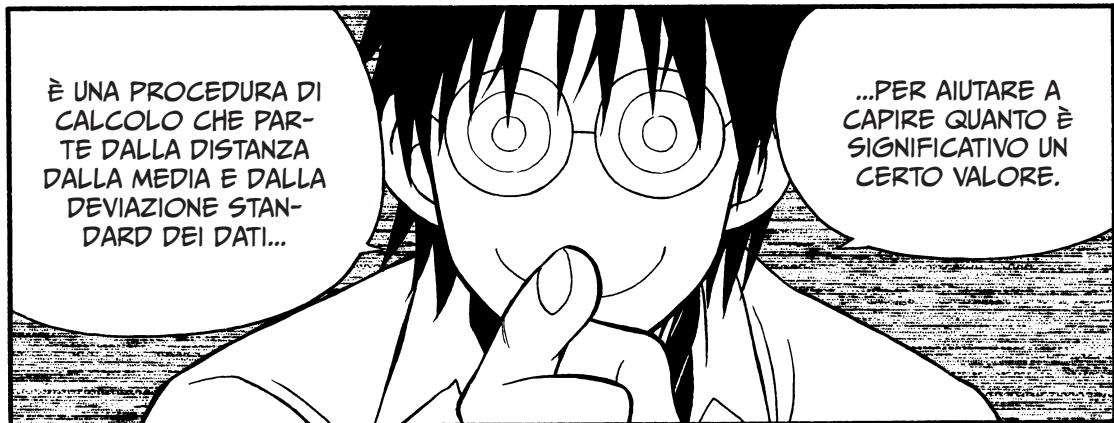
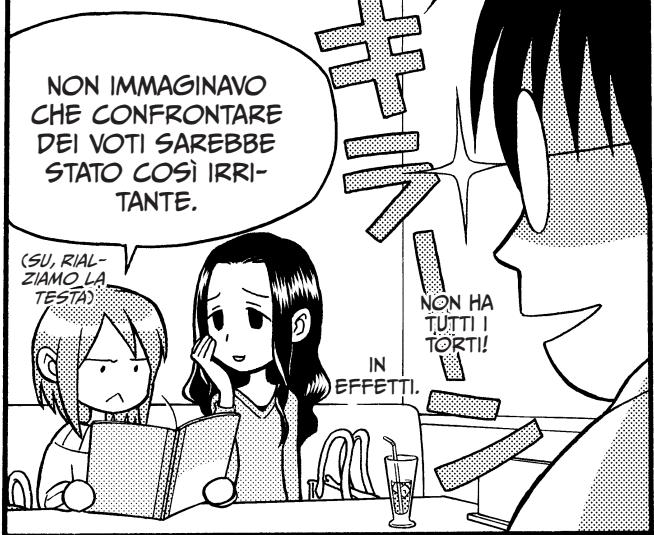
CHE COSA
VUOI DIRE?

SE FOSSI UNA MATRICOLA DI LICEO E VOLESSI POI ISCRIVERSI AL COLLEGE, MI CONCENTREREI SU BIOLOGIA.

PERCHÉ ANCHE SOLO UNO O DUE PUNTI POSSONO INCIDERE MOLTO SUL PIAZZAMENTO.

L'UNIFORME DEL
LICEO GLI STA PROPRIO BENE!

ARGH!



ECCO COME SI CALCOLA LA STANDARDIZZAZIONE. IL DATO STANDARDIZZATO VIENE DETTO VALORE STANDARD*.

(VALORE) - (MEDIA)

DEVIASIONE STANDARD

= VALORE STANDARD

POSSIAMO PENSARE AL VALORE STANDARD COME AL NUMERO DELLE DEVIAZIONI STANDARD IN CUI IL DATO ORIGINALE È AL DI SOPRA O AL DI SOTTO DELLA MEDIA. PER ESEMPIO, SE NORMALIZZANDO UNO DEI VOSTRI VOTI OTTENIAMO 1, QUESTO VUOL DIRE CHE IL VOTO È AL DI SOPRA DELLA MEDIA DELLA CLASSE DI UNA DEVIASIONE STANDARD (IN QUESTO CASO, 22,7)...

WOW!

*IL VALORE STANDARD VIENE ANCHE CHIAMATO "VALORE Z"

...E SE OTTENIAMO -1 VUOL DIRE CHE È UNA DEVIASIONE STANDARD AL DI SOTTO DELLA MEDIA. PROVIAMO AD APPLICARE TUTTO CIÒ AI RISULTATI DEI TEST DI PRIMA.

RICEVUTO!

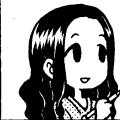
VOTI DI STORIA E BIOLOGIA ORIGINALI E NORMALIZZATI

STUDENTE	STORIA	BIOLOGIA	VALORI NORMALIZZATI DI STORIA	VALORI NORMALIZZATI DI BIOLOGIA
RUI	73	59	0.88	0.33
YUMI	61	73	0.35	1.09
A	14	47	-1.71	-0.33
B	41	38	-0.53	-0.82
C	49	63	-0.18	0.55
D	87	56	1.49	0.16
E	69	15	0.70	-2.08
F	65	53	0.53	0
G	36	80	-0.75	1.48
H	7	50	-2.02	-0.16
I	53	41	0	-0.66
J	100	62	2.07	0.49
K	57	44	0.18	-0.49
L	45	26	-0.35	-1.48
M	56	91	0.13	2.08
N	34	35	-0.84	-0.98
O	37	53	-0.70	0
P	70	68	0.75	0.82
MEDIA	53	53	0	0
DEVIASIONE STANDARD	22.7	18.3	1	1

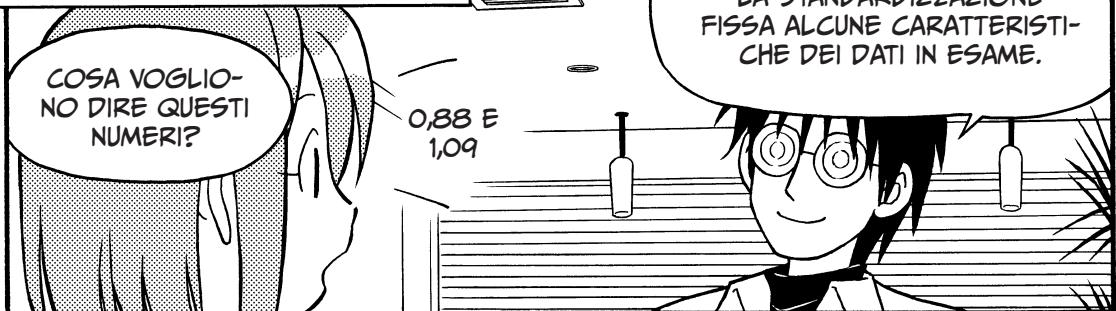
VOTO NORMALIZZATO DI RUI IN STORIA $\frac{73-53}{22.7} = \frac{20}{22.7} = 0.88$

VOTO NORMALIZZATO DI YUMI IN BIOLOGIA $\frac{73-53}{18.3} = \frac{20}{18.3} = 1.09$

QUINDI SONO QUESTI I VALORI!



2. CARATTERISTICHE DEI DATI NORMALIZZATI



(1) Qualunque sia il valore massimo assunto dalla variabile, la media aritmetica dei valori standard è sempre 0, e la deviazione standard è sempre 1.

POSSIAMO CONFRONTARE I RISULTATI DI DUE TEST DIVERSI, CON VALORE MASSIMO RISPETTIVAMENTE 100 E 200.

(2) Qualunque sia l'unità di misura della variabile in esame, la media aritmetica dei dati normalizzati è sempre 0, e la deviazione standard è sempre 1.

POSSIAMO CONFRONTARE VALORI RELATIVI A DIVERSE UNITÀ DI MISURA. SE PARLAMO DI BASEBALL, PER ESEMPIO, LA MEDIA DELLE BATTUTE È IL NUMERO DI HOME RUN.

DAI VALORI STANDARD 0,88 (STORIA) E 1,09 (BIOLOGIA) CAPIAMO QUALE ABBA IL MAGGIOR VALORE RISPETTO AGLI ALTRI RISULTATI RIPORTATI NEL TEST.

OKAY, È UFFICIALE: SONO UNA SCHIAPPA.



3. VALORE DI DEVIASIONE

PASSIAMO ORA ALLA NOZIONE DI DEVIASIONE. È SEMPLICEMENTE UNA MODIFICA DEL VALORE STANDARD: È "CENTRATA" ATTORNO A 50 INVECE CHE ATTORNO ALLO ZERO E HA UNA DEVIASIONE STANDARD DI 10 INVECE CHE DI 1.

OH!

ECCO LA FORMULA.

$$\text{DEVIASIONE} = \text{VALORE STANDARD} \times 10 + 50$$

IN EFFETTI NELLA FORMULA COMPARTE LA STANDARDIZZAZIONE.

ECCO I VOSTRI VALORI DI DEVIASIONE.

RUI
(STORIA)

YUMI
(BIOLOGIA)

$$0.88 \times 10 + 50 = 8.8 + 50 = 58.8$$

$$1.09 \times 10 + 50 = 10.9 + 50 = 60.9$$

SONO ESATTAMENTE QUELLI CHE CI AVEVANO DETTO!

RIASSUMIAMO UN PO' DI FATTI INTERESSANTI.

VALORE STANDARD

(1) Qualunque sia il valore massimo assunto dalla variabile, la media aritmetica dei valori standard è sempre 0, e la deviazione standard è sempre 1.

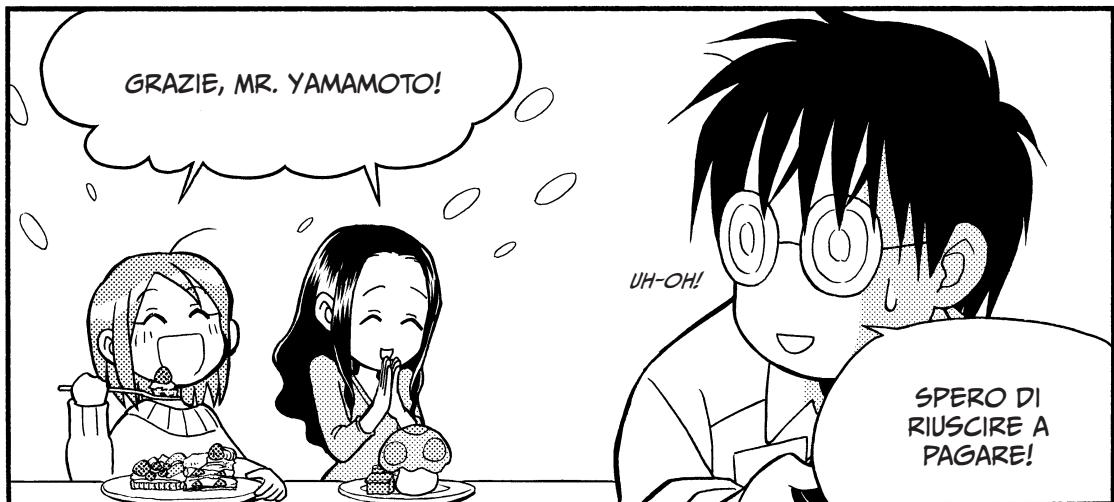
(2) Qualunque sia l'unità di misura della variabile in esame, la media aritmetica dei dati normalizzati è sempre 0, e la deviazione standard è sempre 1.



VALORE DI DEVIASIONE

(1) Qualunque sia il valore massimo assunto dalla variabile, la media aritmetica dei valori di deviazione è sempre 50, e la deviazione standard è sempre 10.

(2) Qualunque sia l'unità di misura della variabile in esame, la media aritmetica dei valori di deviazione è sempre 50, e la deviazione standard è sempre 10.



4. INTERPRETAZIONE DEL VALORE DI DEVIAZIONE



L'interpretazione del valore di deviazione è particolarmente delicata. Come abbiamo visto a pagina 74 la definizione è la seguente:

$$\text{valore di deviazione} = \text{valore standard} \times 10 + 50 = \frac{(\text{valore} - \text{media})}{\text{deviazione standard}} \times 10 + 50$$

Abbiamo visto a pagina 62 che nella classe di Rui ci sono 40 studenti e a pagina 40 che 18 di questi sono ragazze. L'esempio della deviazione a pagina 69 non riguarda l'intera classe ma solo le ragazze. Se riguardasse tutti la media e la deviazione standard sarebbero diverse da quelle per le sole ragazze. Di fatto, considerando tutta la classe, il massimo valore di deviazione è quello di Rui. La tabella 4.1 riporta i risultati dell'intera classe: provate a calcolare le deviazioni.

Vi anticipiamo i risultati: quella di Rui in storia è 59,1 e quella di Yumi in biologia è 56,7.

Supponiamo che gli studenti delle classi 1 e 2 svolgano la medesima prova. La media e la deviazione standard della classe 1 vengono calcolate individualmente, per poi ricavare i valori di deviazione a partire da questi importi. Analogamente per quanto riguarda la media, la deviazione standard e le deviazioni per la classe 2. Sia lo studente A della classe 1 che lo studente B della classe 2 hanno un valore di deviazione di 57. Visti dall'esterno, questo sembra indicare che hanno la stessa preparazione. Ma la media e la deviazione standard utilizzate nei calcoli sono relative a due classi diverse, e sono quindi diverse. In altre parole, non possiamo confrontare il valore di deviazione dei due studenti, a meno che media e deviazione standard siano uguali.

Vediamo un altro esempio. Supponiamo che lo studente A sostenga un esame di ammissione in aprile ottenendo un valore di deviazione di 54. Dopo avere studiato duramente durante uno speciale corso estivo, a settembre sostiene un altro esame d'ammissione in un'altra scuola, con un valore di deviazione di 62. Apparentemente la sua prestazione è migliorata, ma in realtà l'esame di aprile e gli studenti che l'hanno sostenuto sono diversi dall'esame e dagli studenti di settembre. Non possiamo quindi confrontare i due valori di deviazione, perché i dati usati per calcolare media e deviazione standard nelle due situazioni sono diversi. In breve: nella valutazione dei risultati di un esame, possiamo confrontare solo valori di deviazione del gruppo di studenti che ha partecipato a quello stesso esame. Non dimenticate mai queste considerazioni quando cercate di interpretare un valore di deviazione.

RISULTATI DEI TEST DI STORIA E BIOLOGIA (TUTTI STUDENTI DELLA CLASSE DI RUI)

Ragazze	Storia	Biologia	Ragazzi	Storia	Biologia
Rui	73	59	a	54	2
Yumi	61	73	b	93	7
A	14	47	c	91	98
B	41	38	d	37	85
C	49	63	e	44	100
D	87	56	f	16	29
E	69	15	g	12	57
F	65	53	h	44	37
G	36	80	i	4	95
H	7	50	j	17	39
I	53	41	k	66	70
J	100	62	l	53	14
K	57	44	m	14	97
L	45	26	n	73	39
M	56	91	o	6	75
N	34	35	p	22	80
O	37	53	q	69	77
P	70	68	r	95	14
			s	16	24
			t	37	91
			u	14	36
			v	88	76
Media dell'intera classe				48.0	54.9
Deviazione standard dell'intera classe				27.5	26.9

ESERCIZIO CON SOLUZIONE



ESERCIZIO

Ecco i tempi realizzati dalle ragazze del liceo sui 100 metri piani:

Corridore	Tempo sui 100 m piani (secondi)
Ms. A	16.3
Ms. B	22.4
Ms. C	18.5
Ms. D	18.7
Ms. E	20.1
Media	19.2
Deviazione standard	2.01

1. Dimostrare che la media dei tempi normalizzati è 0.
2. Dimostrare che la deviazione standard dei tempi normalizzati è 1.

SOLUZIONE

1. Media dei valori standard dei tempi sui 100 metri:

$$\begin{aligned}
 &= \frac{\left(\frac{16.3 - 19.2}{2.01}\right) + \left(\frac{22.4 - 19.2}{2.01}\right) + \left(\frac{18.5 - 19.2}{2.01}\right) + \left(\frac{18.7 - 19.2}{2.01}\right) + \left(\frac{20.1 - 19.2}{2.01}\right)}{5} \\
 &= \frac{\left\{ \frac{(16.3 - 19.2) + (22.4 - 19.2) + (18.5 - 19.2) + (18.7 - 19.2) + (20.1 - 19.2)}{2.01} \right\}}{5} \\
 &= \frac{\left\{ \frac{16.3 + 22.4 + 18.5 + 18.7 + 20.1 - 19.2 - 19.2 - 19.2 - 19.2 - 19.2}{2.01} \right\}}{5} \\
 &= \frac{\left\{ \frac{96 - 19.2 \times 5}{2.01} \right\}}{5} \\
 &= \frac{\left\{ \frac{96 - 96}{2.01} \right\}}{5} \\
 &= \frac{0}{5} \\
 &= 0
 \end{aligned}$$

passiamo al denominatore comune.

Riscriviamo il numeratore evidenziando valori e medie (19,2).

2. Deviazione standard dei valori standard (o normalizzati):

$$\begin{aligned}
 &= \sqrt{\frac{\left(\frac{16.3 - 19.2}{2.01} - 0\right)^2 + \left(\frac{22.4 - 19.2}{2.01} - 0\right)^2 + \left(\frac{18.5 - 19.2}{2.01} - 0\right)^2 + \left(\frac{18.7 - 19.2}{2.01} - 0\right)^2 + \left(\frac{20.1 - 19.2}{2.01} - 0\right)^2}{5}} \\
 &= \sqrt{\frac{\left(\frac{16.3 - 19.2}{2.01}\right)^2 + \left(\frac{22.4 - 19.2}{2.01}\right)^2 + \left(\frac{18.5 - 19.2}{2.01}\right)^2 + \left(\frac{18.7 - 19.2}{2.01}\right)^2 + \left(\frac{20.1 - 19.2}{2.01}\right)^2}{5}} \\
 &= \sqrt{\frac{\left\{ \frac{(16.3 - 19.2)^2 + (22.4 - 19.2)^2 + (18.5 - 19.2)^2 + (18.7 - 19.2)^2 + (20.1 - 19.2)^2}{2.01^2} \right\}}{5}} \\
 &= \sqrt{\frac{\frac{1}{2.01^2} \times (16.3 - 19.2)^2 + (22.4 - 19.2)^2 + (18.5 - 19.2)^2 + (18.7 - 19.2)^2 + (20.1 - 19.2)^2}{5}} \\
 &= \frac{1}{2.01} \times \sqrt{\frac{(16.3 - 19.2)^2 + (22.4 - 19.2)^2 + (18.5 - 19.2)^2 + (18.7 - 19.2)^2 + (20.1 - 19.2)^2}{5}} \\
 &= \frac{1}{\text{deviazione standard dei tempi sui 100 metri}} \times \text{deviazione standard dei tempi sui 100 metri} \\
 &= 1
 \end{aligned}$$

il numeratore viene chiarito

il numeratore viene chiarito

Verificate attentamente nella tabella di pagina 78.

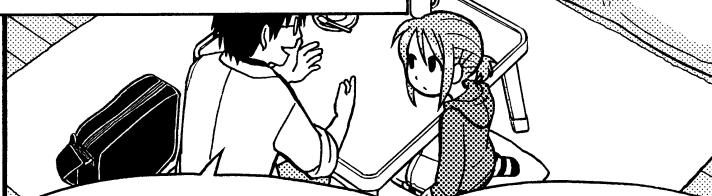
RIASSUMENDO

- La standardizzazione aiuta a interpretare un dato relativamente a tutti gli altri valutando la sua distanza dalla media e la cosiddetta "ampiezza della dispersione" dei dati.
- Potete ricorrere alla normalizzazione per:
 - Confrontare variabili in intervalli diversi.
 - Confrontare variabili espresse in unità di misura diverse.
- Un dato standardizzato viene detto valore standard dell'osservazione. Il valore di deviazione è un'applicazione del valore standard.

5

TROVIAMO LE PROBABILITÀ

1. FUNZIONE DI DENSITÀ DI PROBABILITÀ



IN STATISTICA SI USA SPESO IL TERMINE PROBABILITÀ.

IN ESPRESSIONI COME "LA PROBABILITÀ CHE QUESTO SIA PIÙ GRANDE DI QUELLO ECCETERA ECCETERA".

OGGI TI SPIEGHERÒ CHE COSA CI SERVE SAPERE PER RICAVARE LA PROBABILITÀ "CHE QUESTO SIA PIÙ GRANDE DI QUELLO".

UH?



MI SCUSI! LA PROBABILITÀ CHE DICE LEI È LA STESSA DI CUI SI PARLA NELLE PREVISIONI ATMOSFERICHE?

L'ARGOMENTO DI OGGI È LEGGERMENTE ASTRATTO.

MA CERCA DI STUDIARLO BENE. QUELLO CHE IMPARERAI OGGI TORNERÀ UTILE IN MOLTE AREE DELLA STATISTICA.



PRECISAMENTE.



RUI



RISULTATI DEL TEST DI INGLESE DELLE
MATRICOLE LICEALI DELLA PREFETTURA A

STUDENTE	PUNTEGGIO
1	42
2	91
...	...
10,421	50
MEDIA	53
DEVIAZIONE STANDARD	10

SUPPO-
NIAMO

CHE TUTTE
LE MATRICO-
LE DEI LICEI
DELLA PRE-
FETTURA A...

...SVOGLANO
LA PROVA DI
INGLESE.

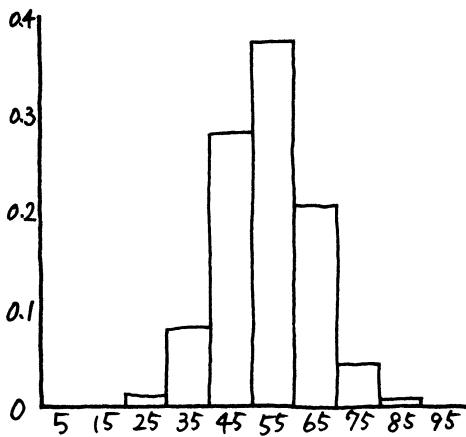
EHI, LA
VEDO
PREPARATA
OGGI.

AH, AH, AH!
QUESTO È
SOLO L'INIZIO.

フフフフフフ...

QUESTO È UN
ISTOGRAMMA RICAVA-
TO DALLA TABELLA...
SULL'ASSE DELLE Y AB-
BIAMO LA PERCENTUALE
DI STUDENTI IN CIASCUNA
CLASSE DI PUNTEGGIO.

ISTOGRAMMA DEI RISULTATI DEL TEST DI INGLESE
(CAMPIEZZA DI CLASSE = 10)



TRASFORMAN-
DO LE TABELLE
IN ISTATOGRAMMI
TUTTO SI CA-
PISCE MOLTO
MEGLIO.

CERTO,
L'APPROC-
CIO VISIVO
È PIÙ IMME-
DIATO.

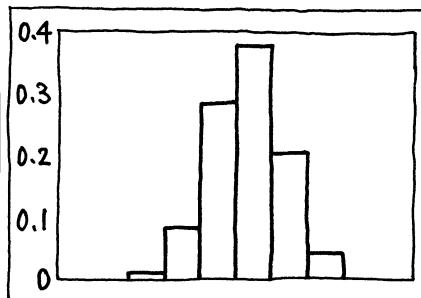
PROVIAMO A IMMAGI-
NARE COSA SUCCIDE
SE RIDUCIAMO L'AM-
PIEZZA DI CLASSE
DELL'ISTOGRAMMA.

CIOÈ?

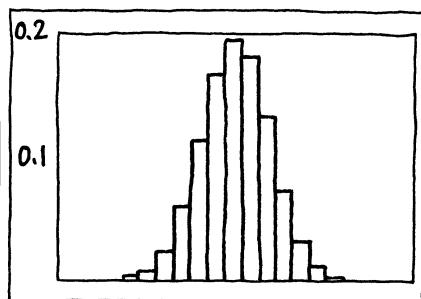
così...

AMPIEZZA DI CLASSE E ISTOGRAMMI
DELLA PROVA DI INGLESE

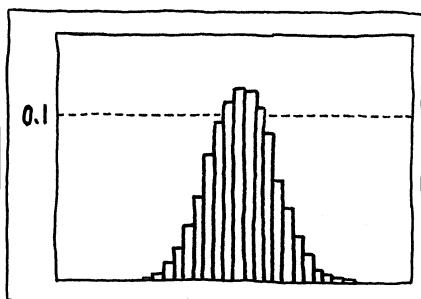
AMPIEZZA=10



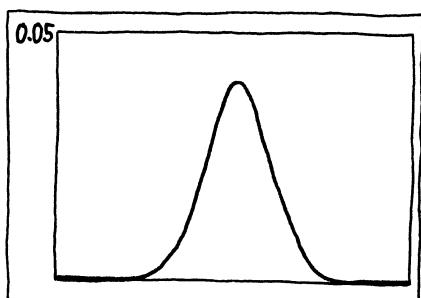
AMPIEZZA=5



AMPIEZZA=3



CURVA



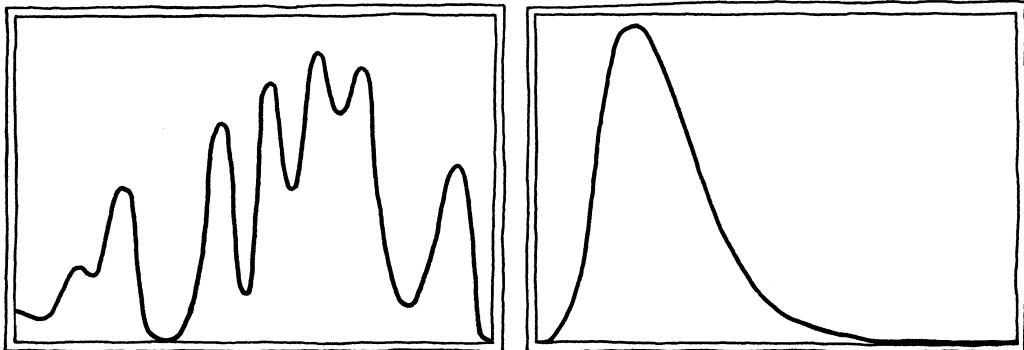
WOW! ALLA FINE DIVENTA UNA LINEA CURVA!



QUANDO FACCIAMO
TENDERE A ZERO L'AM-
PIEZZA DI CLASSE DI UN
ISTOGRAMMA...

...CHIAMIAMO DENSITÀ DI
PROBABILITÀ LA FUNZIO-
NE DI CUI OTTENIAMO IL
GRAFICO.

FUNZIONE DENSITÀ DI PROBABILITÀ



IN TEORIA...

...PER LA DENSITÀ DI
PROBABILITÀ ESISTO-
NO MOLTI POSSIBILI
GRAFICI.

OGGI VEDREMO
ALCUNI DEI PIÙ
IMPORTANTI.

SÌ, PER
FAVORE!

2. DISTRIBUZIONE NORMALE

$$f(x) = \frac{1}{(\text{deviazione standard di } x)\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x - \text{media di } x}{\text{deviazione standard di } x}\right)^2}$$

GUARDA QUA.

E QUESTA ROBA COSA SAREBBE?!

È UNA FUNZIONE DENSITÀ DI PROBABILITÀ MOLTO USATA IN STATISTICA.

COSA DIAVOLO È LA "E" IN CORSIVO?

SI CHIAMA NUMERO DI EULERO E VALE CIRCA 2,71828...*

*E VIENE ANCHE CHIAMATA NUMERO DI NEPERO.

AH!
AH!
AH!

IMMAGINA CHE SIA QUALCOSA TIPO PI GRECO.

OH, CHE BELLO...

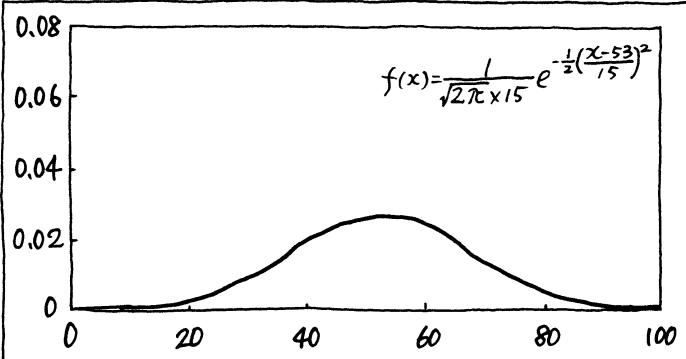
FIN QUI POSSO ARRIVARCI...

IL GRAFICO DELLA DISTRIBUZIONE DI PROBABILITÀ HA DUE CARATTERISTICHE.

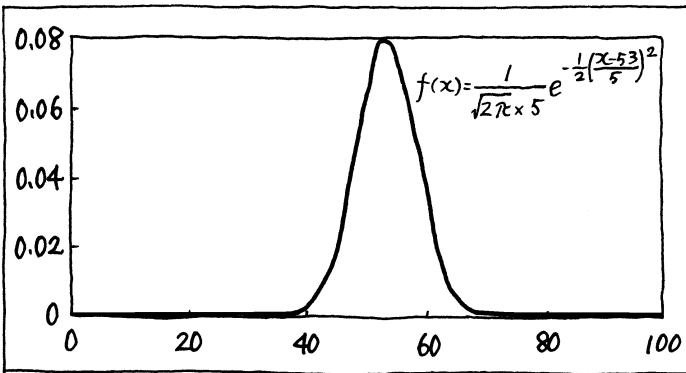
È SIMMETRICO RISPETTO ALLA MEDIA.

DIPENDE DALLA MEDIA E DALLA DEVIAZIONE STANDARD.

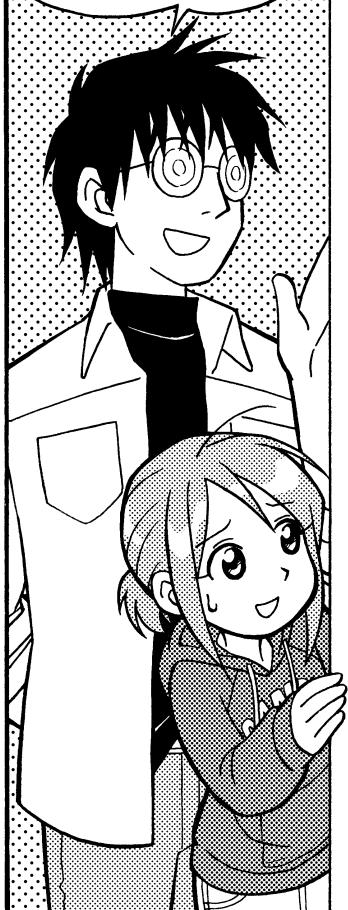
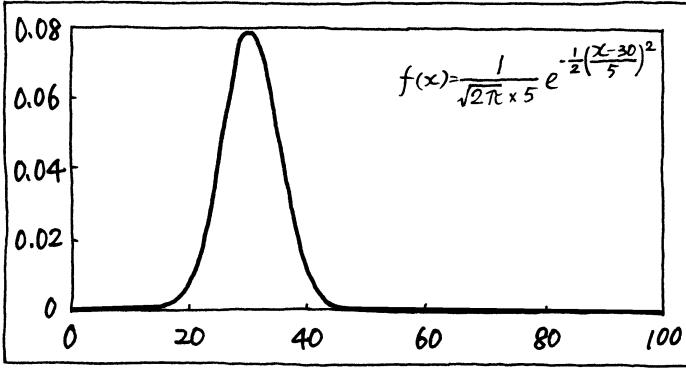
MEDIA=35, DEVIAZIONE STANDARD=15



MEDIA=53, DEVIAZIONE STANDARD=5



MEDIA=30, DEVIAZIONE STANDARD=5



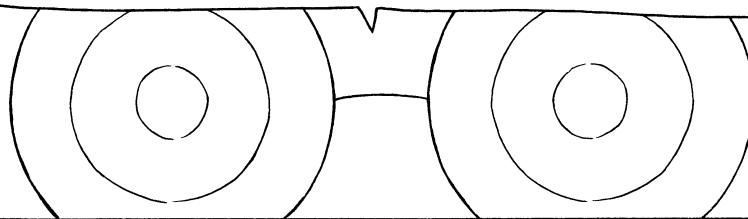
IN STATISTICA TUTTO CIÒ VIENE FORMULATO IN UN MODO PRECISO...



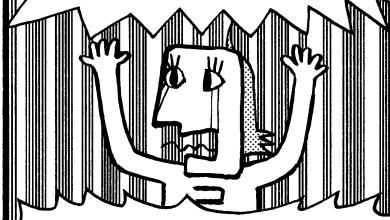
SE LA FORMULA PER LA DENSITÀ DI PROBABILITÀ È

$$f(x) = \frac{1}{(\text{deviazione standard di } x)\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x - \text{media di } x}{\text{deviazione standard di } x} \right)^2}$$

DICIAMO CHE "X SEGUE UNA DISTRIBUZIONE NORMALE CON MEDIA μ E DEVIAZIONE STANDARD σ ".

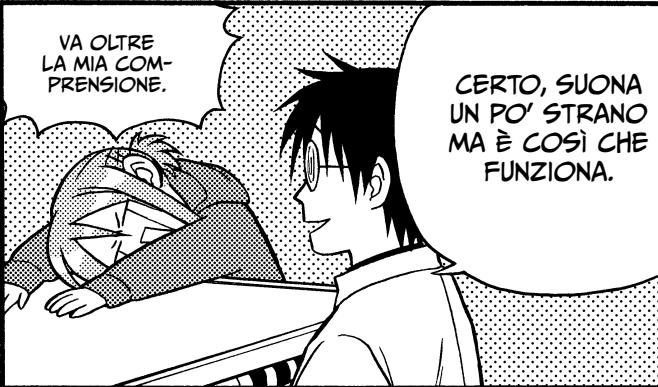


QUESTA FRASE È TROPPO COMPLICATA!



"X SEGUE QUESTO E POI QUELLO E POI QUELL'ALTRO"...?!"

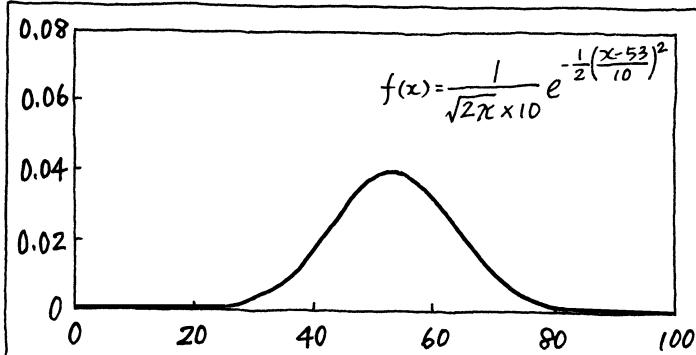
VA OLTRE LA MIA COMPRENSIONE.



TORNIAMO ORA ALLA PROVA D'INGLESE.

SE LA DENSITÀ DI PROBABILITÀ DEI RISULTATI DEL TEST DI INGLESE È QUALCOSA DEL GENERE...

DISTRIBUZIONE NORMALE CON MEDIA 53 E DEVIAZIONE STANDARD 10



POSSIAMO DIRE CHE I RISULTATI DEL TEST DI INGLESE SEGUONO UNA DISTRIBUZIONE NORMALE CON MEDIA 53 E DEVIAZIONE STANDARD 10.

FORSE COMINCIÒ A CAPIRE!

3. DISTRIBUZIONE NORMALE STANDARD

PASSIAMO AL PROSSIMO ARGOMENTO.

AGLI ORDINI!

SE LA FORMULA PER LA DENSITÀ DI PROBABILITÀ DI X È

$$f(x) = \frac{1}{(\text{deviazione standard di } x)\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x - \text{media di } x}{\text{deviazione standard di } x}\right)^2}$$
$$= \frac{1}{1 \times \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-0}{1}\right)^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

NON DIREMO CHE "X SEGUE UNA DISTRIBUZIONE NORMALE CON MEDIA ZERO E DEVIAZIONE STANDARD 1": IN STATISTICA SI PARLA DI DISTRIBUZIONE NORMALE STANDARD, O ANCHE STANDARDIZZATA.

...?!

COME ESEMPIO, USIAMO NUOVAMENTE LA PROVA DI INGLESE.

SUPPONIAMO CHE I RISULTATI SEGUANO LA DISTRIBUZIONE NORMALE CON MEDIA 53 E DEVIAZIONE STANDARD 10.

53

OKAY.

STUDENTE	PUNTEGGIO
1	42
2	91
:	:
10,421	50
MEDIA	53
DEVIAZIONE STANDARD	10

VALORE Z DEI RISULTATI

-1.1

3.8

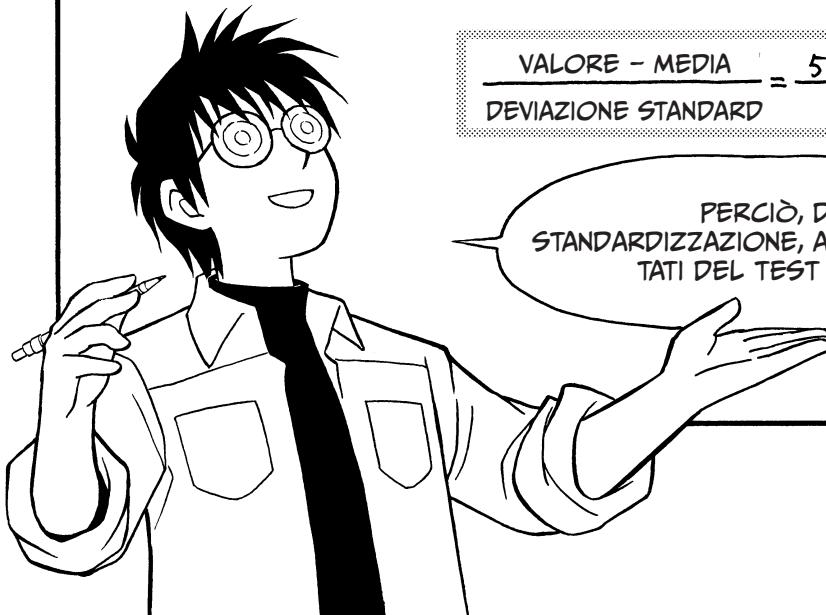
-0.3

0

1

$$\frac{\text{VALORE} - \text{MEDIA}}{\text{DEVIAZIONE STANDARD}} = \frac{50 - 53}{10} = \frac{-3}{10} = -0.3$$

PERCIÒ, DOPO LA STANDARDIZZAZIONE, ABBIAMO CHE I RISULTATI DEL TEST DI INGLESE...



DISTRIBUZIONE NORMALE STANDARD

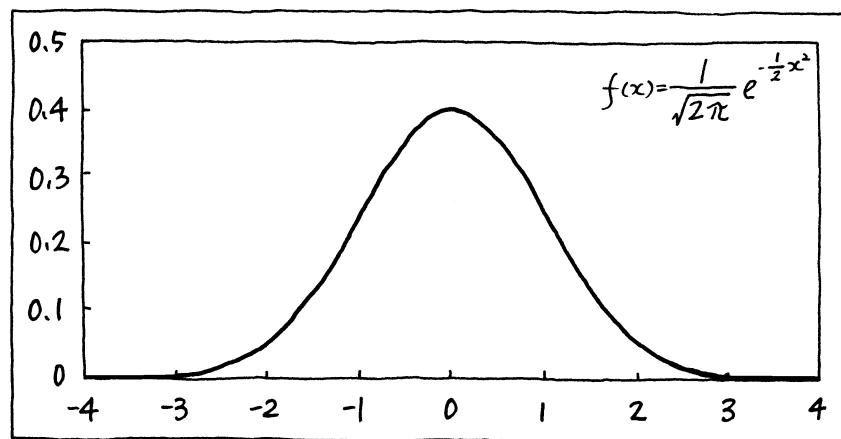
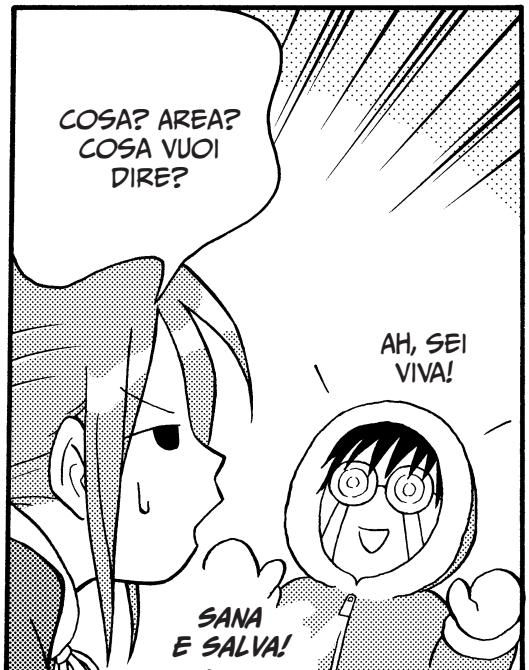
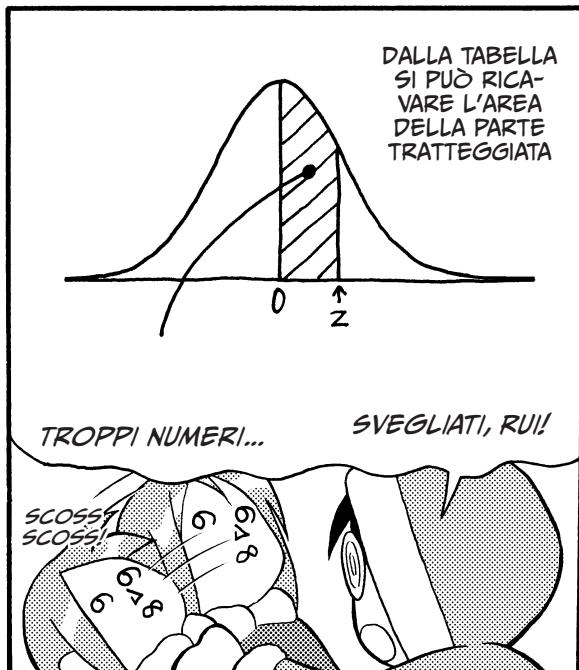
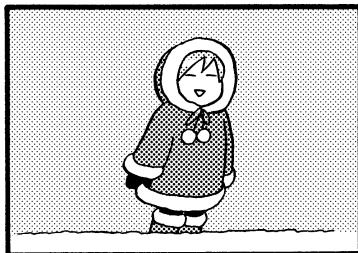
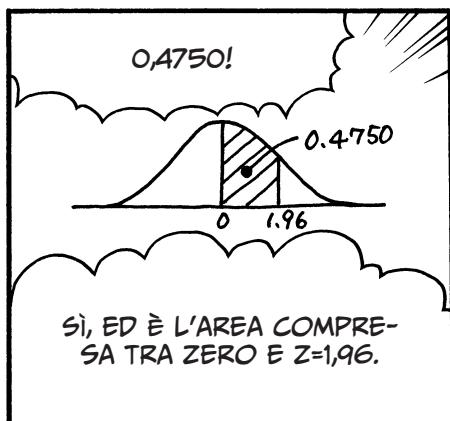
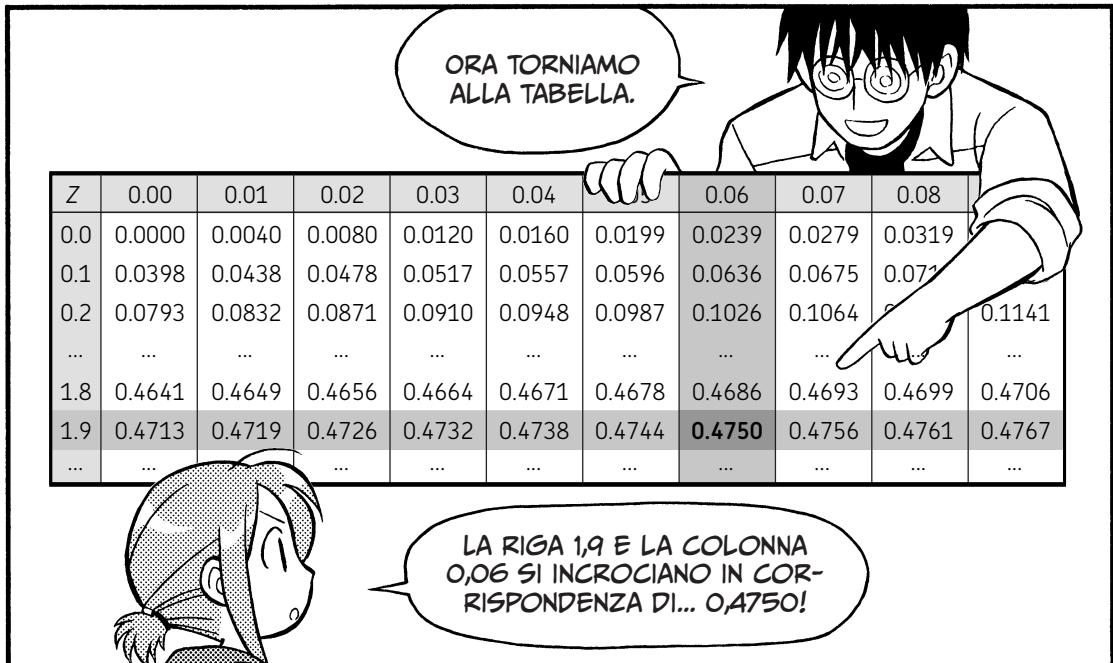
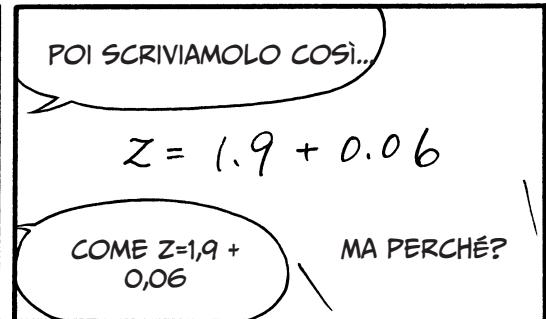


TABELLA DELLA DISTRIBUZIONE NORMALE STANDARD

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
...
0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706	
0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767	
...	





E ORA FAI ATTENZIONE, PERCHÉ
STO PER SPIEGARE IL PIATTO
DEL GIORNO.

NON VEDO
L'ORA DI
MANGIARLO!

L'AREA DELIMITATA DALLA DISTRIBUZIONE
NORMALE STANDARD E L'ASSE ORIZ-
ZONTALE È LA PROBABILITÀ!

CO...?

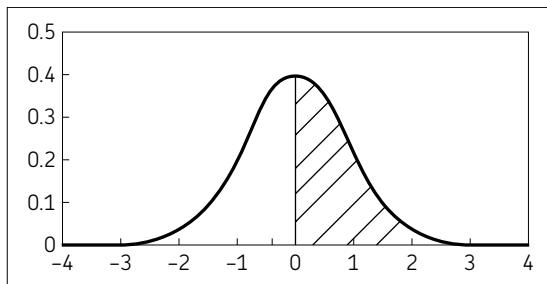
CERCHERÒ DI MOSTRARTELLO CON DUE ESEMPI...
CERCA DI SEGUIRMI!

ESEMPIO 1

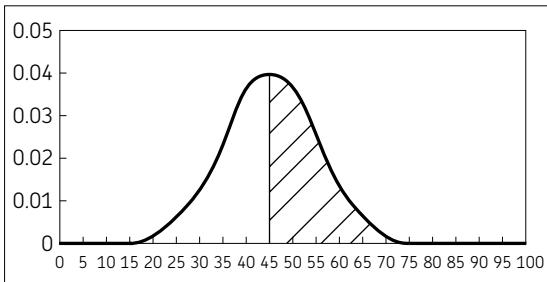


Tutte le matricole dei licei della Prefettura B svolgono la prova di matematica. I risultati seguono una distribuzione normale con media 45 e deviazione standard 10. E ora, attenta, perché le cinque affermazioni che seguono vogliono dire la stessa cosa.

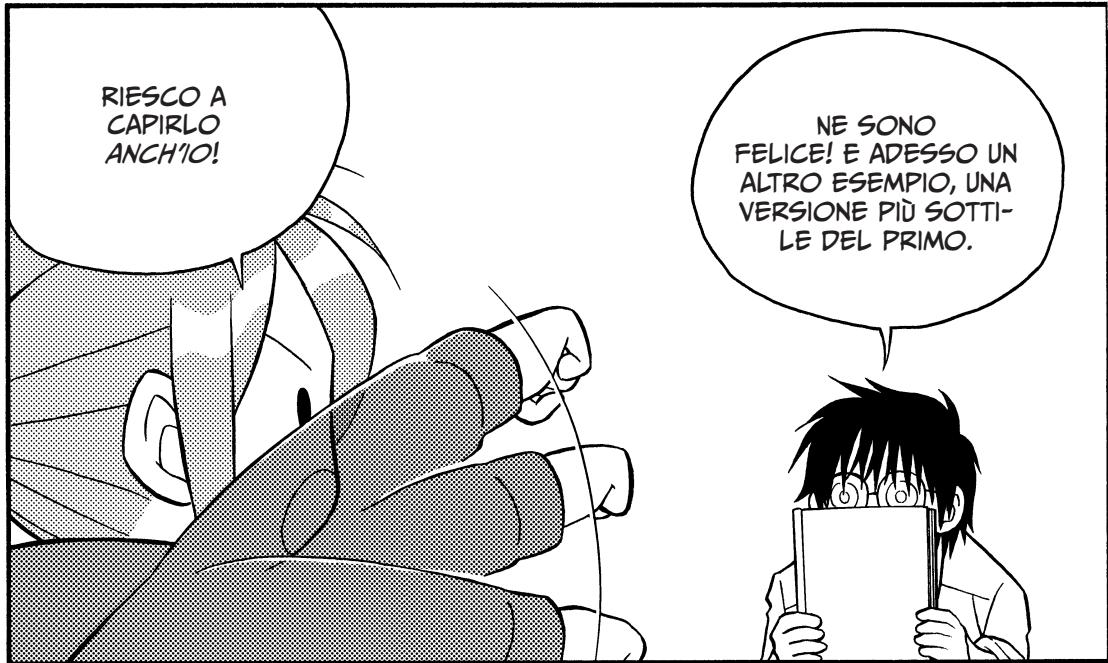
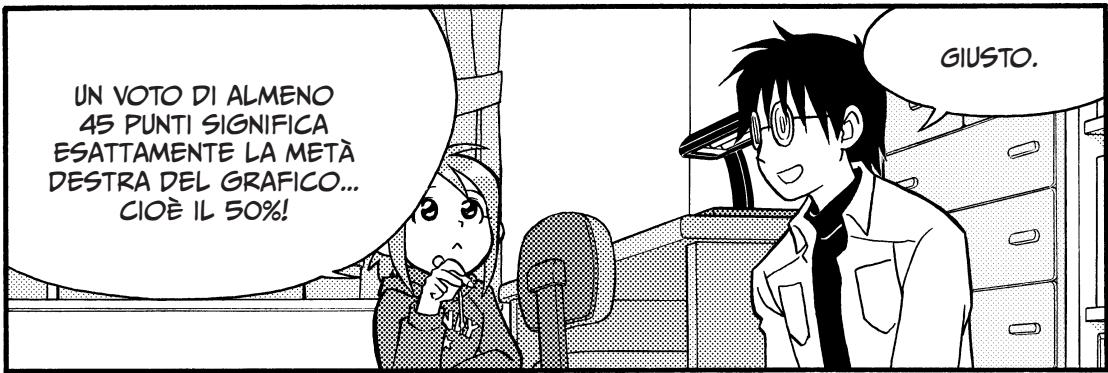
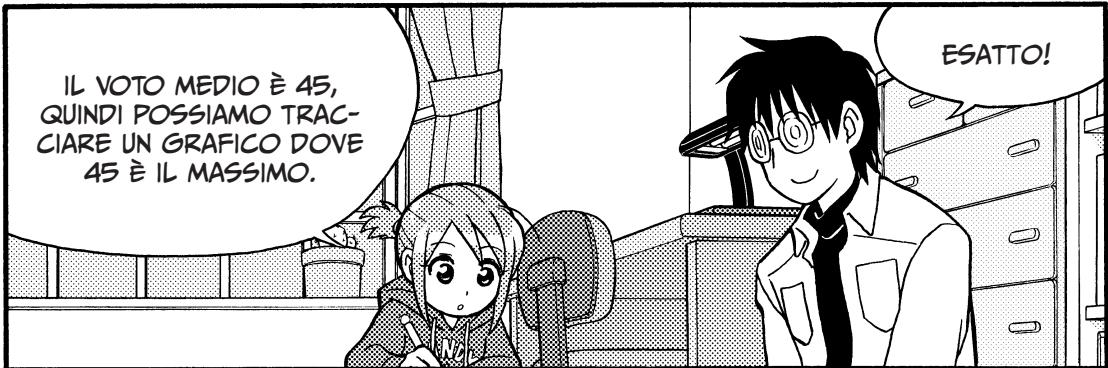
1. In una distribuzione normale di media 45 e deviazione standard 10, l'area tratteggiata sotto il grafico è 0,5.



2. La percentuale di studenti che ha ottenuto almeno 45 è 0,5 (il 50% di tutti gli studenti).
3. Scegliendo a caso uno studente tra tutti coloro che hanno sostenuto la prova, la probabilità che il suo voto sia almeno 45 è 0,5 (il 50%).
4. In una distribuzione normale di "risultati di matematica standardizzati" la percentuale di studenti con valore standard di almeno 0 è 0,5 il 50% di tutti gli studenti).



5. Scegliendo a caso uno studente tra tutti coloro che hanno sostenuto la prova, in una distribuzione normale di "risultati di matematica standardizzati" la probabilità che il suo valore standard sia almeno 0 è 0,5 (il 50%).

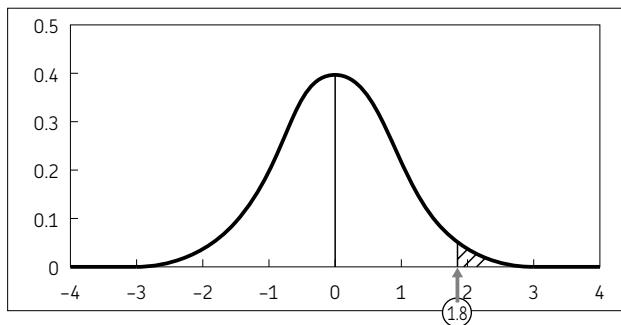


ESEMPIO II

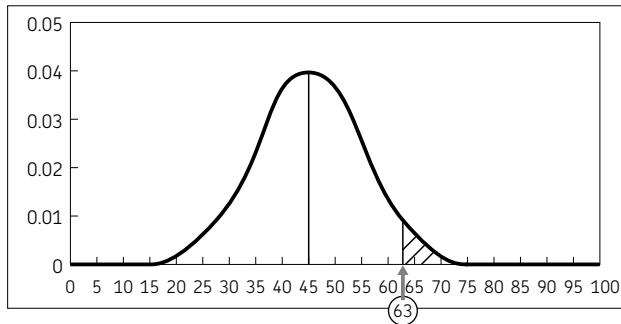
Tutte le matricole dei licei della Prefettura B hanno fatto il test di matematica.
Attenzione ora: le cinque affermazioni seguenti sono equivalenti.



1. In una distribuzione normale di media 45 e deviazione standard 10, l'area tratteggiata sotto il grafico è $0,5 - 0,4641 = 0,0359$.



2. La percentuale di studenti che ha ottenuto almeno 63 è $0,5 - 0,4641 = 0,0359$ (il 3,59% di tutti gli studenti).
3. Scegliendo a caso uno studente tra tutti coloro che hanno sostenuto la prova, la probabilità che il suo voto sia almeno 63 è $0,5 - 0,4641 = 0,0359$ (il 3,59%).
4. In una distribuzione normale di risultati standardizzati la percentuale di studenti con valore standard (o valore z) di almeno 1,8 [(valore - media) / deviazione standard = $(63 - 45) / 10 = 18 / 10 = 1,8$] è 3,59% ($0,5 - 0,4641 = 0,0359$). Il valore si può ricavare anche da una tabella della distribuzione normale standard.

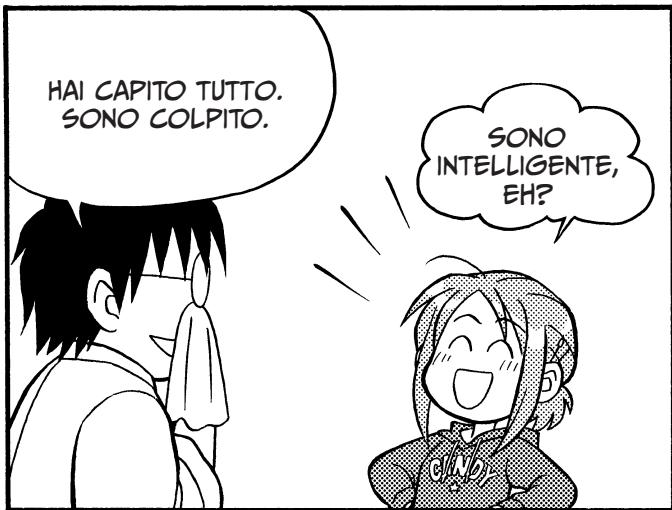


5. Scegliendo a caso uno studente tra tutti coloro che hanno sostenuto la prova, in una distribuzione normale di "risultati di matematica standardizzati" la probabilità che il suo valore standard sia almeno 1,8 è $0,5 - 0,4641 = 0,0359$ (il 3,59%).

WOW! QUINDI AREA, PERCENTUALE E PROBABILITÀ SONO LA STESSA COSA.

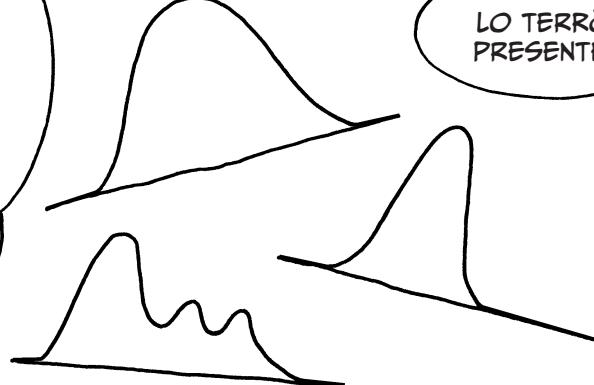


HAI CAPITO TUTTO.
SONO COLPITO.



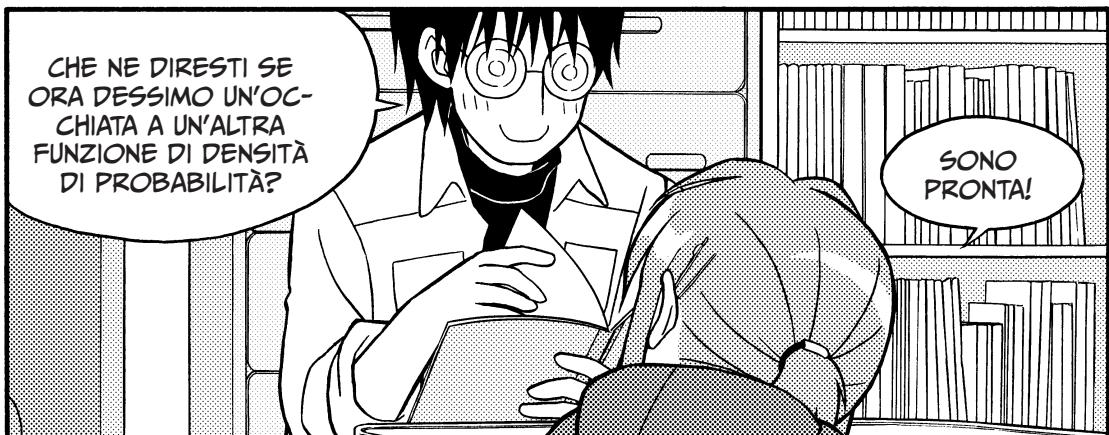
E NON DIMENTICARE
CHE LA RELAZIONE
AREA = PERCENTUALE =
PROBABILITÀ VALE PER
QUALSIASI DENSITÀ DI
PROBABILITÀ, NON SOLO
PER LA DISTRIBUZIONE
NORMALE STANDARD.

LO TERRÒ
PRESENTE.



CHE NE DIRESTI SE
ORA DESSIMO UN'OCCHIATA
A UN'ALTRA
FUNZIONE DI DENSITÀ
DI PROBABILITÀ?

SONO
PRONTA!



4. DISTRIBUZIONE CHI-QUADRO.



QUANDO LA DENSITÀ DI PROBABILITÀ È...

$$f(x) = \frac{1}{2^{\frac{df}{2}} \times \int_0^{\infty} x^{\frac{df}{2}-1} e^{-x} dx} \times x^{\frac{df}{2}-1} \times e^{-\frac{x}{2}}$$

PER $x > 0 \dots$

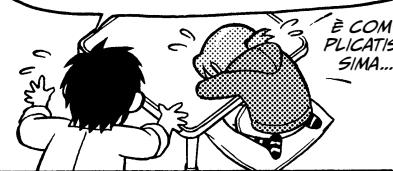
$$f(x) = 0$$

PER $x \leq 0 \dots$

IN STATISTICA DICHIAMO CHE "X SEGUE UNA DISTRIBUZIONE CHI-QUADRO CON N GRADI DI LIBERTÀ".



NIENTE PAURA. NON AVRAI MAI BISOGNO DI IMPARARE QUESTA FORMULA, SE NON DIVENTERAI UNA MATEMATICA.



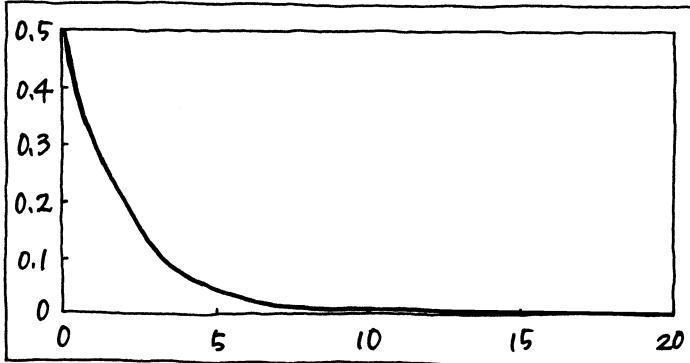
L'HO SCRITTA SOLO PER SPAVENTARTI.



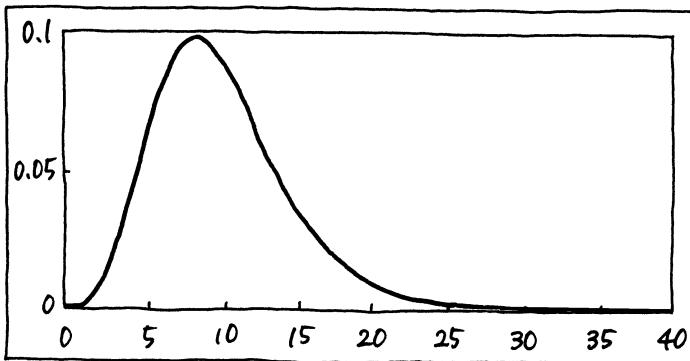
PER COMINCIARE, DIAMO UN'OCCASIONE AI GRAFICI CON 2, 10 E 20 GRADI DI LIBERTÀ.



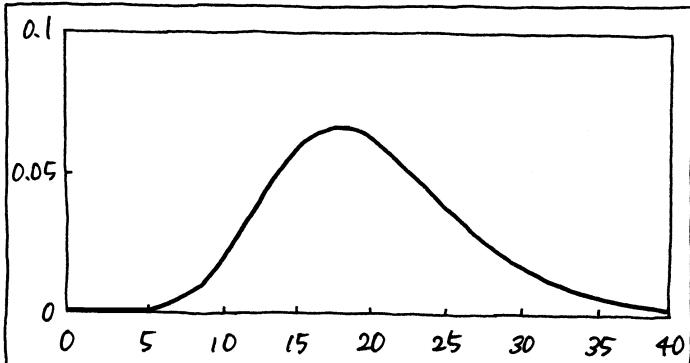
2 GRADI DI LIBERTÀ



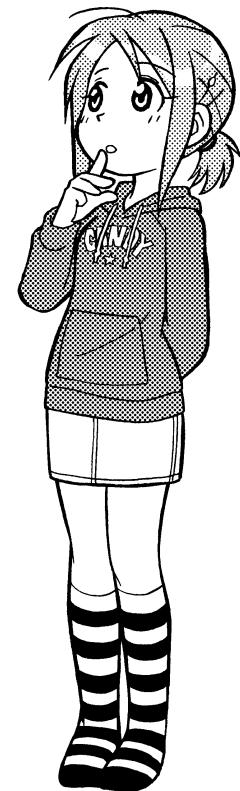
10 GRADI DI LIBERTÀ

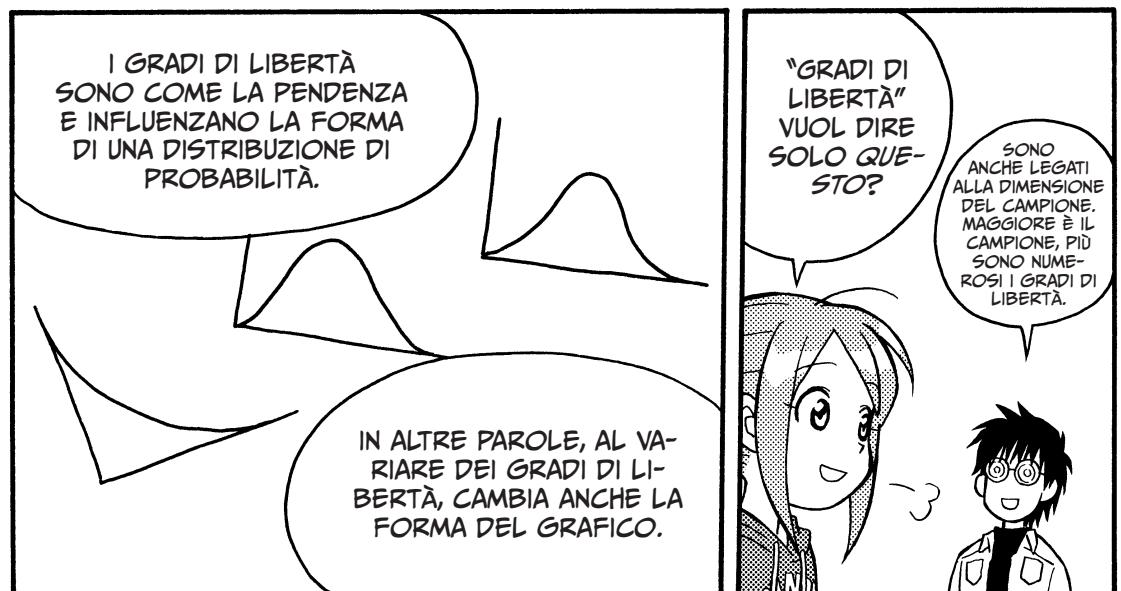
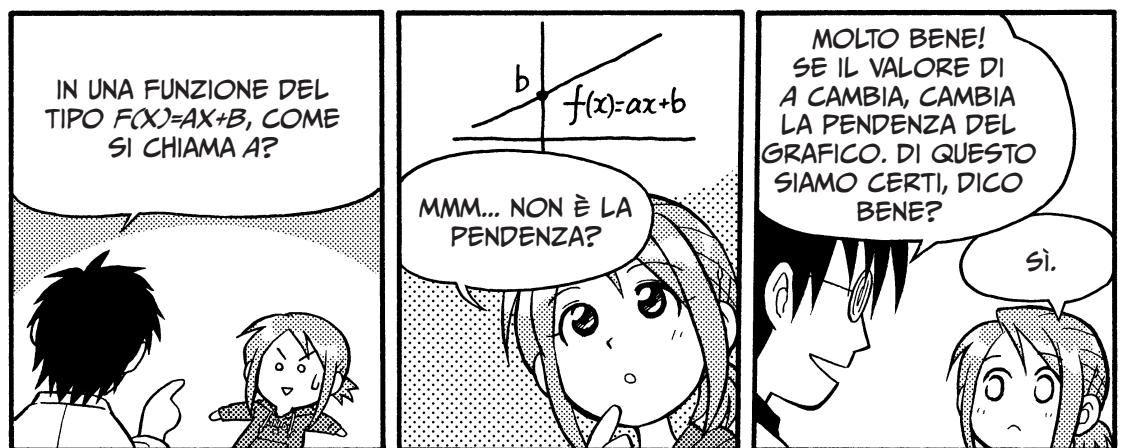


20 GRADI DI LIBERTÀ



LA FORMA
DEL GRAFICO
CAMBIA COL NU-
MERO DEI GRADI
DI LIBERTÀ.



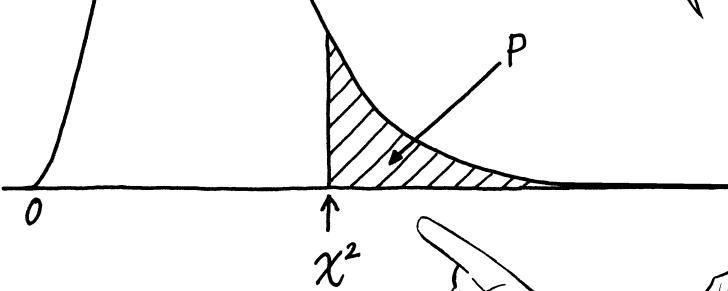


PROPRIO COME ESISTE UNA TABELLA DI PROBABILITÀ PER LA DISTRIBUZIONE NORMALE STANDARD...

...CE N'È UNA PER LA PROBABILITÀ DELLA DISTRIBUZIONE CHI-QUADRO.

SI TRATTA DI UNA TABELLA...

...CHE RIPORTA I VALORI DI χ^2 SULL'ASSE DELLE X, IN CORRISPONDENZA DELLA PROBABILITÀ (E SAPPIAMO GIÀ CHE CORRISPONDE ALL'AREA E ALLA PERCENTUALE) DELLA SUPERFICIE P.





P	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
GRADI DI LIBERTÀ								
1	0.000039	0.0002	0.0010	0.0039	3.8415	5.0239	6.6349	7.8794
2	0.0100	0.0201	0.0506	0.1026	5.9915	7.3778	9.2104	10.5965
3	0.0717	0.1148	0.2158	0.3518	7.8147	9.3484	11.3449	12.8381
4	0.2070	0.2971	0.4844	0.7107	9.4877	11.1433	13.2767	14.8602
5	0.4118	0.5543	0.8312	1.1455	11.0705	12.8325	15.0863	16.7496
6	0.6757	0.8721	1.2373	1.6354	12.5916	14.4494	16.8119	18.5475
7	0.9893	1.2390	1.6899	2.1673	14.0671	16.0128	18.4753	20.2777
8	1.3444	1.6465	2.1797	2.7326	15.5073	17.5345	20.0902	21.9549
9	1.7349	2.0879	2.7004	3.3251	16.9190	19.0228	21.6660	23.5893
10	2.1558	2.5582	3.2470	3.9403	18.3070	20.4832	23.2093	25.1881
...



CON LA TABELLA DELLA DISTRIBUZIONE NORMALE STANDARD, DATO UN CERTO VALORE DELLA COORDINATA X OTTENIAMO LA PROBABILITÀ ASSOCIATA.

PROBABILITÀ
(= AREA = PERCENTUALE)

CON LA TABELLA DELLA DISTRIBUZIONE CHI-QUADRO, SUCCIDE IL CONTRARIO: A PARTIRE DALLA PROBABILITÀ, RICAVIAMO LA COORDINATA X ASSOCIATA.

QUESTO VALORE!

ODDIO, CHE CONFUSIONE...!

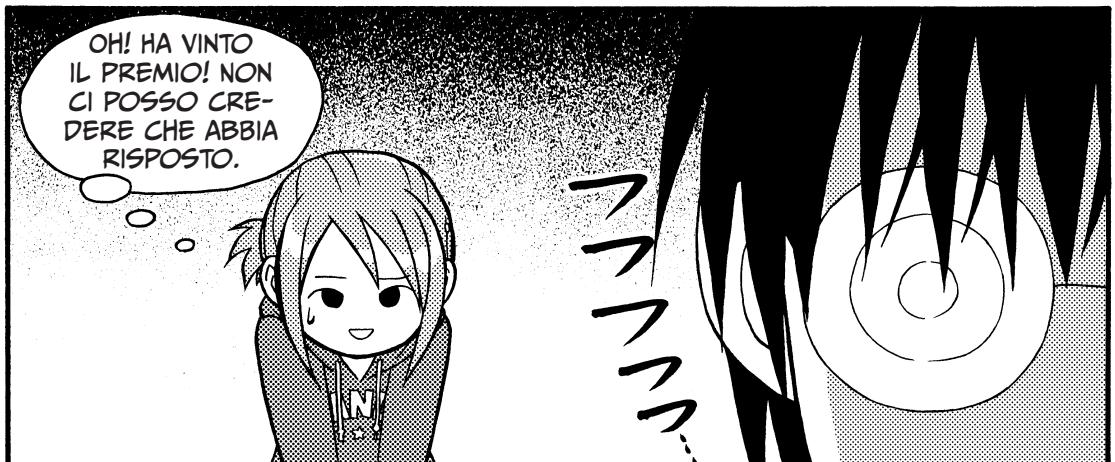
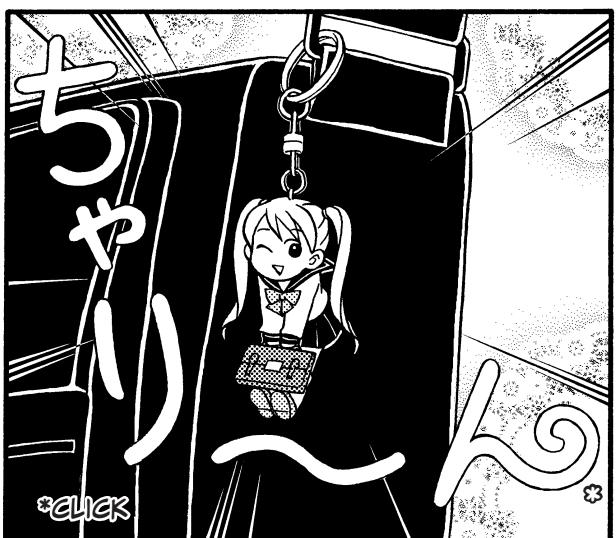
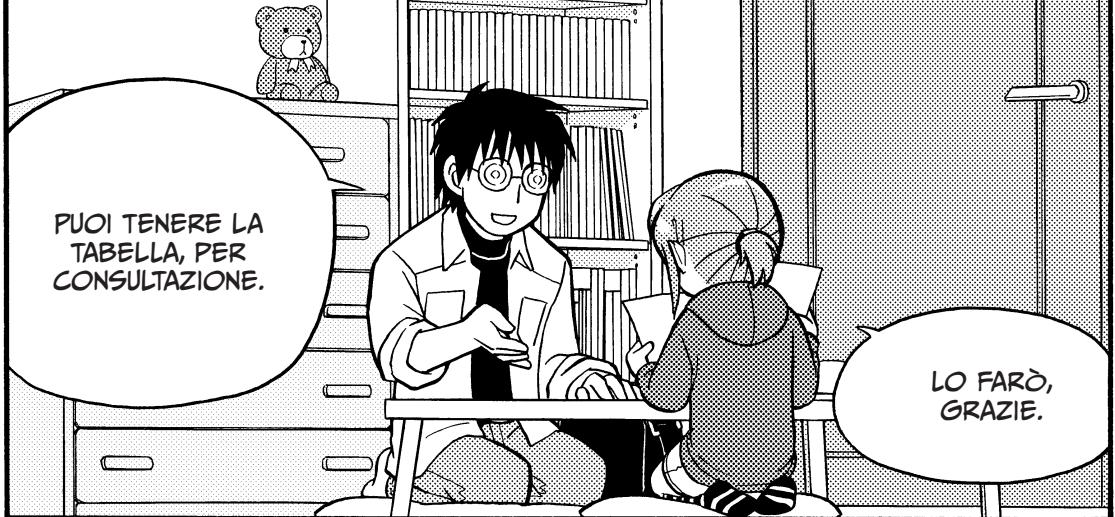
CALMA!

TROVIAMO IL VALORE DI χ^2 NEL CASO DI UN UNICO GRADO DI LIBERTÀ E DI $P=0,05$.

*TABELLA DELLA DISTRIBUZIONE CHI-QUADRO

ALL'INCROCIO DELLA RIGA DI 1 E DELLA COLONNA DI 0,05 TROVIAMO...

3,8415.



5. DISTRIBUZIONE T

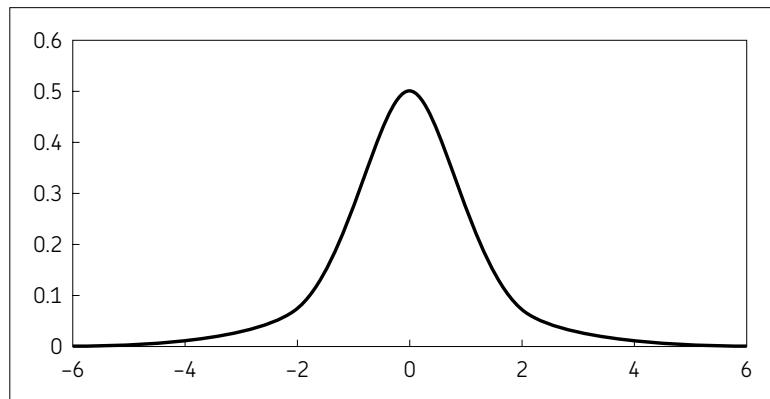


Questa funzione di densità è particolarmente importante in statistica.

$$f(x) = \frac{\int_0^{\infty} x^{\frac{df+1}{2}-1} e^{-x} dx}{\sqrt{df \times \pi} \times \int_0^{\infty} x^{\frac{df}{2}-1} e^{-x} dx} \times \left(1 + \frac{x^2}{df}\right)^{-\frac{df+1}{2}}$$

Quando una funzione di densità in x ha questo aspetto, diciamo che “ x segue una distribuzione t con n gradi di libertà”.

Ecco il caso di 5 gradi di libertà.



6. DISTRIBUZIONE F

Anche questa funzione di densità è importante.

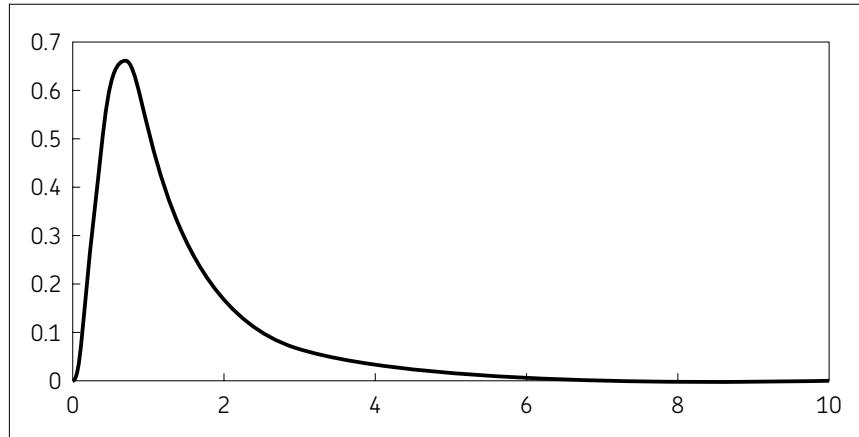
per $x > 0$:

$$f(x) = \frac{\left(\int_0^{\infty} x^{\frac{\text{prima df}+}{2}-1} e^{-x} dx\right) \times \left(\frac{\text{prima df}}{2}\right)^{\frac{\text{prima df}}{2}} \times \left(\frac{\text{seconda df}}{2}\right)^{\frac{\text{seconda df}}{2}}}{\left(\int_0^{\infty} x^{\frac{\text{prima df}}{2}-1} e^{-x} dx\right) \times \left(\int_0^{\infty} x^{\frac{\text{seconda df}}{2}-1} e^{-x} dx\right)} \times \frac{x^{\frac{\text{prima df}}{2}-1}}{\left(\frac{\text{prima df}+\text{seconda df}}{2}\right)^{\frac{\left(\text{prima df}+\text{seconda df}\right)}{2}}}$$

per $x \leq 0$: $f(x) = 0$

Con funzioni di densità di questo tipo, diciamo che “ x segue una distribuzione F con primo grado di libertà m e secondo grado di libertà n ”.

Ecco il caso con primo grado di libertà 10 e secondo grado di libertà 5:



7. DISTRIBUZIONI E FOGLI ELETTRONICI

Fino alla diffusione dei personal computer (più o meno all'inizio degli anni Novanta) non era facile calcolare le probabilità senza l'aiuto delle tabelle della distribuzione normale standard o della distribuzione chi-quadro. Oggi le tabelle sono superate e al loro posto possiamo usare un foglio di calcolo: nei fogli elettronici esistono delle funzioni che permettono di trovare gli stessi valori delle tabelle. I nomi possono variare a seconda del programma ma si tratta di funzioni standard sempre disponibili in ogni foglio dotato di funzioni statistiche.

TABELLA 5.1 – FUNZIONI RELATIVE A VARIE DISTRIBUZIONI

Distribuzione	Funzione	Descrizione della funzione
Normale*	DISTRIB.NORM	Calcola la probabilità corrispondente a un valore dell'ascissa
Normale	INV.NORM	Calcola l'ascissa corrispondente alla probabilità
Normale standard	DISTRIB.NORM.ST	Calcola la probabilità corrispondente a un valore dell'ascissa
Inversa	INV.NORM.ST	Calcola l'ascissa corrispondente alla probabilità
Chi-quadro	DISTRIB.CHI	Calcola la probabilità corrispondente a un valore dell'ascissa
Chi-quadro inversa	INV.CHI	Calcola l'ascissa corrispondente alla probabilità
t	DISTRIB.T	Calcola la probabilità corrispondente a un valore dell'ascissa
t inversa	INV.T	Calcola l'ascissa corrispondente alla probabilità
F	DISTRIB.F	Calcola la probabilità corrispondente a un valore dell'ascissa
F inversa	INV.F	Calcola l'ascissa corrispondente alla probabilità

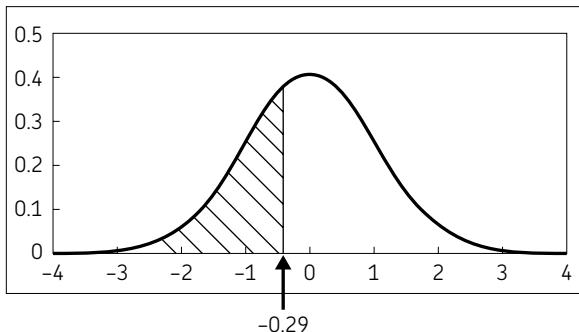
* La densità di probabilità dipende dalla media e dalla deviazione standard e quindi non esiste una singola "tabella della distribuzione normale". Con un foglio elettronico possiamo però calcolarne comodamente i valori e costruire di volta in volta la tabella relativa alla distribuzione normale del caso.

ESERCIZIO CON SOLUZIONE



ESERCIZIO

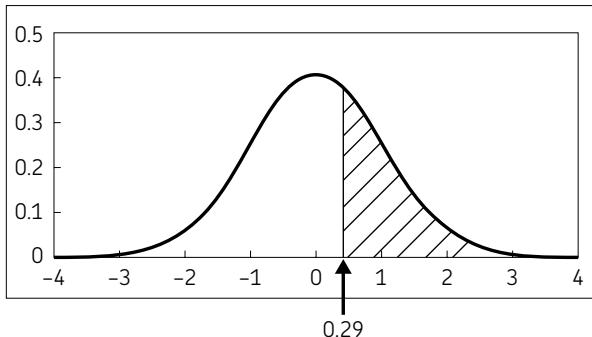
- Calcolare la probabilità (l'area tratteggiata del grafico) usando la tabella della distribuzione normale standard di pagina 93.



- Calcolare il valore di χ^2 nel caso di due gradi di libertà e di $P=0,05$, usando la tabella della distribuzione chi-quadro di pagina 103.

SOLUZIONE

- La distribuzione normale standard è simmetrica rispetto all'asse delle y quindi la probabilità in questione è uguale a quella tratteggiata in questo grafico.



Per $z = 0.29 = 0.2 + 0.09$, secondo la tabella della distribuzione normale standard la probabilità è 0,1141. La probabilità che cerchiamo è quindi $0,5 - 0,1141 = 0,3859$.

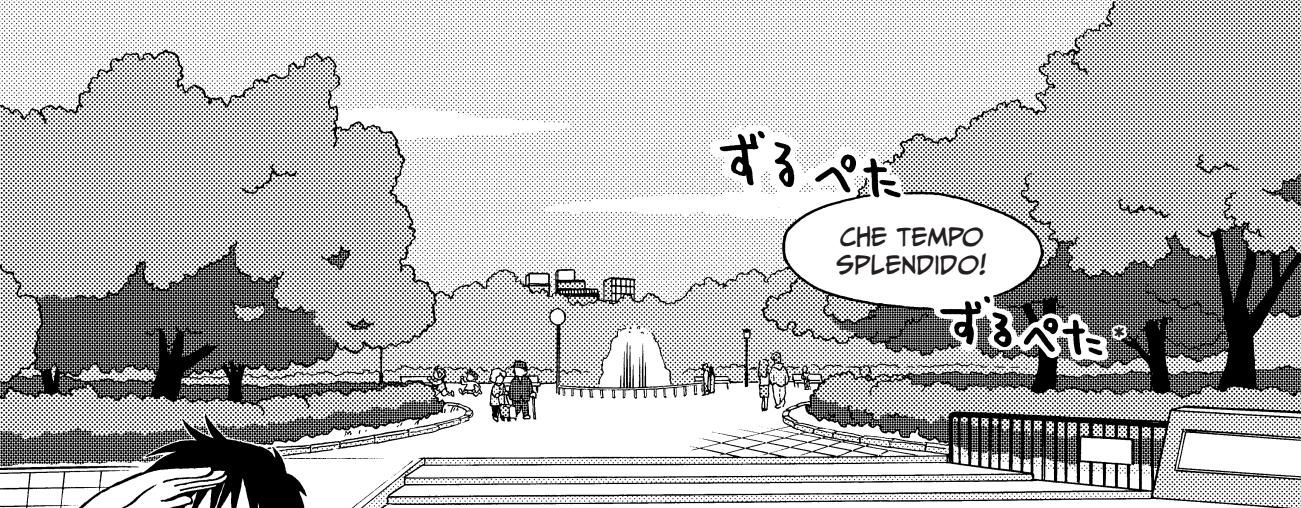
- Secondo la tabella della distribuzione chi-quadro, il valore di χ^2 cercato è 5,9915.

RIASSUMENDO

- Alcune delle (funzioni di) densità di probabilità più comuni sono:
 - Distribuzione normale
 - Distribuzione normale standard
 - Distribuzione chi-quadro
 - Distribuzione t
 - Distribuzione F
- L'area compresa tra il grafico della (funzione di) densità di probabilità e l'asse orizzontale vale 1 e corrisponde alla probabilità, cioè a una percentuale.
- Usando le funzioni di un foglio elettronico o una tavola di probabilità per la distribuzione del caso, si possono calcolare:
 - La probabilità corrispondente a un dato punto sull'asse delle x
 - Il punto sull'asse delle x corrispondente a una data probabilità

6

CHE RELAZIONE C'È
TRA DUE VARIABILI?



ざるペト

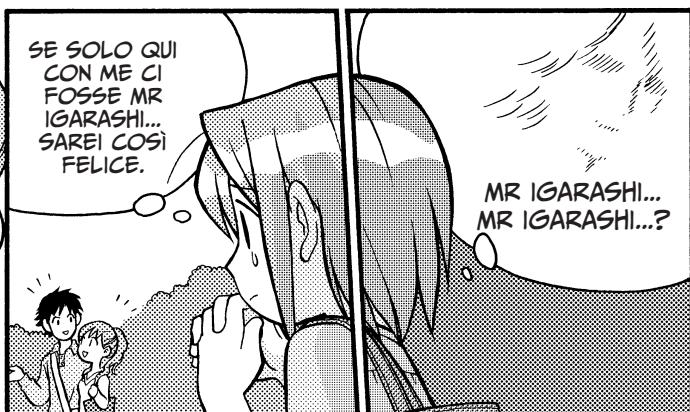
CHE TEMPO
SPLENDIDO!

ざるペト*

*SCREECH SCREECH



UN'OTTIMA
GIORNATA
PER STUDIARE
ALL'APERTO.



*SCREECH
SCREECH

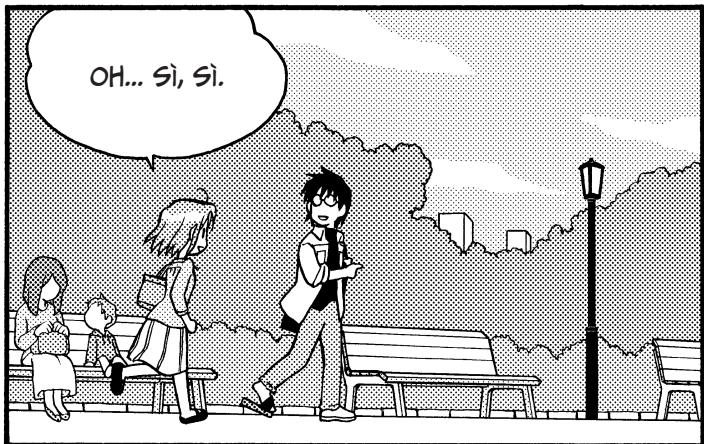
ざるペト

MR YAMAMOTO
È TROPPO
IMPEGNAVITO
PER ME... E MI
STA FACENDO
DIMENTICARE
MR IGARASHI!



MI ASCOLTI?

OH... SÌ, SÌ.



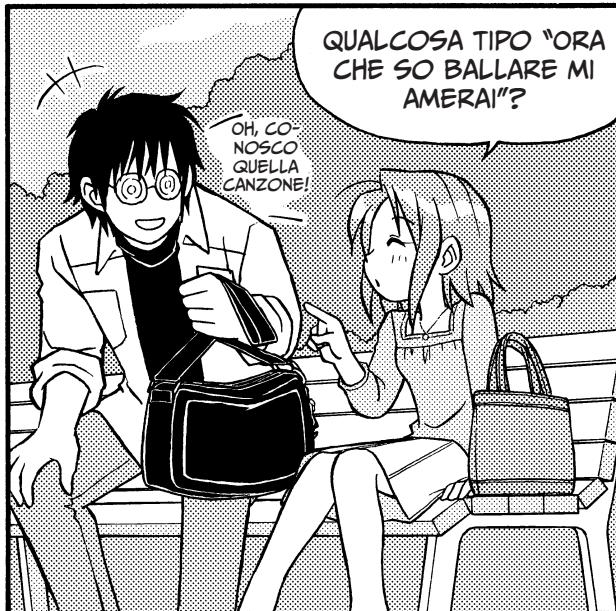
PER ESEMPIO, UNA PER-
SONA PIÙ ALTA HA ANCHE
UN PESO MAGGIORE?
OPPURE: DUE PERSONE A
CUI PIACCIONO BIBITE DI-
VERSE HANNO ANCHE ETÀ
DIVERSE?

PERSONE CHE ABITA-
NO IN LUOGHI DIVERSI
HANNO ANCHE DIVER-
SE POSIZIONI
POLITICHE?

OH, GRAZIE! HA PULITO LA
PANCHINA PER ME!

QUALCOSA TIPO "ORA
CHE SO BALLARE MI
AMERAI"?

OH, CO-
NOSCO
QUELLA
CANZONE!



VEDIA-
MO...

COMINCIA
LA LEZIO-
NE!

SPALANG!

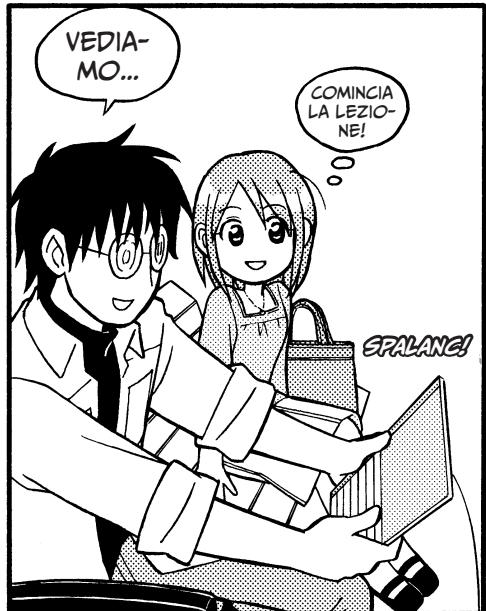
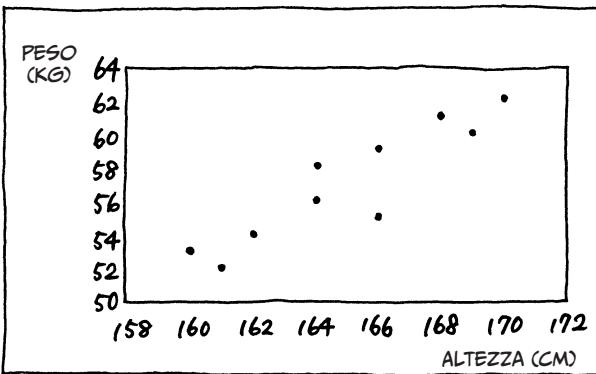
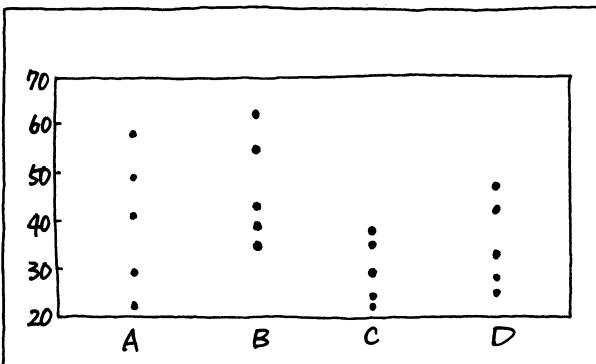


GRAFICO A DISPERSIONE DEL PESO E DELLE ALTEZZE



DUE DATI NUMERICI

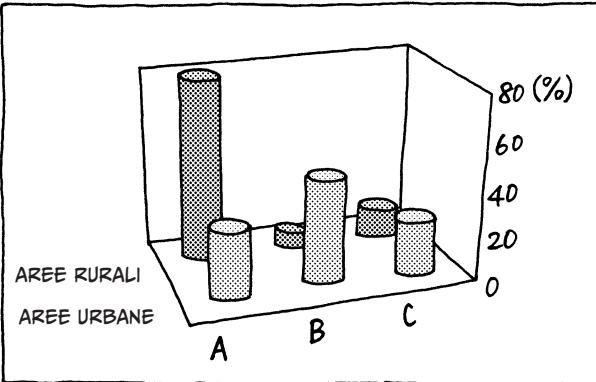
GRAFICO A DISPERSIONE DELLE ETÀ E DELLE BIBITE



UN DATO NUMERICO E UNO CATEGORICO

TRACCIANDO UN GRAFICO PUOI CAPIRE SE TRA LE DUE VARIABILI C'È UNA RELAZIONE OPPURE NO.

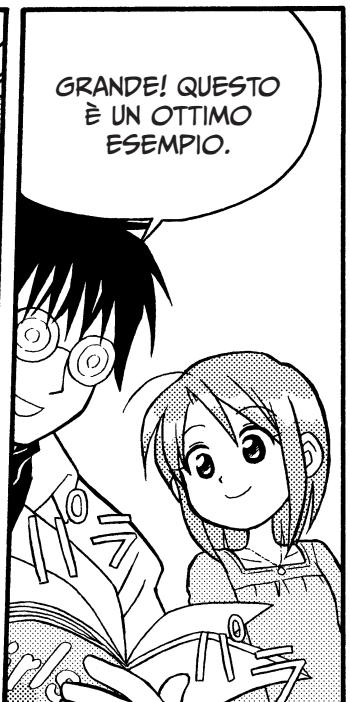
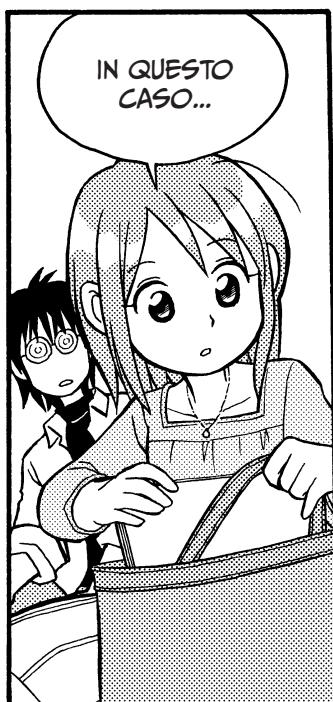
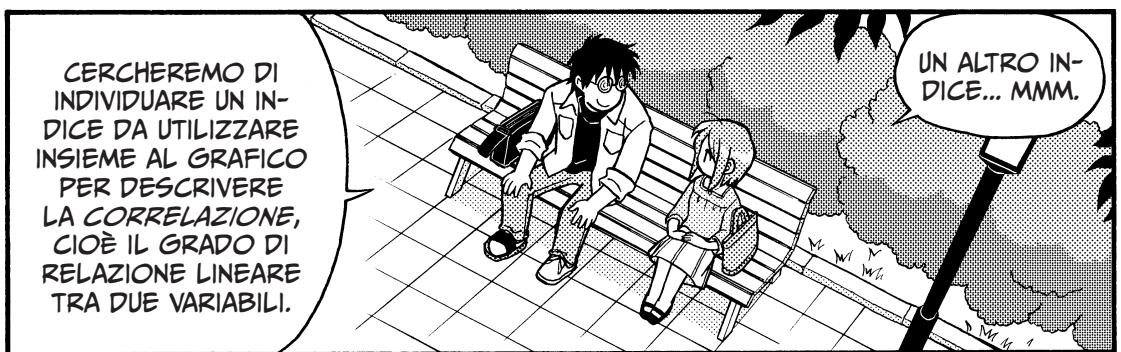
GRAFICO A CILINDRO DEL LUOGO DI RESIDENZA E DELLE PREFERENZE PER IL PARTITO POLITICO X



DUE DATI CATEGORICI

AH!





1. COEFFICIENTE DI CORRELAZIONE

ECCO QUA UN BEL SON-
DAGGIO SULLE SPESE
PER COSMETICI E ABBI-
GLIAMENTO.

ENTRAM-
BE SONO
VARIABILI
NUMERICHE!

Sondaggio!

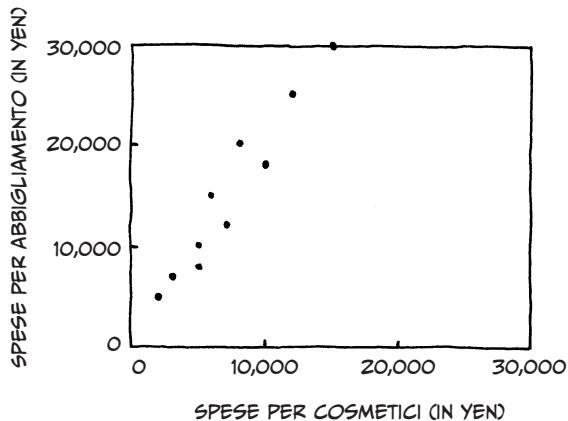
Dieci ragazze dai 20 anni in su hanno risposto su
Spese mensili per cosmetici e abbigliamento

Intervistata	Spese per cosmetici (¥)	Spese per abbigliamento (¥)
Ms. A	3,000	7,000
Ms. B	5,000	8,000
Ms. C	12,000	25,000
Ms. D	2,000	5,000
Ms. E	7,000	12,000
Ms. F	15,000	30,000
Ms. G	5,000	10,000
Ms. H	6,000	15,000
Ms. I	8,000	20,000
Ms. J	10,000	18,000

COMINCIAMO
COL FARE UN
GRAFICO.

SISSI-
GNORE!

GRAFICO A DISPERSIONE DELLE SPESE MEN-
SILI PER COSMETICI E ABBIGLIAMENTO

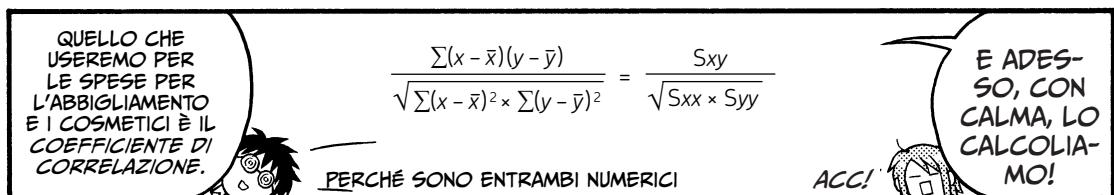


OVVIAMENTE, CHI SPEN-
DE DI PIÙ IN COSMETICI
SPENDE DI PIÙ ANCHE IN
ABBIGLIAMENTO.

PERCHÉ NON CER-
CHIAMO DI CAPIRE
IL GRADO DELLA
RELAZIONE?

Tipi di dati	Indice	Intervallo dei valori	Formula
Numerico e numerico	Coefficiente di correlazione	-1 - 1	$\frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \times \sum(y - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}}$
Numerico e categorico	Rapporto di correlazione*	0 - 1	varianza interclasse (varianza intraclass + varianza interclasse)
Categorico e categorico	Coefficiente di Cramer*	0 - 1	χ^2_0 $\sqrt{\min(\text{numero delle righe della tabella di contingenza}, \text{numero delle colonne della tabella di contingenza}) - 1}}$

*v. pagina 121, "Rapporto di correlazione" e pagina 127, "Coefficiente di Cramer".



	Spese per cosmetici (¥)	Spese per abbigliamento (¥)	PROCEDURA DI CALCOLO DEL COEFFICIENTE DI CORRELAZIONE TRA LE SPESSE MENSILI PER COSMETICI E ABBIGLIAMENTO				
	x	y	x - \bar{x}	y - \bar{y}	(x - \bar{x}) ²	(y - \bar{y}) ²	(x - \bar{x})(y - \bar{y})
Ms. A	3,000	7,000	-4,300	-8,000	18,490,000	64,000,000	34,400,000
Ms. B	5,000	8,000	-2,300	-7,000	5,290,000	49,000,000	16,100,000
Ms. C	12,000	25,000	4,700	10,000	22,090,000	100,000,000	47,000,000
Ms. D	2,000	5,000	-5,300	-10,000	28,090,000	100,000,000	53,000,000
Ms. E	7,000	12,000	-300	-3,000	90,000	9,000,000	900,000
Ms. F	15,000	30,000	7,700	15,000	59,290,000	225,000,000	115,500,000
Ms. G	5,000	10,000	-2,300	-5,000	5,290,000	25,000,000	11,500,000
Ms. H	6,000	15,000	-1,300	0	1,690,000	0	0
Ms. I	8,000	20,000	700	5,000	490,000	25,000,000	3,500,000
Ms. J	10,000	18,000	2,700	3,000	7,290,000	9,000,000	8,100,000
Somma	73,000	150,000	0	0	148,100,000	606,000,000	290,000,000
Media	7,300	15,000			S _{xx}	S _{yy}	S _{xy}
	\bar{x}	\bar{y}					

E ADESSO INSE-
RIAMO I VALORI
NELLA FORMULA.

$$\frac{s_{xy}}{\sqrt{s_{xx} \times s_{yy}}} = \frac{290,000,000}{\sqrt{148,100,000 \times 606,000,000}} = 0,9680$$

CON UNA
CALCOLATRICE È
FACILE.

IL COEFFICIENTE DI
CORRELAZIONE È...
0,9680!



IL COEFFICIENTE È
TANTO PIÙ VICINO A +1
O -1 QUANTO PIÙ LA
RELAZIONE LINEARE
TRA LE DUE VARIABILI
È FORTE.

SE LA RELAZIONE
SI INDEBOLISCE, SI
AVVICINA A ZERO.



INTERES-
SANTE.

IL NOSTRO RISULTATO È
ABBASTANZA VICINO A 1, IL
CHE IMPLICA CHE LE SPESE
PER I COSMETICI E QUELLE
PER L'ABBIGLIAMENTO SONO
FORTEMENTE CORRELATE.



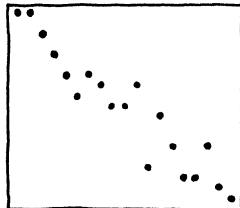
HAI DECISAMENTE
RAGIONE!

E QUANDO SI
AVVICINA A -1?

QUESTO CAPITA SE UNA DELLE
DUE GRANDEZZE AUMENTA
QUANDO L'ALTRA CALA.

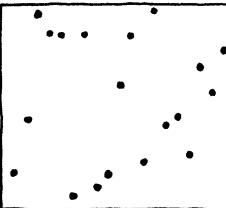


CORRELAZIONE
NEGATIVA



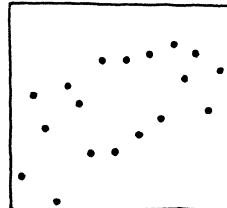
CIRCA -1

NESSUNA
CORRELAZIONE

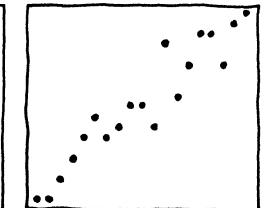


CIRCA 0

CORRELAZIONE POSITIVA



CIRCA 0,5



CIRCA 1

COEFFICIENTE DI CORRELAZIONE



SE IL COEFFICIENTE DI CORRELAZIONE È POSITIVO, COME IN QUESTO CASO, DICHIAMO CHE "C'È UNA CORRELAZIONE POSITIVA" E SE È NEGATIVO CHE "C'È UNA CORRELAZIONE NEGATIVA".

SE È ZERO, DICHIAMO CHE "SONO SCORRELATE".

HO CAPITO TUTTO.



SEMPRE A PROPOSITO DEL COEFFICIENTE DI CORRELAZIONE...

PURTROppo, NON ESISTONO STANDARD STATISTICI CHE ASSICURINO QUANTO FORTEMENTE LE DUE VARIABILI SIANO CORRELATE.

CHE INDICE INAFFIDABILE...



VALORI CONVENZIONALI DEL COEFFICIENTE DI CORRELAZIONE

Valore assoluto del coefficiente di correlazione	Valutazione	Valutazione di massima
1,0-0,9	⇒ Molto fortemente correlati	
0,9-0,7	⇒ Fortemente correlati	Correlati
0,7-0,5	⇒ Debolmente correlati	
Meno di 0,5	⇒ Molto debolmente correlati	Non correlati



PER TUA INFORMAZIONE, QUESTI SONO ALCUNI INTERVALLI INFORMALI PER I VALORI DEL COEFFICIENTE DI CORRELAZIONE.

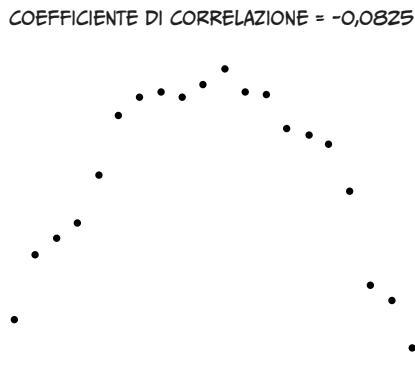


ATTENZIONE

Come ho detto, il coefficiente di correlazione è un indice che misura il grado di relazione lineare tra due variabili numeriche.



CAMPIONE INADATTO AL COEFFICIENTE DI CORRELAZIONE



Per esempio, tra le due variabili del grafico c'è chiaramente una qualche relazione ma il coefficiente di correlazione è vicino allo 0 perché tale relazione è *non lineare*.

2. RAPPORTO DI CORRELAZIONE

ANDIAMO AVANTI!
C'È UN SONDAGGIO
ANCHE SUI MARCHI
PREFERITI IN BASE
ALL'ETÀ!

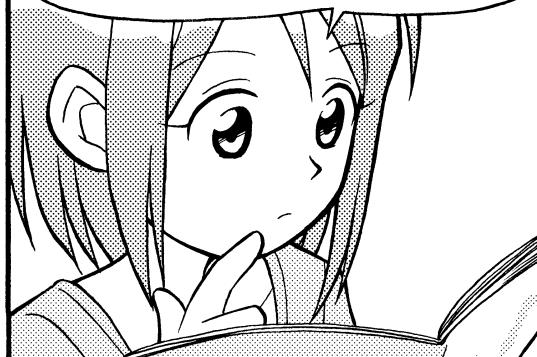


Sondaggio a Everyhills

Età e marchio preferito

Intervistata	Età	Marchio
Ms. A	27	Theremes
Ms. B	33	Channelior
Ms. C	16	Bureperry
Ms. D	29	Bureperry
Ms. E	32	Channelior
Ms. F	23	Theremes
Ms. G	25	Channelior
Ms. H	28	Theremes
Ms. I	22	Bureperry
Ms. J	18	Bureperry
Ms. K	26	Channelior
Ms. L	26	Theremes
Ms. M	15	Bureperry
Ms. N	29	Channelior
Ms. O	26	Bureperry

PER DATI NUMERICI E CATEGORICI
USIAMO IL RAPPORTO DI CORRELAZIONE, CHE ASSUME VALORI...
TRA ZERO E 1.



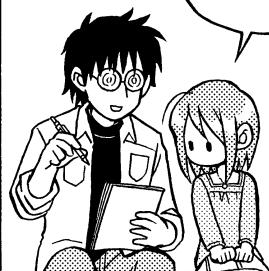
ANCHE IN QUESTO CASO
IL VALORE È PIÙ VICINO A
1 QUANDO È PIÙ FORTE LA
CORRELAZIONE?



ETÀ E MARCHIO PREFERITO

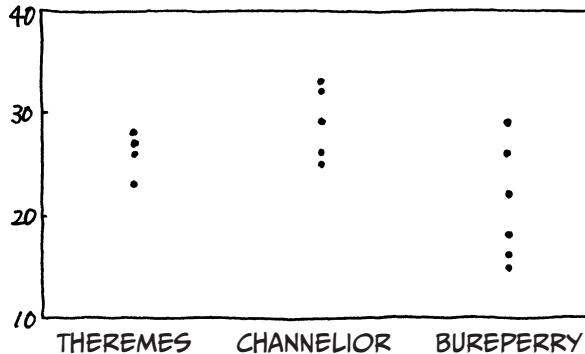
RIORGANIZZIAMO LA TABELLA.

MMM...



	THERMES	CHANNELIOR	BUREPERRY	
23	25	15		
26	26	16		
27	29	18		
28	32	22		
	33	26		
		29		
SOMMA	104	145	126	375
MEDIA	26	29	21	25

GRAFICO A DISPERSIONE DI ETÀ E MARCHIO PREFERITO

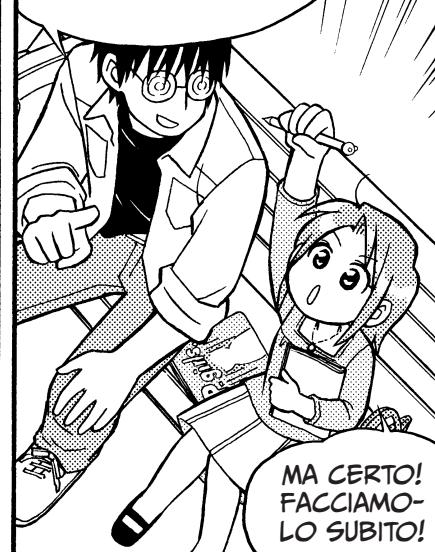


E ADESSO,
FACCIAMO
UN GRAFICO.

WOW! SEMBRA
PROPRIO CHE
CI SIANO DELLE
CORRELAZIONI.



A QUESTO PUNTO
POSSIAMO CAL-
COLARE IL VALORE
DEL RAPPORTO DI
CORRELAZIONE.



MA CERTO!
FACCIAMO-
LO SUBITO!

Lo troveremo seguendo i passi da 1 a 4.



Passo 1

Svolgiamo i calcoli di questa tabella

	Somma
(Theremes - media di Theremes) ²	$(23 - 26)^2 = (-3)^2 = 9$ $(26 - 26)^2 = 0^2 = 0$ $(27 - 26)^2 = 1^2 = 1$ $(28 - 26)^2 = 2^2 = 4$ 14 $\curvearrowleft S_{TT}$
(Channelior - media di Channelior) ²	$(25 - 29)^2 = (-4)^2 = 16$ $(26 - 29)^2 = (-3)^2 = 9$ $(29 - 29)^2 = 0^2 = 0$ $(32 - 29)^2 = 3^2 = 9$ $(33 - 29)^2 = 4^2 = 16$ 50 $\curvearrowleft S_{CC}$
(Bureperry - media di Bureperry) ²	$(15 - 21)^2 = (-6)^2 = 36$ $(16 - 21)^2 = (-5)^2 = 25$ $(18 - 21)^2 = (-3)^2 = 9$ $(22 - 21)^2 = 1^2 = 1$ $(26 - 21)^2 = 5^2 = 25$ $(29 - 21)^2 = 8^2 = 64$ 160 $\curvearrowleft S_{BB}$

Passo 2

Calcoliamo la varianza intraclasse ($STT+SCC+SBB$ =quanto variano i dati all'interno delle categorie).

$$S_{TT} + S_{CC} + S_{BB} = 14 + 50 + 160 = 224$$

Passo 3

Calcoliamo la varianza interclasse, cioè quanto i gruppi differiscono tra loro.

$$\begin{aligned} & (\text{numero di preferenze per Theremes}) \times (\text{media di Theremes} - \text{media di tutti i dati})^2 \\ & + (\text{numero di preferenze per Channelior}) \times (\text{media di Channelior} - \text{media di tutti i dati})^2 \\ & + (\text{numero di preferenze per Bureperry}) \times (\text{media di Bureperry} - \text{media di tutti i dati})^2 \end{aligned}$$

$$4 \times (26 - 25)^2 + 5 \times (29 - 25)^2 + 6 \times (21 - 25)^2$$

$$= 4 \times 1 + 5 \times 16 + 6 \times 16$$

$$= 4 + 80 + 96$$

$$= 180$$

Passo 4

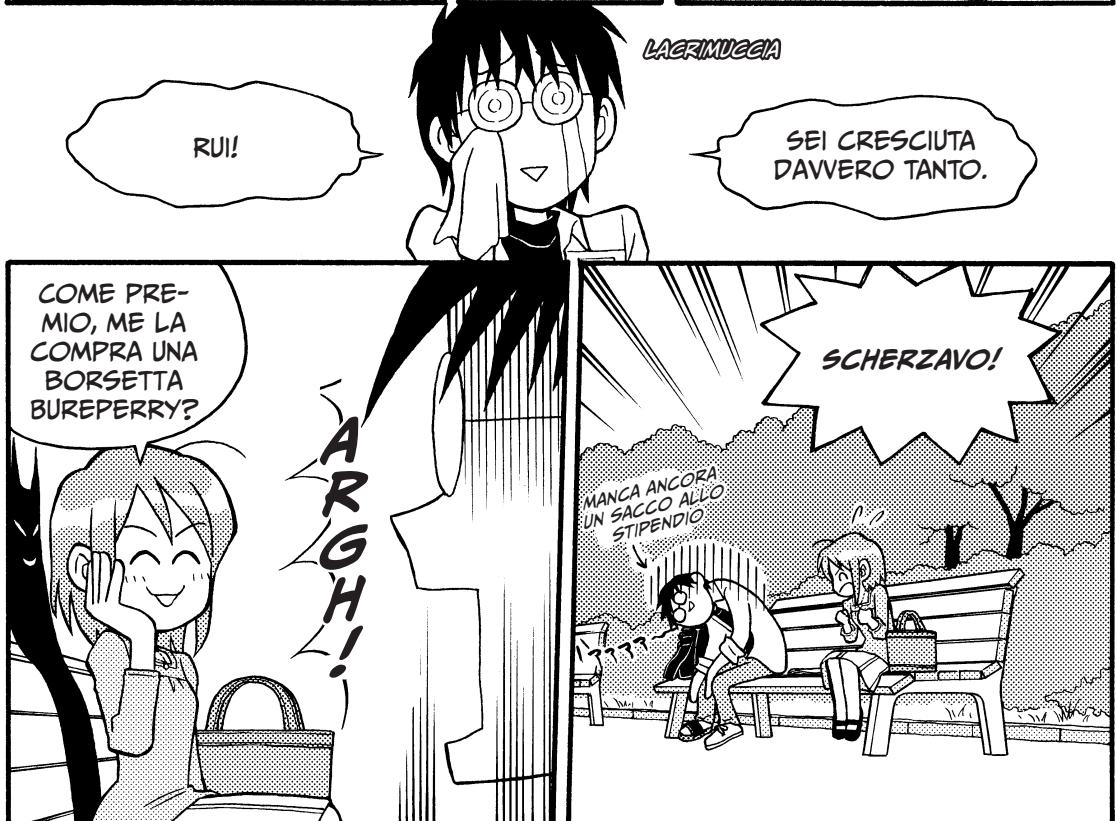
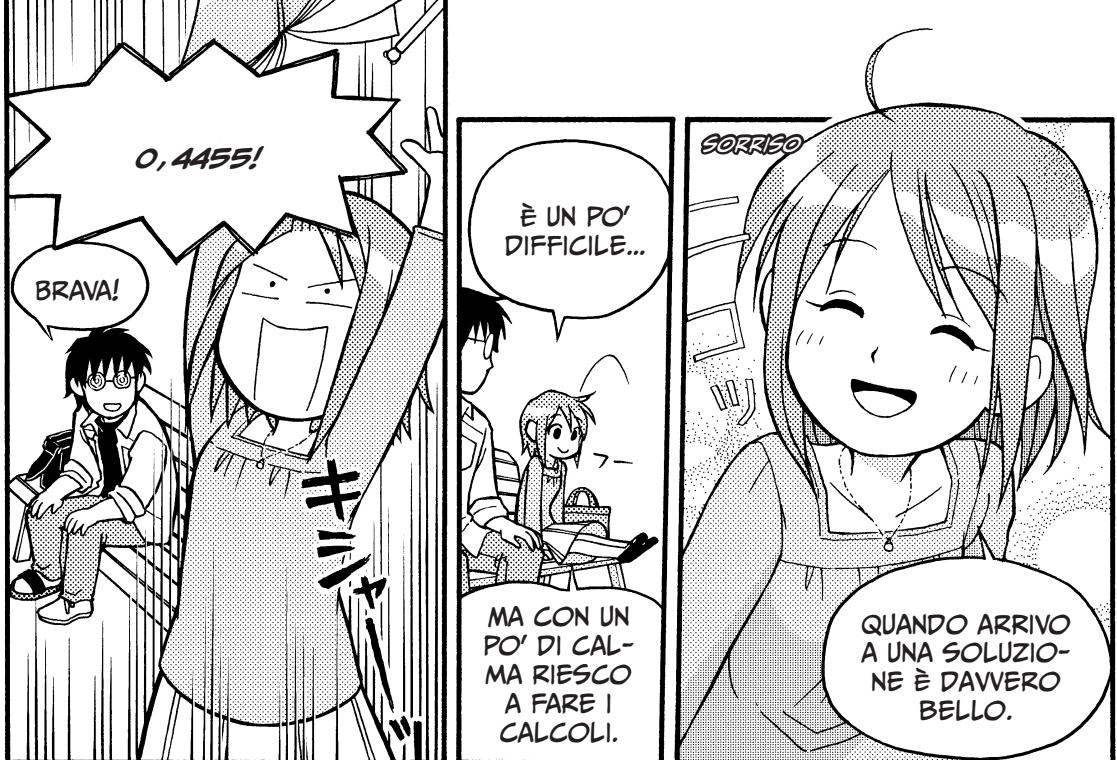
Calcoliamo il Rapporto di correlazione.

$$\frac{\text{varianza interclasse}}{(\text{varianza intraclassica} + \text{varianza interclasse})}$$

$$\frac{180}{224 + 180} = \frac{180}{404} = 0.4455$$

QUINDI... IL RAPPORTO DI CORRELAZIONE PER ETÀ E MARCHIO PREFERITO È...

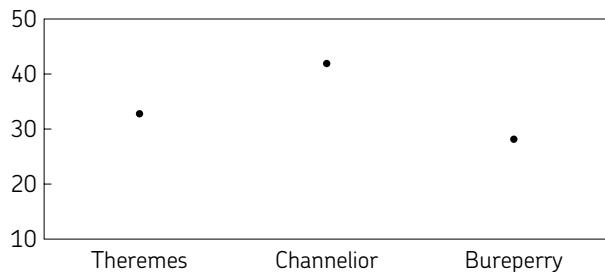




Come abbiamo visto, il rapporto di correlazione varia tra 0 e 1: più la correlazione tra due variabili è forte e più si avvicina a 1, più la correlazione è debole e più si avvicina allo 0. In questi diagrammi trovate un po' di dettagli in più.

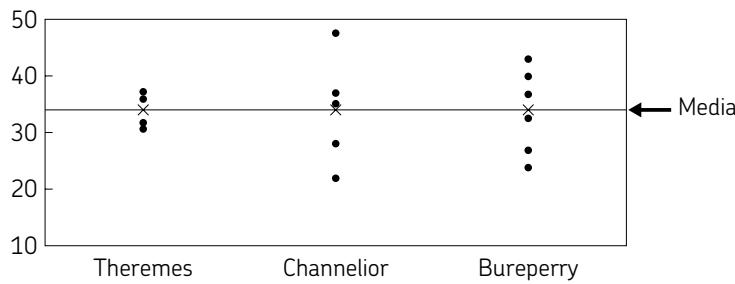


Questo è un grafico a dispersione di età e marchio preferito (con rapporto di correlazione uguale a 1).



Rapporto di correlazione = 1 \Leftrightarrow i dati all'interno di ciascun gruppo sono uguali \Leftrightarrow la varianza intraclasse è 0.

Questo è un grafico a dispersione di età e marchio preferito (con Rapporto di correlazione uguale a 0).



Rapporto di correlazione = 0 \Leftrightarrow le medie dei gruppi sono uguali \Leftrightarrow la varianza interclasse è 0.



Ahimè, non esistono veri e propri standard statistici del tipo "due variabili sono fortemente correlate se il rapporto di correlazione è al di sopra di un certo valore". Ma alcuni standard informali in uso sono i seguenti.

VALORI CONVENZIONALI DEL RAPPORTO DI CORRELAZIONE

Rapporto di correlazione	Valutazione	Valutazione di massima
1.0-0.8	⇒ Molto fortemente correlati	
0.8-0.5	⇒ Fortemente correlati	Correlati
0.5-0.25	⇒ Debolmente correlati	
Meno di 0.25	⇒ Molto debolmente correlati	Non correlati

Quindi nel nostro caso, i calcoli danno 0,4455 e possiamo dire che le variabili sono debolmente correlate.



3. IL COEFFICIENTE DI CRAMER

MI CHIEDO SE NELLA RIVISTA CI SIA UN BUON ESEMPIO PER SPIEGARE LA CORRELAZIONE TRA DUE VARIABILI CATEGORICHE.



CHE NE DICE DI QUESTO?

告白されるとした
どの方法でされたい?

ABBIAMO CHIESTO A 300 LICEALI
"COME VORRESTE CHE VI CHIEDESSERO DI USCIRE?".

MMM... "IL MIO
MODO IDEALE
SAREBBE PER TE-
LEFONO, EMAIL, DI
PERSONA" ...?

MI SEMBRA UN
BUON ESEMPIO.

LE ASSURDITÀ CHE SI
TROVANO NELLE RIVISTE
FEMMINILI NON CESSANO
MAI DI STUPIRMI.

NON
SONO
ASSUR-
DITA!

TABELLA DI CONTINGENZA TRA SESSO E MODALITÀ D'INVITO PREFERITA

		MODALITÀ D'INVITO PREFERITA			SOMMA
SESSO	FEMMINE	TELEFONO	E-MAIL	DI PERSONA	
	MASCHI	38	40	74	152
	SOMMA	72	101	127	300

Da qui si vede che su 152 maschi, 74 hanno risposto che preferirebbero se si chiedesse loro di persona di uscire insieme.

TABELLA DI CONTINGENZA TRA SESSO E MODALITÀ D'INVITO PREFERITA
(TABELLA ORIZZONTALE DELLE PERCENTUALI)

		MODALITÀ D'INVITO PREFERITA			SOMMA
SESSO	FEMMINE	TELEFONO	E-MAIL	DI PERSONA	
	MASCHI	23%	41%	36%	100%
	SOMMA	25%	26%	49%	100%

DOVENDO
CONFRONTARE
DUE VARIABILI,
CI SIAMO CO-
STRUTTI ANCORA
UNA VOLTA UNA
TABELLA DI
CONTIN-
GENZA.



Da qui si vede che il 49% ($\frac{74}{152} \times 100$) dei 152 maschi preferirebbe se si chiedesse loro di persona di uscire insieme.

INTERESSANTE... LE
RAGAZZE TENDONO A
PREFERIRE LE PROPO-
STE VIA EMAIL...

...MENTRE I RAGAZZI
QUELLE FATTE DI
PERSONA.



IN ALTRE PAROLE, ESISTE UNA CORRELAZIONE TRA IL GENERE E LE PREFERENZE SUL MODO IN CUI RICEVERE UNA PROPOSTA.



CERTO!



IL COEFFICIENTE DI CRAMER!



PROPRIO NON CE LA FACCIO A FICCARMI IN TESTA PAROLE DEL GENERE!



NON DEVI ESSERE GENTILE A TUTTI I COSTI...

Calcoliamo il coefficiente di Cramer seguendo i passi da 1 a 5.



Passo 1

Impostiamo una tabella di contingenza. I valori in corrispondenza dei bordi evidenziati vengono detti "frequenze osservate".

		Modalità d'invito preferita			Somma
		Telefono	E-mail	Di persona	
Sesso	Femmine	34	61	53	148
	Maschi	38	40	74	152
Somma		72	101	127	300

Passo 2

Svolgiamo i calcoli nella tabella seguente. I valori in corrispondenza dei bordi evidenziati vengono detti "frequenze attese".

		Modalità d'invito preferita			Somma
		Telefono	E-mail	Di persona	
Sesso	Femmine	$\frac{148 \times 72}{300}$	$\frac{148 \times 101}{300}$	$\frac{148 \times 127}{300}$	148
	Maschi	$\frac{152 \times 72}{300}$	$\frac{152 \times 101}{300}$	$\frac{152 \times 127}{300}$	152
Somma		72	101	127	300

$$\frac{\text{somma di "maschi" + somma di "di persona"}}{\text{totale dei valori}}$$

Formula A

Se tra genere e modalità d'invito preferita per ricevere una proposta non ci fosse correlazione, i rapporti tra le frequenze attese delle tre modalità dovrebbero essere

$$72 : 101 : 127 = \frac{72}{72 + 101 + 127} : \frac{101}{72 + 101 + 127} : \frac{127}{72 + 101 + 127}$$

$$= \frac{72}{300} : \frac{101}{300} : \frac{127}{300}$$

sia per i maschi che per le femmine (stando alla colonna delle somme nella tabella del Passo 2). Pertanto, la frequenza attesa (Formula A) mostra che – nel caso in cui non vi sia nessuna correlazione tra il genere e il tipo di proposta – il numero dei maschi che preferirebbero una proposta personale è $152 \times (127 \div 300) = (152 \times 127) \div 300$, o

$$152 \times \frac{127}{300} = \frac{152 \times 127}{300} = 64.3$$



Passo 3

Per ogni riquadro calcoliamo $\frac{(frequenza\ osservata - frequenza\ attesa)^2}{frequenza\ attesa}$

		Modalità d'invito preferita			Somma
		Telefono	E-mail	Di persona	
Sesso	Femmine	$\frac{\left(34 - \frac{148 \times 72}{300}\right)^2}{\frac{148 \times 72}{300}}$	$\frac{\left(61 - \frac{148 \times 101}{300}\right)^2}{\frac{148 \times 101}{300}}$	$\frac{\left(53 - \frac{148 \times 127}{300}\right)^2}{\frac{148 \times 127}{300}}$	148
	Maschi	$\frac{\left(38 - \frac{152 \times 72}{300}\right)^2}{\frac{152 \times 72}{300}}$	$\frac{\left(40 - \frac{152 \times 101}{300}\right)^2}{\frac{152 \times 101}{300}}$	$\frac{\left(74 - \frac{152 \times 127}{300}\right)^2}{\frac{152 \times 127}{300}}$	152
Somma		72	101	127	300



Maggiore è il divario tra frequenze osservate e frequenze attese e maggiori sono i valori in ciascun riquadro.

Passo 4

Calcoliamo la somma dei valori all'interno del riquadro evidenziato nella tabella del Passo 3. Questo numero si chiama *Statistica chi-quadro di Pearson*. Da questo momento lo scriveremo così: χ_0^2

$$\begin{aligned}\chi_0^2 &= \frac{\left(34 - \frac{148 \times 72}{300}\right)^2}{\frac{148 \times 72}{300}} + \frac{\left(61 - \frac{148 \times 101}{300}\right)^2}{\frac{148 \times 101}{300}} + \frac{\left(53 - \frac{148 \times 127}{300}\right)^2}{\frac{148 \times 127}{300}} \\ &+ \frac{\left(38 - \frac{152 \times 72}{300}\right)^2}{\frac{152 \times 72}{300}} + \frac{\left(40 - \frac{152 \times 101}{300}\right)^2}{\frac{152 \times 101}{300}} + \frac{\left(74 - \frac{152 \times 127}{300}\right)^2}{\frac{152 \times 127}{300}} \\ &= 8.0091\end{aligned}$$

Dal Passo 3 deduciamo che più le frequenze osservate divergono dalle frequenze attese (cioè: maggiore è la correlazione tra genere e modalità d'invito preferita per ricevere una proposta) e maggiore diventa il valore della statistica chi-quadro di Pearson (χ_0^2).



Passo 5

Calcoliamo il Coefficiente di Cramer.

$$\sqrt{\frac{\chi_0^2}{\text{numero dei valori } x}} \quad \chi_0^2 \\ \sqrt{\frac{8.0091}{300 \times \min\{2,3\} - 1}} = \sqrt{\frac{8.0091}{300 \times (2 - 1)}} = \sqrt{\frac{8.0091}{300}} = 0.1634$$

$\min\{a,b\}$ sta per "il più piccolo tra i due valori a, b "

$$\sqrt{\frac{8.0091}{300 \times \min\{2,3\} - 1}} = \sqrt{\frac{8.0091}{300 \times (2 - 1)}} = \sqrt{\frac{8.0091}{300}} = 0.1634$$

QUINDI IL COEFFICIENTE DI CRAMER È 0,1634.

AUTO... MI
GIRA LA
TESTA.

NON CREDO DI
POTER FARE QUESTI
CALCOLI DA SOLA.

ANDRÀ TUTTO
BENE! È UN PO'
COMPLICATO MA
SE LI ESEGUI PAS-
SO DOPO PASSO
CE LA FARAI.

E SE C'È
QUALCOSA CHE
NON CAPISCI, BA-
STA CHE TU ME LO
CHIEDA, OKAY?

!

SGUARDO
FISSO

CHE
C'È?

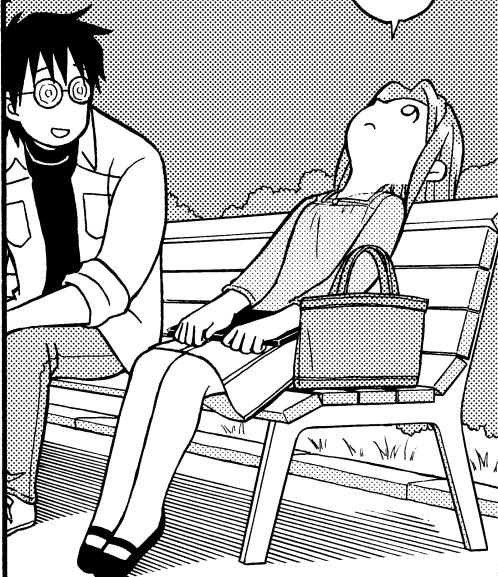
OH!

QUALCOSA
NON VA?

NO, NIENTE.

GRATT.
GRATT.

PER UN
ISTANTE MR
YAMAMOTO MI
È SEMBRATO
addirittura
CARINO.



Come abbiamo visto, il Coefficiente di Cramer varia tra 0 e 1. Più due variabili sono correlate e più il coefficiente è vicino a 1. Più la correlazione è debole e più il coefficiente si avvicina a 0. Per qualche dettaglio in più, potete osservare questi esempi di tabelle di contingenza delle percentuali.



Questa è la tabella di contingenza delle percentuali nel caso in cui il Coefficiente di Cramer vale 1.

		Modalità d'invito preferita			Somma
		Telefono	E-mail	Di persona	
Sesso	Femmina	17%	83%	0%	100%
	Maschio	0%	0%	100%	100%

Il Coefficiente di Cramer è 1 \Leftrightarrow le preferenze di maschi e femmine sono totalmente diverse.

Questa è la tabella di contingenza delle percentuali nel caso in cui il Coefficiente di Cramer vale 0.

		Modalità d'invito preferita			Somma
		Telefono	E-mail	Di persona	
Sesso	Femmine	17%	48%	35%	100%
	Maschi	17%	48%	35%	100%

Il Coefficiente di Cramer è 0 \Leftrightarrow le preferenze di maschi e femmine sono le stesse.



Ahimè, non esistono veri e propri standard statistici del tipo "due variabili sono fortemente correlate se il Coefficiente di Cramer è al di sopra di un certo valore". Ma alcuni standard informali in uso sono i seguenti.

VALORI CONVENZIONALI DEL COEFFICIENTE DI CRAMER

Valore	Valutazione	Valutazione di massima
1.0-0.8	⇒ Molto fortemente correlati	
0.8-0.5	⇒ Fortemente correlati	Correlati
0.5-0.25	⇒ Debolmente correlati	
Meno di 0,25	⇒ Molto debolmente correlati	Non correlati

IN CONCLUSIONE, RELATIVAMENTE AL NOSTRO ESEMPIO, POSSIAMO DIRE CHE LA CORRELAZIONE È ESTREMAMENTE DEBOLE.

E QUESTO È TUTTO PER OGGI.

CAPISCO.

GRAZIE.

IN SINTESI, ABBIAMO PARLATO DEL COEFFICIENTE DI CRAMER.

A PARTIRE DA QUESTI ARGOMENTI, NELLA PROSSIMA LEZIONE AFFRONTEREMO I TEST D'INDIPENDENZA.

TEST D'INDIPENDENZA?

I TEST D'INDIPENDENZA SONO AMPIAMENTE UTILIZZATI NELL'ANALISI DEI DATI STATISTICI.

UNA VOLTA CHE LI AVRAI CAPITI BENE, SARAI IN POSSESSO DELLE BASI DELLA STATISTICA.

QUESTO VUOL DIRE CHE LA PROSSIMA LEZIONE SARÀ L'ULTIMA?

PER IL MOMENTO, SÌ.

ERA ORA!

ESERCIZIO CON SOLUZIONE



ESERCIZIO

La società X gestisce un ristorante che ultimamente non sta andando bene. La società decide di analizzare le necessità dei clienti e commissiona un sondaggio tra persone di nazionalità giapponese selezionate a caso, di almeno 20 anni di età. La tabella riassume i risultati del sondaggio.

Intervistato	Che tipo di menù preferite in questo tipo di ristorante?	Alla fine del pasto vi viene offerta una bevanda gratuita. Che cosa preferite?
1	Cinese	Caffè
2	Europeo	Caffè
...
250	Giapponese	Tè

Ecco una tabella di contingenza derivata da quella precedente

		Preferenza per caffè o tè		Somma
		Caffè	Tè	
Tipo di menù ordinato più spesso	Giapponese	43	33	76
	Europeo	51	53	104
	Cinese	29	41	70
Somma		123	127	250

Calcolate il Coefficiente di Cramer per il tipo di menù preferito in un ristorante e la bevanda omaggio preferita.

SOLUZIONE

Passo 1

Compilate la tabella di contingenza.

		Preferenza per caffè o tè		Somma
		Caffè	Tè	
Tipo di menù ordinato più spesso	Giapponese	43	33	76
	Europeo	51	53	104
	Cinese	29	41	70
	Somma	123	127	250

Passo 2

Calcolate la frequenza attesa.

		Preferenza per caffè o tè		Somma
		Caffè	Tè	
Tipo di menù ordinato più spesso	Giapponese	$\frac{76 \times 123}{250}$	$\frac{76 \times 127}{250}$	76
	Europeo	$\frac{104 \times 123}{250}$	$\frac{104 \times 127}{250}$	104
	Cinese	$\frac{70 \times 123}{250}$	$\frac{70 \times 127}{250}$	70
	Somma	123	127	250

Passo 3

Calcolate

$$\frac{(\text{frequenza osservata} - \text{frequenza attesa})^2}{\text{frequenza attesa}}$$

per ciascun riquadro

		Preferenza per caffè o tè		Somma
		Caffè	Tè	
Tipo di menù ordinato più spesso	Giapponese	$\left(43 - \frac{76 \times 123}{250}\right)^2$ $\frac{76 \times 123}{250}$	$\left(33 - \frac{76 \times 127}{250}\right)^2$ $\frac{76 \times 127}{250}$	76
	Europeo	$\left(51 - \frac{104 \times 123}{250}\right)^2$ $\frac{104 \times 123}{250}$	$\left(53 - \frac{104 \times 127}{250}\right)^2$ $\frac{104 \times 127}{250}$	104
	Cinese	$\left(29 - \frac{70 \times 123}{250}\right)^2$ $\frac{70 \times 123}{250}$	$\left(41 - \frac{70 \times 127}{250}\right)^2$ $\frac{70 \times 127}{250}$	70
	Somma	123	127	250

Passo 4

Calcolate la somma dei valori all'interno dei riquadri evidenziati nella tabella al passo 3, cioè il valore della statistica chi-quadrato di Pearson (χ_0^2).

$$\begin{aligned}\chi_0^2 &= \frac{\left(43 - \frac{76 \times 123}{250}\right)^2}{\frac{76 \times 123}{250}} + \frac{\left(33 - \frac{76 \times 127}{250}\right)^2}{\frac{76 \times 127}{250}} \\ &\quad + \frac{\left(51 - \frac{104 \times 123}{250}\right)^2}{\frac{104 \times 123}{250}} + \frac{\left(53 - \frac{104 \times 127}{250}\right)^2}{\frac{104 \times 127}{250}} \\ &\quad + \frac{\left(29 - \frac{70 \times 123}{250}\right)^2}{\frac{70 \times 123}{250}} + \frac{\left(41 - \frac{70 \times 127}{250}\right)^2}{\frac{70 \times 127}{250}} \\ &= 3.3483\end{aligned}$$

Passo 5

Calcolate il coefficiente di Cramer.

$$\sqrt{\frac{\chi_0^2}{\text{numero dei valori} \times (\min\{\text{numero delle righe della tabella di contingenza}, \text{numero delle colonne della tabella di contingenza}\} - 1)}}$$
$$\sqrt{\frac{3.3483}{250 \times (\min\{3, 2\} - 1)}} = \sqrt{\frac{3.3483}{250 \times (2 - 1)}} = \sqrt{\frac{3.3483}{250}} = 0.1157$$

RIASSUMENDO



- L'indice che misura il grado di correlazione tra due tipi di dati numerici è il *coefficiente di correlazione*.
- L'indice che misura il grado di correlazione tra dati numerici e dati categorici è il *rapporto di correlazione*.
- L'indice che misura il grado di correlazione tra due tipi di dati categorici è il *Coefficiente di Cramer* (detto anche la *V di Cramer* o *coefficiente indipendente*).
- La tabella seguente riassume le caratteristiche dei tre indici.

	Minimo	Massimo	Valore quando le due variabili sono del tutto scorrelate	Valore quando le due variabili sono correlate in massimo grado
Coefficiente di correlazione	-1	1	0	-1 or 1
Rapporto di correlazione	0	1	0	1
Coefficiente di Cramer	0	1	0	1

- Non esistono standard statistici per i valori assunti da *coefficiente di correlazione*, *rapporto di correlazione* e *Coefficiente di Cramer* nella forma “le due variabili sono fortemente correlate se il valore è al di sopra di una certa soglia”.

7

APPROFONDIAMO UN PO' I TEST D'IPOTESI STATISTICA

1. TEST D'IPOTESI STATISTICA

EHI! NON MI DICI NIENTE?

ALLO-
RA, LA
LEZIONE
DI OGGI
SARÀ...

TA-TA!

OH, BE', ECCO, CIOÈ,
SCUSA SE NON ME
N'ERO ACCORTO. È
QUELLA LA NUOVA
UNIFORME DI CUI
PARLAVI L'ALTRO
GIORNO?

ESATTO! È ANCORA
UN MODELLO DI
PROVA, MA IN VIA
ECCEZIONALE PUOI
VEDERLO IN ANTE-
PRIMA.

BE', GRAZIE...
TI STA MOLTO
BENE.

GRAZIE.

ALLORA,
COS'ABBIAMO
OGGI A
LEZIONE?

RICORDI L'ULTIMA VOLTA, QUANDO ABBIANO PARLATO DEL COEFFICIENTE DI CRAMER?

* ABBIAMO CHIESTO A 300 LICEALI "COME PREFERISCI CHE TI VENGA CHIESTO DI USCIRE?"

高校生300人にさました
告白するといたら
どうの方法でやれたい!

VUOI DIRE QUEL SONDAGGIO SU COME CHIEDERE DI USCIRE?

NELL'ESEMPIO, IL COEFFICIENTE DI CRAMER ERA 0,1634 E QUESTO CORRISPONDEVÀ A UNA CORRELAZIONE PIUTTOSTO DEBOLE.

ORA STAI BENE ATTENTA, RUI.

IL SONDAGGIO FU EFFETTUATO SU UN CAMPIONE DI 300 PERSONE...

SÌ, ME LO RICORDO.

...SELEZIONATE A CASO TRA TUTTI I LICEALI GIAPPONESI.

SCEGLIENDO 300 PERSONE DIVERSE AVREMMO OTTENUTO UN DIVERSO COEFFICIENTE DI CRAMER... NON SAREBBE STATO 0,1634.

ORA CHE CI PENSO, HAI RAGIONE.

HAI IDEA DI QUALE
POTREBBE ESSERE
IL COEFFICIENTE DI
CRAMER PER L'INTERA
POPOLAZIONE DEL NO-
STRO ESEMPIO, CIOÈ
PER TUTTI I LICEALI DEL
GAIPPONE?

E COME PO-
TREI MAI SA-
PERLO?

UNA RISPOSTA NATURALE.

SFORTUNATAMENTE,
NESSUNO PUÒ SAPERLO,
PERCHÉ È IMPOSSIBILE
SOTTOPORRE AL SON-
DAGGIO OGNI SINGOLO
STUDENTE GIAPPONESE.

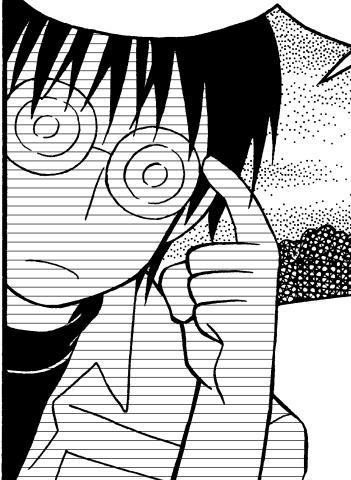
UH-UH.

NON CAPITA SOLO IN
QUESTO ESEMPIO,
OVIAMENTE. IN GENE-
RALE, NON È POSSIBILE
CALCOLARE IL COEFFI-
CIENTE DI CRAMER PER
L'INTERA POPOLAZIONE.

PER QUESTO
NON ABBIAMO
ALTRA SCELTA
CHE...

...EFFETTUARE UNA
SCELTA INFORMA-
TA A PROPOSITO
DEL COEFFICIENTE,
COME PER ESEM-
PIO...

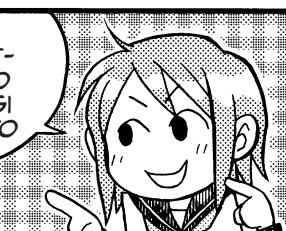
"POICHÉ IL VALORE OTTENUTO CON UN
CAMPIONE DI 300 PERSONE SELEZIONATE
CASUALMENTE È 0,1634..."



...IL COEFFICIENTE
CALCOLATO RELATIVAMENTE
ALL'INTERA POPOLAZIONE
NON SI DISCOSTERÀ MOLTO
DA QUESTO VALORE."



NON SUO-
NA AFFATTO
CONVINCENTE.

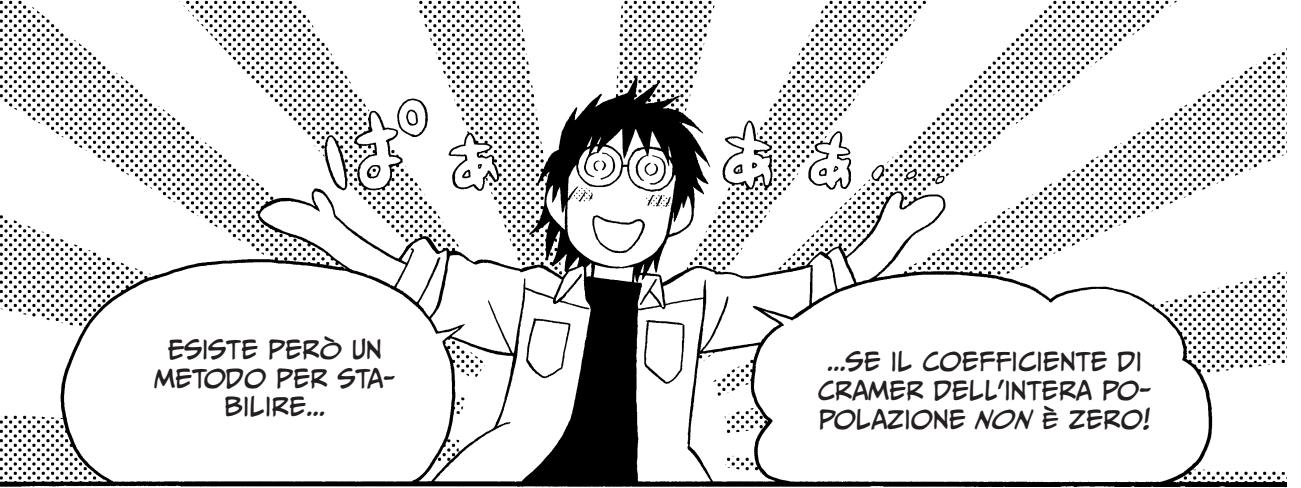


COME IN ALTRI
CASI, SCOMMET-
TO CHE USANDO
LA STATISTICA SI
PUÒ FARE MOLTO
MEGLIO.
ピ"!"



qualsiasi tecnica sta-
tistica utiliziamo,
resterà impossibile
calcolare l'effettivo
coefficiente di Cramer,
perché non potremo
mai intervistare ogni
singolo componente
della popolazione.
Quindi... mi dispiace de-
luderti ma la
risposta è "no".

NO?!



ESISTE PERÒ UN
METODO PER STA-
BILIRE...

...SE IL COEFFICIENTE DI
CRAMER DELL'INTERA PO-
POLAZIONE NON È ZERO!



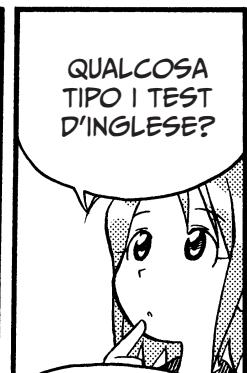
NON MI SEMBRA 'STA
GRAN COSA...

CERTO CHE LO È. È
UN DATO OGGETTIVO
RELATIVO ALL'INTERA
POPOLAZIONE.



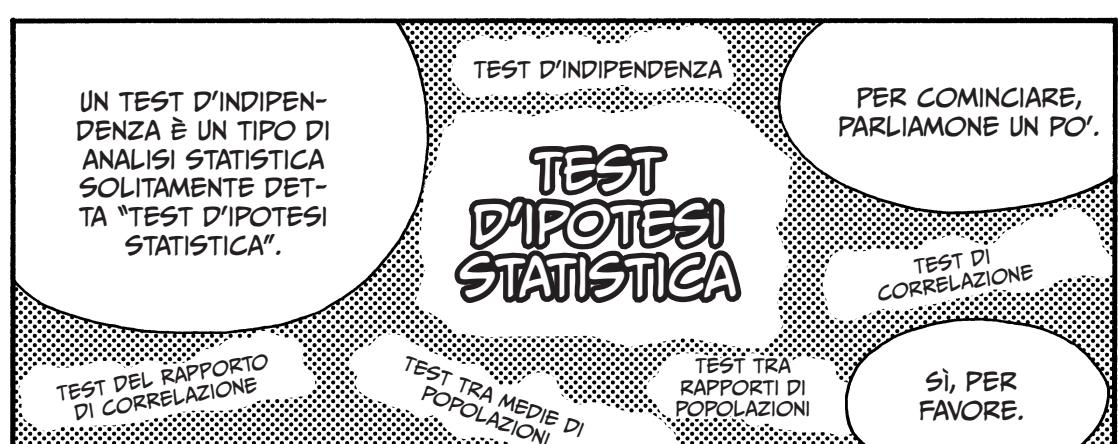
E COME SI
FA?

NELLA LEZIONE
PRECEDENTE HO
ANTICIPATO CHE OGGI
STUDIEREMO I TEST
D'INDIPENDENZA.
SONO GLI STRUMENTI
CHE UTILIZZEREMO.



QUALCOSA
TIPO I TEST
D'INGLESE?

NO, NO, È UNA
COSA COM-
PLETAMENTE
DIVERSA.



UN TEST D'INDIPEN-
DENZA È UN TIPO DI
ANALISI STATISTICA
SOLITAMENTE DET-
TA "TEST D'IPOTESI
STATISTICA".

TEST D'INDIPENDENZA

PER COMINCIARE,
PARLIAMONE UN PO'.

TEST D'IPOTESI STATISTICA

TEST DEL RAPPORTO
DI CORRELAZIONE

TEST TRA MEDIE DI
POPOLAZIONI

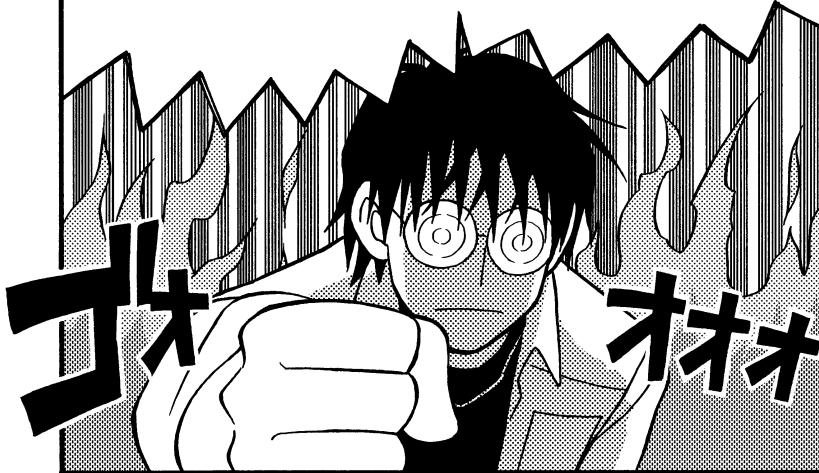
TEST TRA
RAPPORTI DI
POPOLAZIONI

TEST DI
CORRELAZIONE

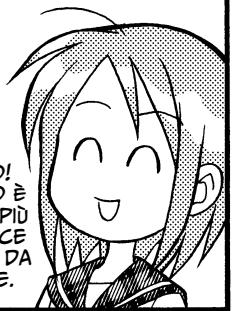
SÌ, PER
FAVORE.

UN TEST D'IPOTESI CERCA DI CAPIRE SE L'ANALISTA HA FATTO IPOTESI CORRETTE SULLA POPOLAZIONE A PARTIRE DA UN DETERMINATO CAMPIONE.

IL NOME FORMALE È TEST DI VERIFICA D'IPOTESI.



BRAVO!
QUESTO È
MOLTO PIÙ
SEMPLICE
PER ME DA
CAPIRE.



ESISTONO DIVERSI TIPI DI TEST D'IPOTESI STATISTICA.

ESEMPI DI TEST D'IPOTESI STATISTICA

Nome	Esempi di utilizzo
Test d'indipendenza	Stima se il valore del Coefficiente di Cramer per genere e preferenze nell'essere invitati sia 0 per una popolazione.
Test del rapporto di correlazione	Stima se il valore del rapporto di correlazione per l'età e il marchio di moda preferito è 0 per una popolazione.
Test di correlazione	Stima se il coefficiente di correlazione tra la spesa in cosmetici e quella in abbigliamento è 0 per una popolazione.
Test tra medie di popolazioni	Stima se le paghette sono diverse tra le liceali di Osaka e quelle di Tokyo*.
Test di rapporti tra popolazioni	Stima se il tasso di popolarità del Governo X è diverso tra elettori delle aree urbane e delle aree rurali*.

* Si noti che vengono considerate due popolazioni.

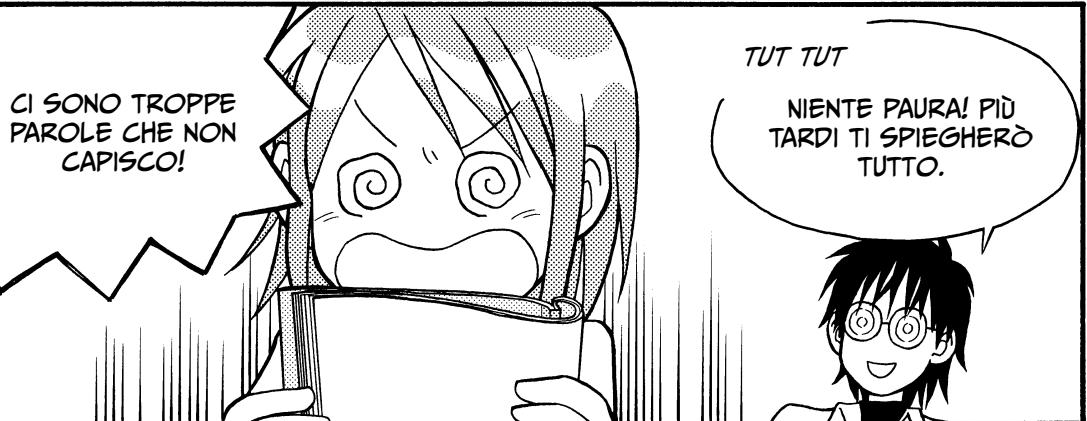


ESISTONO TANTI TIPI DI TEST D'IPOTESI STATISTICA MA SEGUONO TUTTI LA MEDESIMA PROCEDURA.

SAPERLO È UN VERO PIACERE.

PROCEDIMENTO PER UN TEST DI VERIFICA D'IPOTESI

-
- Passo 1** Definire la popolazione.
-
- Passo 2** Definire un'ipotesi nulla e un'ipotesi alternativa.
-
- Passo 3** Scegliere il test d'ipotesi statistica da effettuare.
-
- Passo 4** Fissare un livello di significatività.
-
- Passo 5** Calcolare la statistica relativa ai dati.
-
- Passo 6** Stabilire se la statistica del passo 5 ricade nella regione critica.
-
- Passo 7** Se nel passo 6 la statistica ricade nella regione critica, scartare l'ipotesi nulla. Viceversa, non si potrà scartarla.
-



2. IL TEST D'INDIPENDENZA CHI-QUADRO

L'ARGOMENTO PRINCIPALE DI OGGI È UN TEST D'INDIPENDENZA.



COME HO DETTO, UN TEST D'INDIPENDENZA È UNA TECNICA D'ANALISI DEI DATI IMPIEGATA PER STIMARE SE IL COEFFICIENTE DI CRAMER DI UNA POPOLAZIONE È ZERO.

È VERO,
L'HA
DETTO.

IN ALTRI TERMINI, È UNA TECNICA PER STABILIRE SE LE DUE VARIABILI DI UNA TABELLA DI CONTINGENZA SONO CORRELATE.

		MODALITÀ D'INVITO PREFERITA			SOMMA
		TELEFONO	E-MAIL	DI PERSONA	
SESSO	F	34	61	53	
	M	38	40	74	148
SOMMA		72	101	127	152

QUESTO PARTICOLARE TEST SI CHIAMA TEST CHI-QUADRO.

OH, NO!
ANCORA QUEL
CHI-QUALCOSA!

SPIEGAZIONE

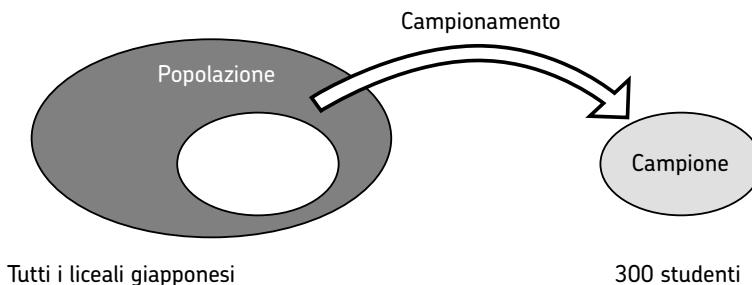
La statistica chi-quadro di Pearson (χ^2_0) e la distribuzione chi-quadro



Prima di un effettivo esempio di test d'indipendenza, vorrei spiegare una cosa davvero fondamentale. Supponiamo di effettuare il seguente esperimento, che nella realtà è impossibile.

Passo 1

Selezioniamo un campione casuale di 300 studenti dalla popolazione “tutti i liceali giapponesi”.



Passo 2

Sulle 300 persone selezionate al Passo 1, conduciamo il sondaggio di pagina 127 per ricavarne χ^2_0 .

Passo 3

Reintegriamo le 300 persone nella popolazione.

Passo 4

Ripetere i passi da 1 a 3 più volte.

In questo esperimento, se il valore del Coefficiente di Cramer per la popolazione “tutti i liceali giapponesi” è 0, allora il grafico del valore della statistica chi-quadro di Pearson (χ^2_0) risulterà essere una distribuzione chi-quadro con due gradi di libertà.

- V. alle pagine 130-133 come ricavare il valore della statistica chi-quadro di Pearson (χ^2_0).
- V. pagina 100 per dettagli sulla distribuzione chi-quadro con due gradi di libertà.

Abbiamo effettivamente svolto l'esperimento, con le seguenti limitazioni.



- È impossibile coinvolgere l'intera popolazione "tutti i liceali giapponesi", che quindi identificheremo con le 10.000 persone della tabella 7.1.
- Supporremo che il Coefficiente di Cramer per "tutti i liceali giapponesi" sia 0. Questo significa che le proporzioni (o i rapporti) tra chi preferisce essere invitato nei vari modi è lo stesso per maschi e femmine
- (v. pagina 135). La tabella di contingenza di 7.1 è la 7.2.
- Ripeteremo i passi 1-3 per 10.000 volte.

TABELLA 7.1 - PREFERENZA SUL MODO DI ESSERE INVITATI

(TUTTI I LICEALI GIAPPONESI)

Intervistato	Sesso	Preferenza
1	Femmine	Di persona
2	Femmine	Telefono
...
10.000	Maschi	E-mail

TABELLA 7.2 - TABELLA DI CONTINGENZA TRA SESSO E PREFERENZA SUL MODO DI ESSERE INVITATI.

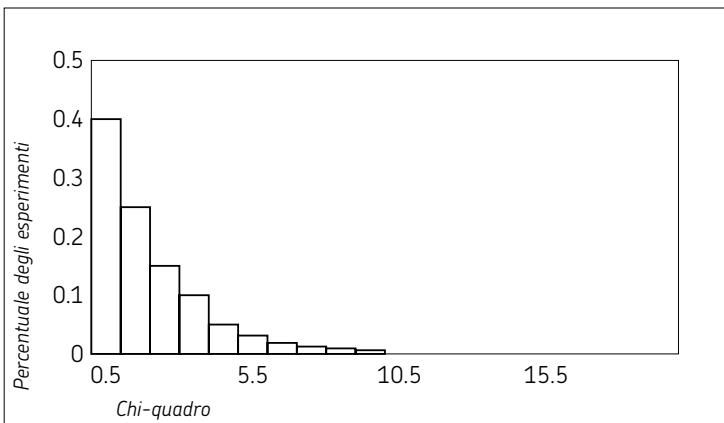
		Preferenza			Somma
		Telefono	E-mail	Di persona	
Sesso	Femmine	400	1.600	2.000	4.000
	Maschi	600	2.400	3.000	6.000
Somma		1.000	4.000	5.000	10.000

La tabella 7.3 illustra i risultati dell'esperimento. La Figura 7.1 è un istogramma costruito a partire dalla Tabella 7.3.

TABELLA 7.3 – RISULTATI DELL'ESPERIMENTO

Esperimento	La statistica chi-quadro di Pearson (χ^2_0)
1	0.8598
2	0.7557
...	...
10,000	2.7953

FIGURA 7.1 – ISTOGRAMMA DELLA TABELLA 7.3 (AMPIEZZA DI CLASSE=1)



La figura 7.1 assomiglia molto al grafico a pagina 100 “Due gradi di libertà” e sembra quindi corretto affermare che il valore della statistica chi-quadro di Pearson (χ^2_0) segue una distribuzione chi-quadro con due gradi di libertà. Anche se non ha nulla a che vedere con l'esperimento in sé, osserviamo una cosa: i due gradi di libertà vengono da

$$(2 - 1) \times (3 - 1) = 1 \times 2 = 2$$

↑ ↑
2 valori:
maschi e femmine 3 valori:
telefono, e-mail, e di persona



Non spiegherò il perché di questo strano calcolo: sarebbe un argomento troppo avanzato per questo libro. Ma non preoccupatevi, non sarà un problema.

RIASSUMENDO: OSSERO QUINDI CHE IL VALORE DEL COEFFICIENTE DI CRAMER PER "TUTTI I LICEALI GIAPPONESI" È ZERO... E QUESTO VUOL DIRE CHE TRA IL GENERE E LA MODALITÀ DI INVITO PREFERITA NON C'È NESSUNA CORRELAZIONE.

CREDO...

I RAPPORTI TRA LE PREFERENZE SONO GLI STESSI PER I MASCHI E PER LE FEMMINE!

POI EFFETTUO IL SONDAGGIO TRA 300 STUDENTI SELEZIONATI TRA "TUTTI I LICEALI GIAPPONESI".

POI LO RIPETO UN SACCO DI VOLTE!

sondaggi
sondaggi

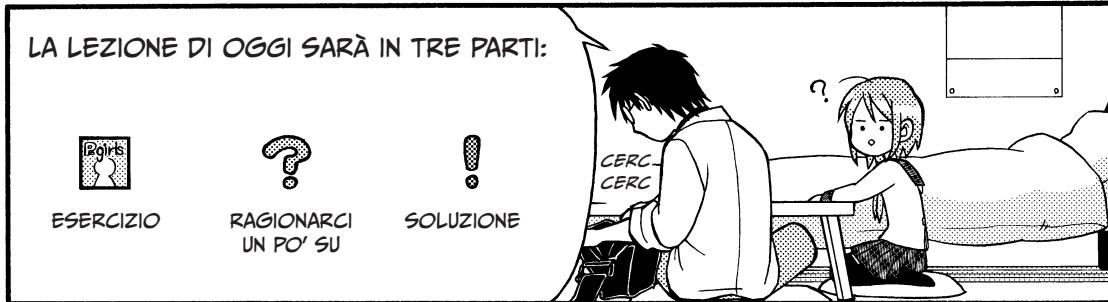
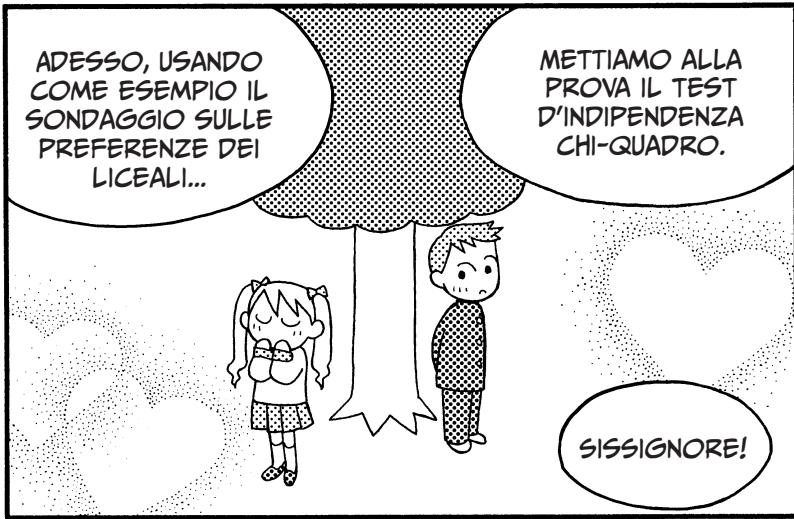
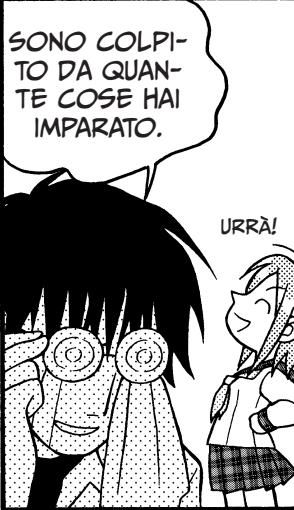
E ALLA FINE CALCOLO IL VALORE DELLA STATISTICA CHI-QUADRO DI PEARSON (χ^2_0)

$$\text{SOMMA DI IN CIASCUN Riquadro} \frac{(\text{FREQUENZA OSSERVATA} - \text{FREQUENZA ATTESA})^2}{\text{FREQUENZA ATTESA}}$$



IL GRAFICO DEI RISULTATI È UNA DISTRIBUZIONE CHI-QUADRO CON DUE GRADI DI LIBERTÀ!

FINALMENTE...
...HO LA SOLUZIONE!



La rivista *P-Girl Magazine* ha pubblicato un articolo intitolato "Abbiamo chiesto a 300 liceali 'Come vorreste che vi chiedessero di uscire?'". Per documentare l'articolo, un giornalista ha selezionato 300 studenti a caso tra tutti i liceali giapponesi: la tabella riassume i risultati del sondaggio.

Intervistato	Preferenza	Età	Sesso
1	Di persona	17	Femmine
2	Telefono	15	Femmine
...
300	E-mail	18	Maschi

Questa invece è la tabella di contingenza tra genere e preferenza.

		Modalità d'invito preferita			Somma
		Telefono	E-mail	Di persona	
Sesso	Femmine	34	61	53	148
	Maschi	38	40	74	152
Somma		72	101	127	300

Col test d'indipendenza chi-quadro, stimiamo se il Coefficiente di Cramer sia maggiore di 0. Questo equivale a stimare con un test d'indipendenza se genere e preferenza sono correlati. Ricordate che il livello di significatività (che vedremo più avanti) è 0,05.

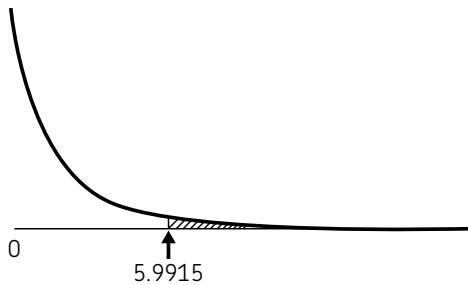


PENSIAMOCI UN PO' SU

Come spiegato alle pagine 152-154, il valore della statistica chi-quadro di Pearson (χ^2_0) segue una distribuzione chi-quadro con due gradi di libertà, nell'ipotesi nulla che il valore del Coefficiente di Cramer per la popolazione "tutti i liceali giapponesi" sia 0. Se questo è vero, allora la probabilità che il valore di χ^2_0 ricavato dalle 300 persone del campione sia 5,9915, o superiore, è 0,005.



FIGURA 7.2 – PROBABILITÀ CHE χ^2_0 SIA 5,9915, O SUPERIORE



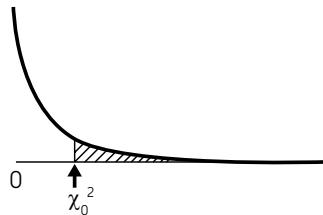
Questo è evidente dalla tabella della distribuzione chi-quadro di pagina 103. Il valore di χ^2_0 per questo esercizio è già stato calcolato a pagina 132, ed è 8,0091. Certo, è relativo al campione di 300 studenti scelti a caso, ma non vi sembra un po' troppo grande? Ricordando il commento a pagina 132, non sarebbe naturale assumere che il Coefficiente di Cramer dell'intera popolazione "tutti i liceali giapponesi" sia maggiore di 0?

Ricordate che in generale (non solo in questo esercizio) per il test d'indipendenza chi-quadro si procede in questo modo:

1. Si assume temporaneamente, come ipotesi nulla, che "il Coefficiente di Cramer della popolazione è 0".
2. Si calcola il valore di χ^2_0 relativo ai dati raccolti.
3. Se χ^2_0 è troppo grande, si può scartare l'ipotesi nulla e concludere che "il Coefficiente di Cramer della popolazione è maggiore di 0".

All'aumentare di χ^2 , la probabilità rappresentata dalla superficie tratteggiata in Figura 7.3 diminuisce.

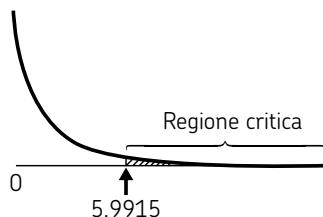
FIGURA 7.3 – PROBABILITÀ RELATIVA A χ^2_0



Nei test d'indipendenza chi-quadro, se la probabilità corrispondente alla superficie tratteggiata nella figura 7.3 è inferiore o uguale a un valore detto *soglia* o *livello di significatività* si scarta l'ipotesi nulla iniziale e si conclude che "il Coefficiente di Cramer della popolazione è maggiore di 0". In generale, il livello di significatività (detta anche *valore alfa* e indicata col simbolo α) è convenzionalmente posto a 0,05 o 0,01.

La scelta del livello di significatività sta all'analista. Supponiamo per esempio di usare il valore 0,05: il livello di significatività è in effetti la misura dell'area tratteggiata in figura 7.3, e l'intervallo dei valori maggiori o uguali a χ^2_0 viene detto *regione critica*.

FIGURA 7.4 – REGIONE CRITICA
(IN CORRISPONDENZA DEL LIVELLO DI SIGNIFICATIVITÀ 0,05)



! SOLUZIONE

Passo 1

Definiamo la popolazione.

La popolazione è:

TUTTI I LICEALI
GIAPPONESI!



In questo particolare esercizio, con questa definizione della popolazione, il Passo 1 non è necessario.

Ma nella tabella a pagina 149, le popolazioni erano definite così: "elettori nelle aree urbane" e "elettori nelle aree rurali".

Quali sono esattamente queste aree? Tokyo e Osaka? I capoluoghi delle prefetture e i rispettivi territori? Questo dev'essere chiarito dall'analista.

Teniamo sempre presente che quando si formula un test d'ipotesi statistica occorre definire con precisione la popolazione. È bene ripeterlo: di qualunque test si tratti, non potete mai esimervi dal definire la popolazione in maniera adeguata.

Altrimenti, il rischio è di ritrovarsi a domandarsi "Che cosa stavo cercando di stimare?". Prestate grande attenzione a questo punto, è una trappola in cui cadono molti statistici.

Passo 2

Definiamo un'ipotesi nulla e un'ipotesi alternativa.

Nel nostro caso, l'ipotesi nulla è "il Coefficiente di Cramer della popolazione è 0, cioè genere e modalità d'invito preferita sono scorrelati."

L'ipotesi alternativa è: "il Coefficiente di Cramer della popolazione è maggiore di 0. in altre parole, genere e modalità d'invito preferita sono correlati."



Trovate una discussione dell'ipotesi nulla e dell'ipotesi alternativa a pagina 170.

Passo 3

Scegiamo il test d'ipotesi statistica.

Farò il test d'indipendenza chi-quadro.



In questo caso particolare, sappiamo già che effettueremo il test d'indipendenza chi-quadro. In generale, sarete voi a dover decidere quale test d'ipotesi statistica è il più adatto allo scopo.



Passo 4

Fissiamo un livello di significatività.

Io come soglia scelgo 0,05.



In generale, dovete fissare di volta in volta un livello di significatività α . Come abbiamo già detto, due valori molto usati sono 0,05 e 0,01. Più il *p-value** ricavato dai dati è piccolo e più l'ipotesi nulla appare da scartare.

* Per la definizione di *p-value* v. pagina 175.

Passo 5

Calcoliamo la statistica relativa ai dati.

Effettuerò un test d'indipendenza chi-quadro, quidi la statistica è quella di Pearson (χ^2_0). Il suo valore per questo esercizio è già stato calcolato a pagina 132 ed è $\chi^2_0 = 8,0091$.



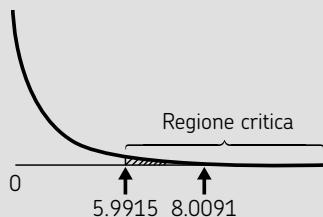
La statistica di un test è una funzione che dai dati del campione ricava un singolo valore. Test d'ipotesi statistiche diversi richiedono funzioni (o statistiche) diverse. Come abbiamo già detto, la statistica per un test d'indipendenza è χ^2_0 e nel caso di test di correlazione (v. pagina 149) la statistica è la seguente:

$$\frac{\text{coefficiente di correlazione}^2 \times \sqrt{\text{numero dei valori} - 2}}{1 - \sqrt{\text{coefficiente di correlazione}^2}}$$

Passo 6

Stabiliamo se il valore della statistica del passo 5 ricade nella regione critica.

Il valore della statistica chi-quadro di Pearson (χ^2_0) è 8,0091. Con una soglia di significatività di 0,05, dalla tabella della distribuzione chi-quadro di pagina 103 ricaviamo che la regione critica comincia a 5,9915. La figura mostra che il valore della statistica ricade nella regione critica.



La regione critica cambia a seconda del livello di significatività α . Se in questo esercizio scegliessimo $\alpha=0,01$, sempre per la tabella della distribuzione chi-quadro la regione critica partirebbe da 9,2104.

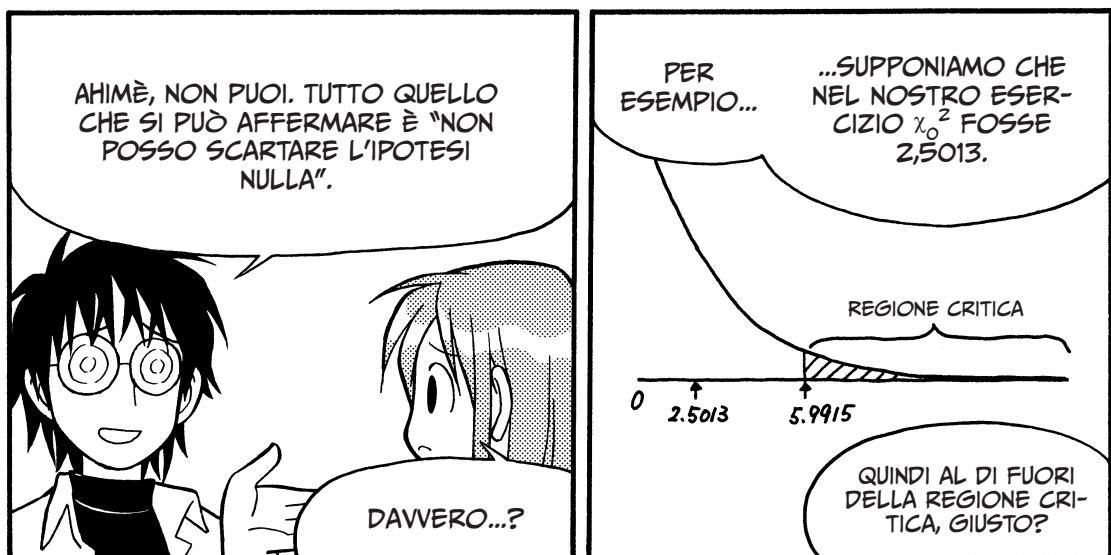
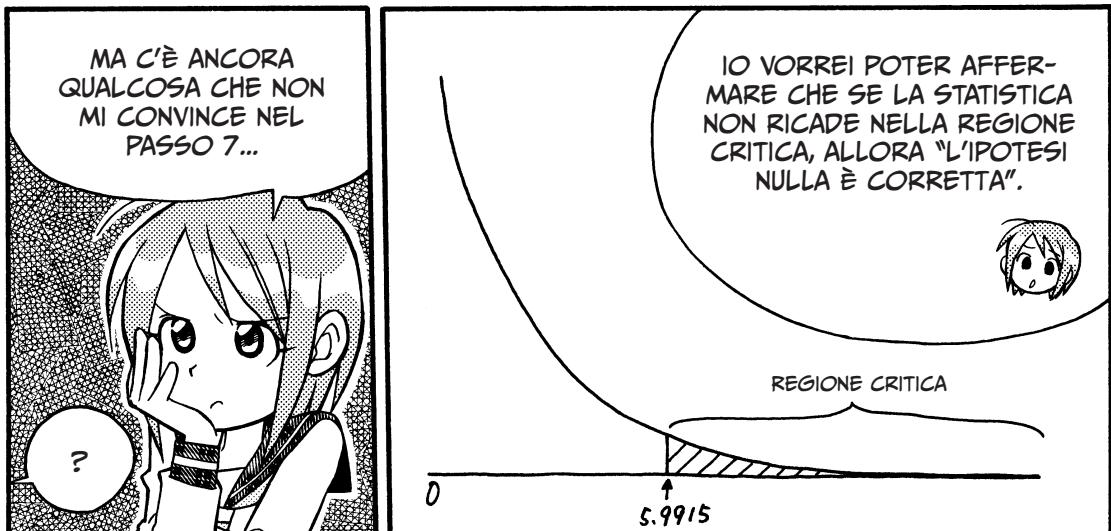
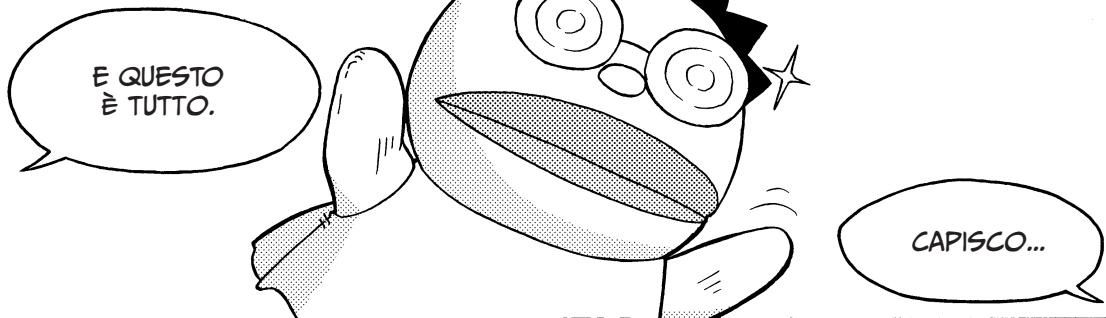
Passo 7

Se nel passo 6 la statistica ricade nella regione critica, scarteremo l'ipotesi nulla. Viceversa, non potremo scartarla.

Dunque è corretta l'ipotesi alternativa: "il Coefficiente di Cramer della popolazione è maggiore di 0"!



Anche se la statistica ricade nella regione critica, in un test d'ipotesi statistica non si può concludere che l'ipotesi alternativa è corretta in termini assoluti. L'unica conclusione possibile è "al massimo, la possibilità che l'ipotesi nulla sia corretta è di $(\alpha \times 100)\%$ "



DI REGOLA, NON PUOI SEMPLICEMENTE ACCETTARE L'IPOTESI ALTERNATIVA PER CUI "IL COEFFICIENTE DI CRAMER DELLA POPOLAZIONE È MAGGIORI DI ZERO".

MA NON È NEANCHE POSSIBILE FARLO CON L'IPOTESI NULLA "IL COEFFICIENTE DI CRAMER DELLA POPOLAZIONE È ZERO".

UNA BREVE STORIETTA DOVREBBE AIUTARTI A CAPIRE.

SUPPONIAMO CHE QUALCUNO TI RUBI UN BUDINO CHE VOLEVI MANGIARE PIÙ TARDI, E SE LO MANGI LUI.

CHI HA RUBATO IL MIO BUDINO?!

TRA I SOSPETTATI C'È YUMI.

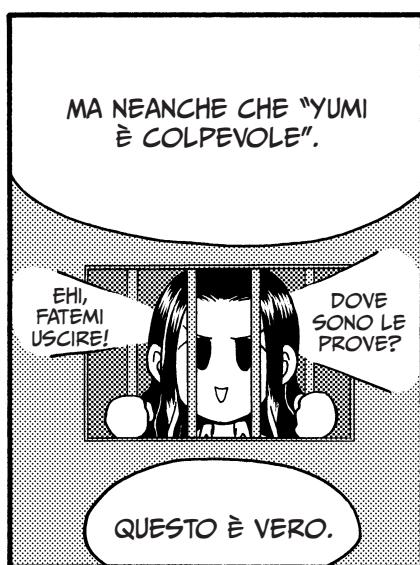
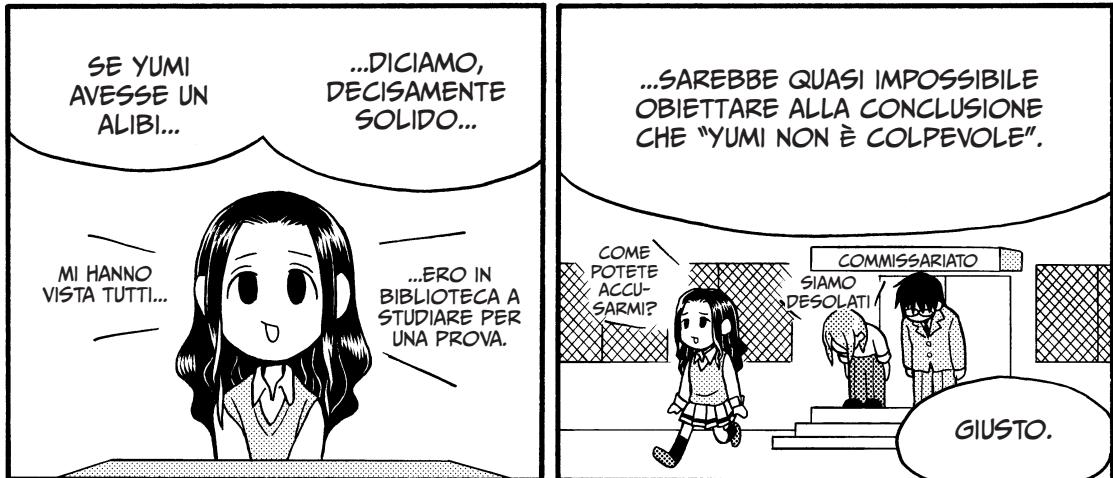
YUMI! E IO CHE TI CREDEVO UN'AMICA!

È SOLO UNA STORIA IMMAGINARIA!

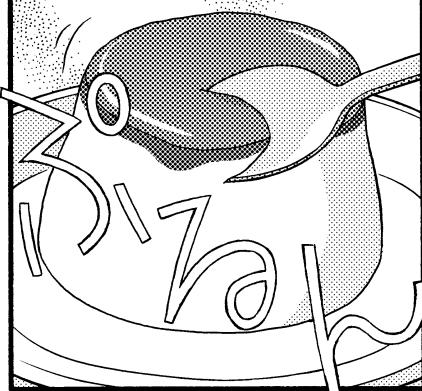
SVOLGIAMO UN TEST D'IPOTESI STATISTICA CON QUESTE DUE IPOTESI.

IPOTESI NULLA	YUMI NON È COLPEVOLE
IPOTESI ALTERNATIVA	YUMI È COLPEVOLE

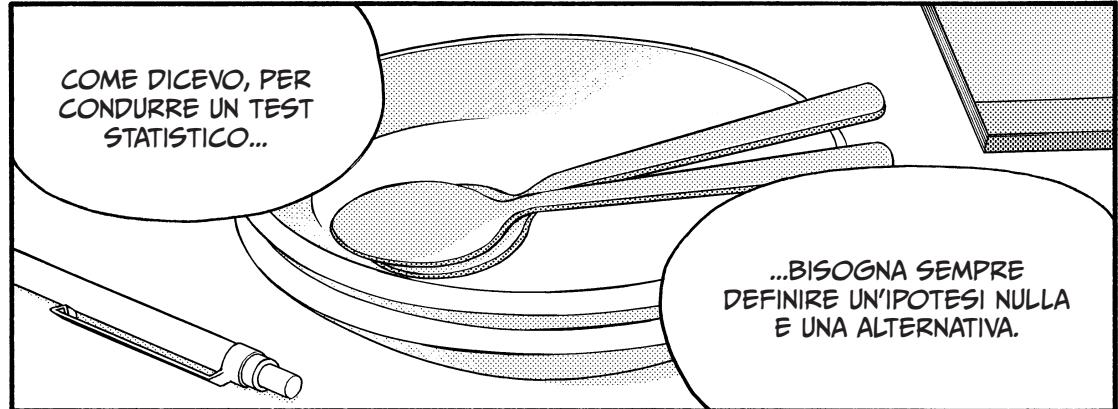
TRASCURIAMO DETTAGLI COME IL TIPO DI TEST D'IPOTESI STATISTICA O IL LIVELLO DI SIGNIFICATIVITÀ...



3. L'IPOTESI NULLA E QUELLA ALTERNATIVA



IL SUO ESEMPIO MI HA RICORDATO CHE IN FRIGO C'ERA DEL BUDINO.



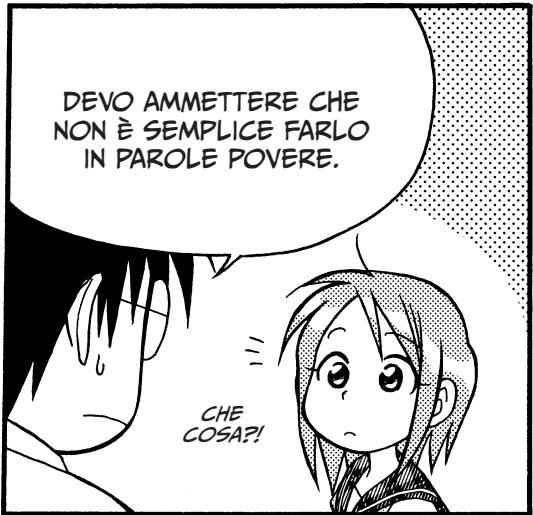
COME DICEVO, PER CONDURRE UN TEST STATISTICO...

...BISOGNA SEMPRE DEFINIRE UN'IPOTESI NULLA E UNA ALTERNATIVA.



D'ACCORDO, MA CHE COSA SONO ESATTAMENTE QUESTE DUE IPOTESI?

AVEVA DETTO CHE LO AVREBBE SPIEGATO, MA NON L'HA ANCORA FATTO.



DEVO AMMETTERE CHE NON È SEMPLICE FARLO IN PAROLE POVERE.

CHE COSA?!



ESEMPI DI TEST D'IPOTESI STATISTICA

Nome	Esempi di utilizzo
Test d'indipendenza	Stima se il valore del Coefficiente di Cramer per genere e preferenze nell'essere invitati sia 0 per una popolazione.
Test del rapporto di correlazione	Stima se il valore del rapporto di correlazione per l'età e il marchio di moda preferito è 0 per una popolazione.
Test di correlazione	Stima se il coefficiente di correlazione tra la spesa in cosmetici e quella in abbigliamento è 0 per una popolazione.
Test tra medie di popolazioni	Stima se le paghette sono diverse tra le liceali di Osaka e quelle di Tokyo*.
Test di rapporti tra popolazioni	Stima se il tasso di popolarità del Governo X è diverso tra elettori delle aree urbane e delle aree rurali*.

* Si noti che vengono considerate due popolazioni.



TEST D'INDIPENDENZA

Ipotesi nulla	il Coefficiente di Cramer per genere e preferenze nell'essere invitati è 0 per la popolazione.
Ipotesi alternativa	il Coefficiente di Cramer per genere e preferenze nell'essere invitati è maggiore di 0 per la popolazione.

TEST DEL RAPPORTO DI CORRELAZIONE

Ipotesi nulla	Il rapporto di correlazione per età e marchio preferito è 0 per la popolazione.
Ipotesi alternativa	Il rapporto di correlazione per età e marchio preferito è maggiore di 0 per la popolazione.

TEST DI CORRELAZIONE

Ipotesi nulla	Il coefficiente di correlazione tra le spese per cosmetici e per abbigliamento è 0 per la popolazione.
Ipotesi alternativa	Il coefficiente di correlazione tra le spese per cosmetici e per abbigliamento non è 0 per la popolazione. oppure Il coefficiente di correlazione tra le spese per cosmetici e per abbigliamento è maggiore di 0 per la popolazione. oppure Il coefficiente di correlazione tra le spese per cosmetici e per abbigliamento è minore di 0 per la popolazione.

TEST TRA MEDIE DI POPOLAZIONI

Ipotesi nulla	Le paghette delle liceali di Tokyo e di Osaka sono uguali.
Ipotesi alternativa	Le paghette delle liceali di Tokyo e di Osaka non sono uguali. oppure Le paghette delle liceali di Tokyo sono maggiori di quelle delle liceali di Osaka. oppure Le paghette delle liceali di Tokyo sono inferiori a quelle delle liceali di Osaka.

TEST TRA RAPPORTI DI POPOLAZIONI

Ipotesi nulla	Il tasso di popolarità del Governo X tra elettori delle aree urbane e delle aree rurali è lo stesso.
Ipotesi alternativa	Il tasso di popolarità del Governo X tra elettori delle aree urbane e delle aree rurali non è lo stesso. oppure Il tasso di popolarità del Governo X tra elettori delle aree urbane è maggiore di quello tra gli elettori delle aree rurali. oppure Il tasso di popolarità del Governo X tra elettori delle aree urbane è minore di quello tra gli elettori delle aree rurali.

UN'OTTIMA
SPIEGAZIONE!



RIESCI A SEGUIRE?

SPERO CHE ORA SIA CHIARO PERCHÉ "IL COEFFICIENTE DI CRAMER DELLA POPOLAZIONE È 'QUASI' ZERO" NON È L'UNICA IPOTESI NULLA POSSIBILE, E NE ESISTONO ALTRE DIFFICILI DA DIMOSTRARE, COME PER ESEMPIO "IL COEFFICIENTE DI CRAMER È ZERO".

SÌ, ALCUNE SONO DAVVERO SPECIFICHE.



INOLTRE, HAI NOTATO COME LE IPOTESI NULLE SIANO FORMULATE COME AFFERMAZIONI POSITIVE E FANNO USO DI ESPRESSIONI COME "...È..." O "...SONO UGUALI"?

LE IPOTESI ALTERNATIVE, INVECE, SONO FORMULATE COME NEGAZIONI E UTILIZZANO ESPRESSIONI COME "...NON È..." O "...NON SONO UGUALI".

È VERO.

DOVRESTI VEDERE LA QUESTIONE IN QUESTO MODO.

SCEGLIERE COME IPOTESI NULLA UN'IPOTESI CHE È POSITIVA ED APPARE DIFFICILE DA DIMOSTRARE.

MI SEMBRA GIUSTO.

POI, COME IPOTESI ALTERNATIVA, L'OPPOSTA DI QUELLA NULLA.

4. P-VALUE E PROCEDURE PER I TEST D'IPOTESI



VUOI DIRE LA PARTE TRATTEGGIATA?

PRIMA DELL'AVVENTO DEI PERSONAL COMPUTER, CALCOLARE IL P-VALUE ERA TERRIBILMENTE DIFFICILE.

DICIAMO, FINO ALL'INIZIO DEGLI ANNI NOVANTA.

DAVVERO...?

ERA QUESTO IL MOTIVO PER CUI IN GENERE SI USAVA IL METODO (1)...

...PER LA MAGGIOR PARTE DELLE CONCLUSIONI NEI TEST D'IPOTESI.

E OGGI?

È FACILE CALCOLARLO CON UN FOGLIO ELETTRONICO O ALTRI PROGRAMMI. PER QUESTO OGGI (2) STA COMINCIANDO A ESSERE USATO SEMPRE PIÙ SPESO.

MI FA DAVVERO PIACERE.

(2) IMPLICA UN PROCEDIMENTO DIVERSO DA QUELLO CHE HO SPIEGATO PER (1), QUINDI...

"LA MARIONETTA STATISTICA COLPISCE ANCORA CON LE SUE ANALISI!"

SMETTILA SUBITO! QUELLA VOCINA DA RAGAZZA È INSOPPORTABILE!

Passo 6p

Cerchiamo di determinare se il p-value relativo alla statistica del passo 5 è inferiore al livello di significatività.

Il livello di significatività è 0,05. Poiché il valore del test chi-quadrato di Pearson χ_0^2 è 8,0091, il p-value è 0,0182 e quindi minore di 0,05.



Come abbiamo detto, possiamo calcolare il p-value con un foglio elettronico (anche se questo dipende dal tipo di test d'ipotesi che stiamo svolgendo). Per i dettagli, potete andare a pagina 208.

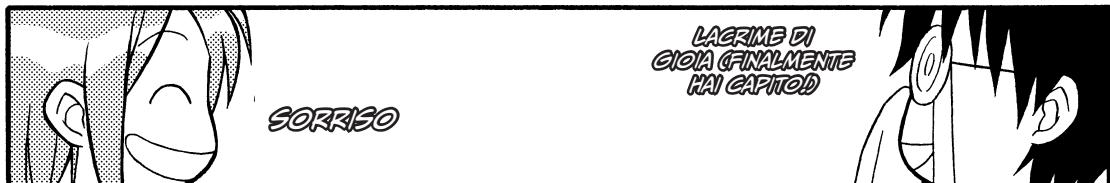
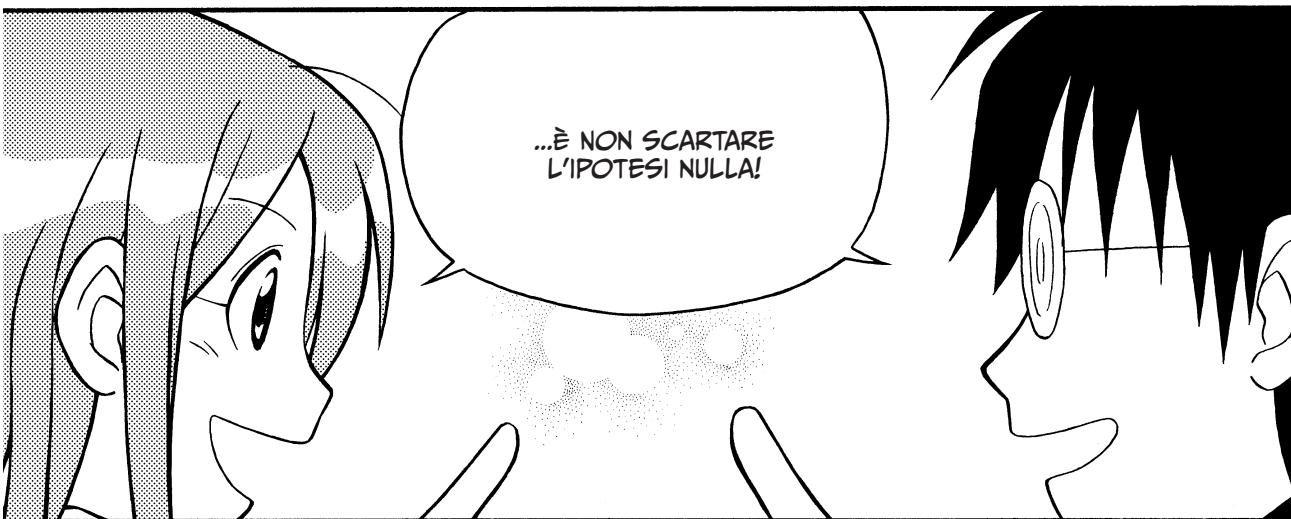
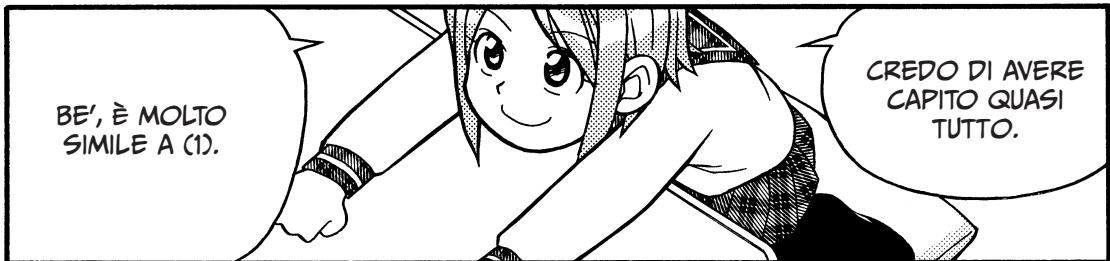
Passo 7p

Se il p-value è inferiore al livello di significatività al passo 6p, rifiutiamo l'ipotesi nulla. Altrimenti non potremo scartarla.

Il p-value era inferiore al livello di significatività. Pertanto optiamo per l'ipotesi alternativa: "il Coefficiente di Cramer della popolazione è maggiore di 0".



In realtà, anche nel caso che il p-value sia minore del livello di significatività, non possiamo concludere che in un test d'ipotesi l'ipotesi alternativa sia "assolutamente" corretta. L'unica assunzione che possiamo fare è che "c'è una probabilità del ($\alpha \times 100\%$) che l'ipotesi nulla sia corretta".



NEL CORSO
DELLE NOSTRE
LEZIONI HAI LA-
VORATO DAVVE-
RO SODO, RUI.

GRAZIE MILLE, MR
YAMAMOTO.

ALL'INIZIO,
STATISTICA MI
SEMPRAVA DAV-
VERO DIFFICILE,
MA ALLA FINE
HO IMPARATO UN
SACCO DI COSE.

COMINCIO DAVVERO A
PENSARE CHE LA TABU-
LAZIONE DEI SONDAG-
GI E COSE DEL GE-
NERE SIANO PROPRIO
DIVERTENTI.

È BELLO INSEGNARE
A STUDENTI COME TE!

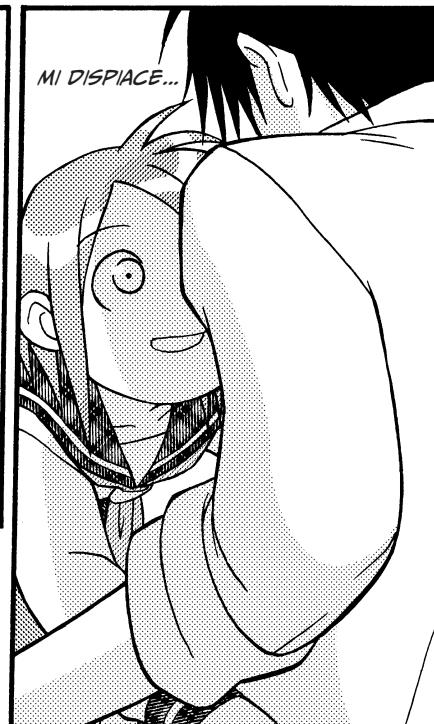
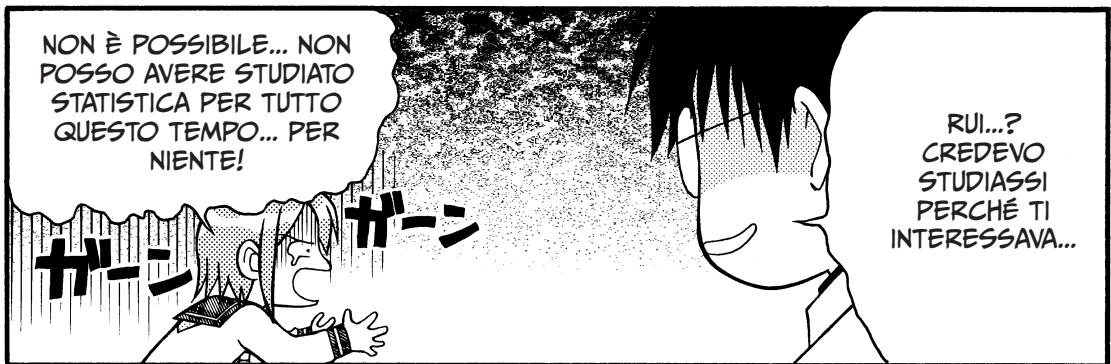
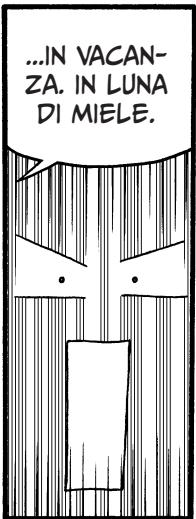
SONO FELICE DI
SENTIRTELLO DIRE.

VOGLIO FARE
ANCHE ALTRI
TIPI DI ANALISI.

SE TI SENTI
COSÌ ENTUSIA-
STA, PERCHÉ
NON COMINCI A
LAVORARE NEL
NOSTRO
SETTORE?

OH! A MOMENTI DI-
MENTICAVO CHE HO
COMINCIATO A STU-
DIARE STATISTICA PER
LAVORARE CON MR
IGARASHI.

A QUESTO
PUNTO, POS-
SO ANCHE
ANDARE





PER FAVORE,
MR YAMAMOTO,
CONTINUI A INSEGNARMI
UN SACCO DI COSE!

E COSÌ LE LEZIONI
PROSEGUITRANNO...

...FORSE... O FORSE NO.

5. TEST D'INDIPENDENZA E TEST DI OMogeneità



Esiste un test molto simile a quello dell'ipotesi nulla, detto *test di omogeneità* e sotto ne vedete riportato un esempio. Studiatelo attentamente e cercate di capire in che cosa sia diverso da un test d'indipendenza.

ESEMPIO

La rivista *P-Girl Magazine* ha pubblicato un articolo intitolato "Abbiamo chiesto a 300 liceali 'Come vorreste che vi chiedessero di uscire?'. Le possibili risposte erano "per telefono", "per email" o "di persona".

IPOTESI: I RAPPORTI TRA LE VARIE RISPOSTE DEI RAGAZZI SONO DIVERSI DA QUELLI DELLE RAGAZZE.

Per verificare l'ipotesi, un giornalista ha selezionato a caso dai due gruppi "tutti i liceali giapponesi maschi" e "tutte le liceali giapponesi femmine". La tabella riassume i risultati:

Intervistato	Preferenza	Età	Sesso
1	Di persona	17	Femmina
...
148	E-mail	16	Femmina
149	Telefono	15	Maschio
...
300	E-mail	18	Maschio

La tabella di contingenza tra sesso e modalità d'invito preferita ha prodotto la tabella seguente:

		Modalità d'invito preferita			Somma
		Telefono	E-mail	Di persona	
Sesso	Femmine	34	61	53	148
	Maschi	38	40	74	152
Somma		72	101	127	300

Stimate se l'ipotesi formulata sia corretta ricorrendo a un test di omogeneità. Come livello di significatività adottate 0,05.

PROCEDIMENTO

Passo 1	Definire la popolazione	In questo caso, la popolazione è “tutti i liceali giapponesi maschi” e “tutte le liceali giapponesi femmine”.
Passo 2	Formulare un’ipotesi nulla e un’ipotesi alternativa	L’ipotesi nulla è “i rapporti tra le risposte ‘per telefono’, ‘per email’ e ‘di persona’ sono gli stessi per i ragazzi e le ragazze”. L’ipotesi alternativa è “i rapporti tra le risposte ‘per telefono’, ‘per email’ e ‘di persona’ sono diversi per i ragazzi e le ragazze”.
Passo 3	Scegliere il test di ipotesi da effettuare	In questo caso, un test d’omogeneità.
Passo 4	Fissare un livello di significatività	In questo caso, 0,05.
Passo 5	Calcolare la statistica del test a partire dai dati del campione	Stiamo svolgendo un test d’omogeneità, pertanto la statistica del test è quella del chi-quadro di Pearson. Abbiamo già calcolato a pagina 132 il valore di χ^2_0 per questo esercizio $\chi^2_0 = 8.0091$
		Questa statistica segue una distribuzione chi-quadro con gradi di libertà $(2-1)(3-1)=2$, se l’ipotesi nulla è vera.
Passo 6	Stabilire se la statistica al passo 5 ricade nella regione critica	Secondo la tabella della distribuzione chi-quadro a pagina 103, con livello di significatività 0,05 la regione di critica va da 5,9915 in su. Quindi per il Passo 5 il valore della statistica ricade nella regione critica.
Passo 7	Se la statistica ricade nella regione critica al passo 6, rifiutare l’ipotesi nulla e adottare l’ipotesi alternativa. In caso contrario, non possiamo scartarla.	La statistica ricade nella regione nulla e pertanto adottiamo l’ipotesi alternativa: “i rapporti tra le risposte ‘per telefono’, ‘per email’ e ‘di persona’ sono diversi per i ragazzi e le ragazze”.



Non sembra anche a voi che esempio e procedura siano sostanzialmente analoghi a quelli per il test d'indipendenza? Diamo un'occhiata alle differenze e, per cominciare, osserviamo tre cose.

In primo luogo, le popolazioni in esame sono diverse. Nel primo test, è una sola (“tutti i liceali giapponesi”), nel secondo sono due (“tutti i liceali giapponesi maschi” e “tutte le liceali giapponesi femmine”).

Sono diverse anche le ipotesi. Nel primo test sono:

Ipotesi nulla	Il Coefficiente di Cramer della popolazione è 0, cioè genere e modalità d'invito preferita sono scorrelati.
Ipotesi alternativa	Il Coefficiente di Cramer della popolazione è maggiore di 0. In altre parole, genere e modalità d'invito preferita sono correlati.

Nel secondo:

Ipotesi nulla	I rapporti tra le risposte ‘per telefono’, ‘per email’ e ‘di persona’ sono gli stessi per i ragazzi e le ragazze.
Ipotesi alternativa	I rapporti tra le risposte ‘per telefono’, ‘per email’ e ‘di persona’ sono diversi per i ragazzi e le ragazze.

Infine, l'ordine dei passi della procedura è diverso: nel primo test l'ipotesi viene fissata dopo la raccolta dei dati, mentre nel secondo viene decisa prima.

Ci sono quindi delle differenze, ma spesso capita che i due tipi di test vengano confusi e che si effettui l'uno per l'altro... quindi attenzione!

6. CONCLUSIONI

Finora abbiamo formulato la conclusione di un test d'ipotesi in questo modo:

SE LA STATISTICA DEL TEST RICADE NELLA REGIONE CRITICA, POSSIAMO SCARTARE L'IPOTESI NULLA. VICEVERSA, NON POSSIAMO SCARTARLA.

Possiamo esprimere le stesse conclusioni in altri modi, riassunti dalla tabella seguente.

TABELLA 7.4 - FORMULAZIONE DELLE CONCLUSIONI DI UN TEST D'IPOTESI

Quando la statistica ricade nella regione critica	Quando la statistica non ricade nella regione critica
<ul style="list-style-type: none">• Si adotta l'ipotesi alternativa• Il risultato è statisticamente significativo• Si scarta l'ipotesi nulla	<ul style="list-style-type: none">• Non possiamo scartare l'ipotesi nulla• Il risultato non è statisticamente significativo• Accettiamo l'ipotesi nulla

Le espressioni “è statisticamente significativo” e “non è statisticamente significativo” sono decisamente popolari nei testi d'avviamento alla statistica. Perché quindi abbiamo utilizzato deliberatamente un'espressione meno nota? Probabilmente molte persone ancora inesperte di test d'ipotesi usano l'espressione “è statisticamente significativo” senza comprenderne effettivamente il significato, come se semplicemente confermassero la statistica del test, o p-value. Senza fissare un'ipotesi nulla e una alternativa adeguate, il termine *significativo* diventa ambiguo. Un altro problema degli inesperti è che anche la loro definizione della popolazione lascia spesso a desiderare.

Una volta pensavo di non dovere essere così preciso e severo con chi cominciava a studiare statistica, ma senza ipotesi formulate con precisione non è possibile trarre conclusioni valide. In questo libro ho pertanto utilizzato le espressioni “scartare l'ipotesi nulla” e “non scartare l'ipotesi nulla” in modo da suggerire la necessità di definire il più accuratamente possibile quali sono le ipotesi.

ESERCIZIO CON SOLUZIONE

ESERCIZIO

La tabella seguente è la stessa di pagina 138.

		Preferenza per caffè o tè		Somma
		Caffè	Tè	
Tipo di menù ordinato più spesso	Giapponese	43	33	76
	Europeo	51	53	104
	Cinese	29	41	70
Somma		123	127	250

Usare un test d'indipendenza chi-quadro per valutare se il Coefficiente di Cramer per il menu preferito e la preferenza tra tè e caffè nella popolazione “persone di nazionalità giapponese di almeno 20 anni di età” sia maggiore di 0. Questo equivale a stabilire se esiste una correlazione tra il menù preferito e la preferenza tra tè e caffè. Fissate il livello di significatività in 0,01.

SOLUZIONE

Passo 1 Definire la popolazione	In questo caso, la popolazione è “persone di nazionalità giapponese di almeno 20 anni di età”.
Passo 2 Formulare un'ipotesi nulla e un'ipotesi alternativa	L'ipotesi nulla è “il menù preferito e la preferenza tra tè e caffè non sono correlati”. L'ipotesi alternativa è “il menù preferito e la preferenza tra tè e caffè sono correlati”.
Passo 3 Scegliere il test d'ipotesi da effettuare	In questo caso, un test d'indipendenza chi-quadro.
Passo 4 Fissare un livello di significatività	In questo caso, 0,01.
Passo 5 Calcolare la statistica del test a partire dai dati del campione	Stiamo svolgendo un test d'indipendenza chi-quadro, pertanto la statistica del test è quella del chi-quadro di Pearson (χ_0^2). Abbiamo già calcolato a pagina 141 il valore di χ_0^2 per questo esercizio: $\chi_0^2 = 3,3483$.
Passo 6 Stabilire se la statistica al passo 5 ricade nella regione critica	Poiché $\chi_0^2 = 3,3483$, secondo la tabella della distribuzione chi-quadro a pagina 103, con livello di significatività $\alpha=0,01$, la regione critica va da 9,2104 in su. La statistica pertanto non ricade nella regione critica.
Passo 7 Se la statistica ricade nella regione critica al passo 6, rifiutare l'ipotesi nulla. In caso contrario, non possiamo scartare l'ipotesi nulla.	In questo caso, quindi, non possiamo scartare l'ipotesi nulla: “il menù preferito e la preferenza tra tè e caffè non sono correlati”.

RIASSUMENDO



- Un *test d'ipotesi* è una tecnica d'analisi statistica utilizzata per stimare la correttezza di un'ipotesi formulata dall'analista utilizzando un campione di dati.
- Il nome formale del test è *test di verifica d'ipotesi*.
- Le statistiche si ottengono tramite una funzione che fornisce un singolo valore a partire dai dati campione.
- In generale, come livello di significatività si utilizzano i valori 0,05 o 0,01.
- La *regione critica* è un'area che corrisponde al livello di significatività (detto anche valore alfa e indicata col simbolo α).
- Un *test d'indipendenza chi-quadro* è una tecnica d'analisi utilizzata per valutare se il Coefficiente di Cramer di una popolazione è 0. Si può anche dire che serve a stimare se le due variabili di una tabella di contingenza sono correlate.
- Se il Coefficiente di Cramer di una popolazione è 0, la statistica del chi-quadro di Pearson segue una distribuzione chi-quadro.
- In un test d'indipendenza il *p-value* è la probabilità di ottenere un valore della statistica chi-quadro di Pearson uguale o maggiore del valore corrispondente al caso in cui l'ipotesi nulla è vera.
- La conclusione di un test d'ipotesi, può essere espressa in due modi:
 1. Se la statistica si trova nella regione critica.
 2. Se il p-value è minore del livello di significatività.
- La procedura da seguire in qualsiasi test d'ipotesi è la stessa per un test d'indipendenza, o qualsiasi altro tipo di test:

Passo 1 Definire la popolazione.

Passo 2 Formulare un'ipotesi nulla e un'ipotesi alternativa.

Passo 3 Scegliere il test d'ipotesi da effettuare.

Passo 4 Fissare un livello di significatività.

Passo 5 Calcolare la statistica del test a partire dai dati del campione.

Passo 6 Stabilire se la statistica al passo 5 ricade nella regione critica.

Passo 7 Se la statistica ricade nella regine critica al passo 6, rifiutare l'ipotesi nulla. In caso contrario, non possiamo scartarla.

Passo 6p Determinare se il p-value relativo alla statistica del passo 5 è inferiore al livello di significatività.

Passo 7p Se il p-value è inferiore al livello di significatività al passo 6p, rifiutare l'ipotesi nulla. Altrimenti non possiamo scartarla.

APPENDICE:
FACCIAMO UN PO' DI CALCOLI
CON UN FOGLIO ELETTRONICO



Questa appendice fornisce alcuni suggerimenti per calcolare diverse grandezze statistiche utilizzando un foglio elettronico.

Qui imparerete a:

1. Compilare una tabella delle frequenze.
2. Calcolare media aritmetica, mediana e deviazione standard.
3. Compilare una tabella di contingenza.
4. Calcolare il valore normale e il valore di deviazione.
5. Calcolare la probabilità della distribuzione normale standard.
6. Calcolare il punto sull'asse orizzontale della distribuzione chi-quadro.
7. Calcolare il coefficiente di correlazione.
8. Effettuare dei test d'indipendenza.

I lettori meno esperti di fogli elettronici dovrebbero prima leggere.

"Calcoliamo Media Aritmetica, Mediana e Deviazione Standard" a pagina 195.

1. COMPILARE UNA TABELLA DELLE FREQUENZE

Questo esercizio riprende i prezzi dei ristoranti di ramen a pagina 33.

Passo 1

Selezzionate la cella J3.

A	B	C	D	E	F	G	H	I	J
1	Prezzo (yen)			Prezzo (yen)		Maggiore o uguale di	Minore di	Minore o uguale	Frequenza
2	Ristorante di ramen 1	700	Ristorante di ramen 26	780		500	600	599	
3	Ristorante di ramen 2	850	Ristorante di ramen 27	590		600	700	699	
4	Ristorante di ramen 3	600	Ristorante di ramen 28	650		700	800	799	
5	Ristorante di ramen 4	650	Ristorante di ramen 29	580		800	900	899	
6	Ristorante di ramen 5	980	Ristorante di ramen 30	750		900	1000	999	
7	Ristorante di ramen 6	750	Ristorante di ramen 31	800					
8	Ristorante di ramen 7	500	Ristorante di ramen 32	550					
9	Ristorante di ramen 8	890	Ristorante di ramen 33	750					
10	Ristorante di ramen 9	880	Ristorante di ramen 34	700					
11	Ristorante di ramen 10	700	Ristorante di ramen 35	600					
12	Ristorante di ramen 11	890	Ristorante di ramen 36	800					
13	Ristorante di ramen 12	720	Ristorante di ramen 37	800					
14	Ristorante di ramen 13	680	Ristorante di ramen 38	880					
15	Ristorante di ramen 14	650	Ristorante di ramen 39	790					
16	Ristorante di ramen 15	790	Ristorante di ramen 40	790					
17	Ristorante di ramen 16	670	Ristorante di ramen 41	780					
18	Ristorante di ramen 17	680	Ristorante di ramen 42	600					
19	Ristorante di ramen 18	900	Ristorante di ramen 43	670					
20	Ristorante di ramen 19	880	Ristorante di ramen 44	680					
21	Ristorante di ramen 20	720	Ristorante di ramen 45	650					
22	Ristorante di ramen 21	850	Ristorante di ramen 46	890					
23	Ristorante di ramen 22	700	Ristorante di ramen 47	930					
24	Ristorante di ramen 23	780	Ristorante di ramen 48	650					
25	Ristorante di ramen 24	850	Ristorante di ramen 49	777					
26	Ristorante di ramen 25	750	Ristorante di ramen 50	700					

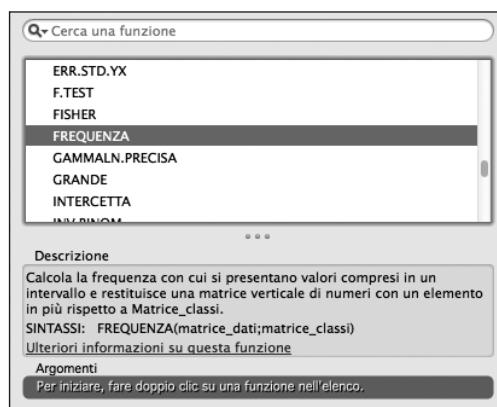
Passo 2

Selezionate **Inserisci ▶ Funzione.**



Passo 3

Selezionate **FREQUENZA** per il nome della funzione.



Passo 4

Selezionate l'area evidenziata in figura e cliccate **OK**.

A	B	C	D	E	F	G	H	I	J
1		Prezzo (yen)		Prezzo (yen)					
2	Ristorante di ramen 1	700	Ristorante di ramen 26	780		Maggiore o uguale di	Minore di	Minore o uguale	Frequenza
3	Ristorante di ramen 2	850	Ristorante di ramen 27	590		500	600	599	6;1;17)
4	Ristorante di ramen 3	600	Ristorante di ramen 28	650		600	700	699	
5	Ristorante di ramen 4	650	Ristorante di ramen 29	580		700	800	799	
6	Ristorante di ramen 5	980	Ristorante di ramen 30	750		800	900	899	
7	Ristorante di ramen 6	750	Ristorante di ramen 31	800		900	1000	999	
8	Ristorante di ramen 7	500	Ristorante di ramen 32	550					
9	Ristorante di ramen 8	890	Ristorante di ramen 33	750					
10	Ristorante di ramen 9	880							
11	Ristorante di ramen 10	700							
12	Ristorante di ramen 11	890							
13	Ristorante di ramen 12	720							
14	Ristorante di ramen 13	680							
15	Ristorante di ramen 14	650							
16	Ristorante di ramen 15	790							
17	Ristorante di ramen 16	670							
18	Ristorante di ramen 17	680							
19	Ristorante di ramen 18	900							
20	Ristorante di ramen 19	880							
21	Ristorante di ramen 20	720							
22	Ristorante di ramen 21	850							
23	Ristorante di ramen 22	700							
24	Ristorante di ramen 23	780							
25	Ristorante di ramen 24	850	Ristorante di ramen 49	777					
26	Ristorante di ramen 25	750	Ristorante di ramen 50	700					
27									

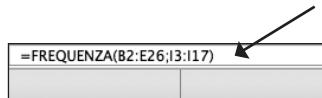
Passo 5

Partendo dalla cella J3, selezionate fino alla cella J7 compresa, come in figura.

G	H	I	J
Maggiore o uguale di	Minore di	Minore o uguale	Frequenza
500	600	599	4
600	700	699	
700	800	799	
800	900	899	
900	1000	999	

Passo 6

Cliccate nella barra della formula.



Passo 7

Premete **INVIO** tenendo premuti contemporaneamente il tasto delle maiuscole e **CTRL**.

Passo 8

Adesso avete la frequenza di ogni classe!

G	H	I	J
Maggiore o uguale di	Minore di	Minore o uguale	Frequenza
500	600	599	4
600	700	699	13
700	800	799	18
800	900	899	12
900	1000	999	3

2. CALCOLARE MEDIA ARITMETICA, MEDIANA E DEVIAZIONE STANDARD

I dati sono quelli del torneo di bowling delle compagne di Rui visti a pagina 41.



Passo 1

Selezzionate la cella B10.

A	B
1	Squadra A
2 Rui-Rui	86
3 Jun	73
4 Yumi	124
5 Shizuka	111
6 Touko	90
7 Kaede	38
8	
9	
10 Media	
11 Mediana	
12 Deviazione standard	
13	

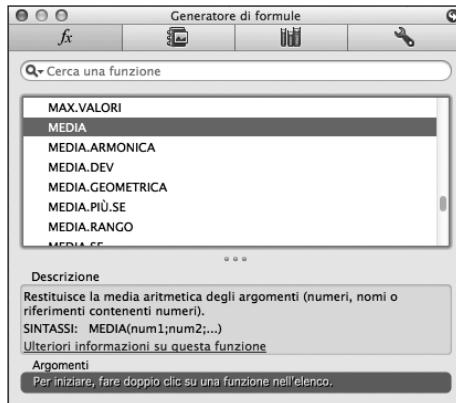
Passo 2

Selezzionate **Inserisci ▶ Funzione...**



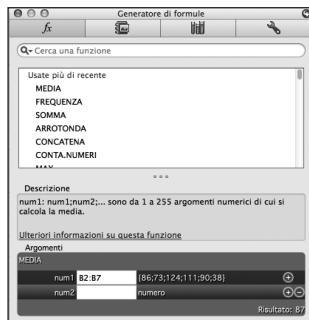
Passo 3

Selezzionate **STATISTICA** dal menù a tendina e poi **MEDIA** per il nome della funzione.



Passo 4

Inserite l'intervallo in figura, poi cliccate **OK**.



Passo 5

Ora avete ottenuto il punteggio medio della squadra.

	A	B
1		Squadra A
2	Rui-Rui	86
3	Jun	73
4	Yumi	124
5	Shizuka	111
6	Touko	90
7	Kaede	38
8		
9		
10	Media	87
11	Mediana	
12	Deviazione standard	

Calcolate media e deviazione standard ripetendo i passi da 1 a 5 e, al passo 2, selezionando le funzioni **MEDIANA** e **DEV.STAND.P**.

3. COMPILARE UNA TABELLA DI CONTINGENZA



I dati sono quelli delle risposte dei compagni di classe di Rui sulle nuove uniformi, che trovate a pagina 61.

Passo 1

Selezionate la cella F20, poi **Inserisci ▶ Funzione**.

A	B	C	D	E	F	G	H
1	Risposta			Risposta			Risposta
2	1	piace	16	non so	31	non so	
3	2	non so	17	piace	32	non so	
4	3	piace	18	piace	33	piace	
5	4	non so	19	piace	34	non piace	
6	5	non piace	20	piace	35	piace	
7	6	piace	21	piace	36	piace	
8	7	piace	22	piace	37	piace	
9	8	piace	23	non piace	38	piace	
10	9	piace	24	non so	39	non so	
11	10	piace	25	piace	40	piace	
12	11	piace	26	piace			
13	12	piace	27	non piace			
14	13	non so	28	piace			
15	14	piace	29	piace			
16	15	piace	30	piace			
17							
18							
19					Frequenza		
20				piace			
21				non so			
22				non piace			
23							

Passo 2

Selezionate **Statistica** al menù a tendina e poi **CONTA.SE** per il nome della funzione.

Passo 3

Selezzionate l'area evidenziata in figura, scrivete “piace” (senza le virgolette!) alla voce **CRITERI**, poi cliccate **OK**.

A	B	C	D	E	F	G	H
1	Risposta			Risposta			Risposta
2	1 piace		16	non so		31	non so
3	2 non so		17	piace		32	non so
4	3 piace					33	piace
5	4 non so					34	non piace
6	5 non piace					35	piace
7	6 piace					36	piace
8	7 piace					37	piace
9	8 piace					38	piace
10	9 piace					39	non so
11	10 piace					40	piace
12	11 piace						
13	12 piace						
14	13 non so						
15	14 piace						
16	15 piace						
17							
18							
19							Frequenza
20				piace		28	(;piace)
21				non so			
22				non piace			
23							

Passo 4

Adesso avete il totale dei compagni di classe di Rui a cui piacciono le nuove uniformi.

Passo 5

Potete ricavare la frequenza di “non piace” e “non so” ripetendo i passi da 1 a 4 e scrivendo i rispettivi termini al passo 3.

A	B	C	D	E	F	G	H
1	Risposta			Risposta			Risposta
2	1 piace		16	non so		31	non so
3	2 non so		17	piace		32	non so
4	3 piace		18	piace		33	piace
5	4 non so		19	piace		34	non piace
6	5 non piace		20	piace		35	piace
7	6 piace		21	piace		36	piace
8	7 piace		22	piace		37	piace
9	8 piace		23	non piace		38	piace
10	9 piace		24	non so		39	non so
11	10 piace		25	piace		40	piace
12	11 piace		26	piace			
13	12 piace		27	non piace			
14	13 non so		28	piace			
15	14 piace		29	piace			
16	15 piace		30	piace			
17							
18							
19							Frequenza
20				piace		28	(;piace)
21				non so			
22				non piace			
23							

4. CALCOLARE IL VALORE STANDARD E QUELLO DI DEVIAZIONE



I dati utilizzati per questo esercizio sono quelli visti a pagina 72.

I passi dall'1 all'8 mostrano come ricavare il valore standard.

Con i passi dal 9 all'11 ricaverete quello di deviazione. Esiste una funzione apposita per il primo, ma non per il secondo, che però può essere calcolato abbastanza facilmente a partire dai valori standard.

Passo 1

Selezzionate la cella E2.

	A	B	C	D	E	F
1		Storia			Valore standard	Valore di deviazione
2	Rui	73		Rui		
3	Yumi	61		Yumi		
4	A	14		A		
5	B	41		B		
6	C	49		C		
7	D	87		D		
8	E	69		E		
9	F	65		F		
10	G	36		G		
11	H	7		H		
12	I	53		I		
13	J	100		J		
14	K	57		K		
15	L	45		L		
16	M	56		M		
17	N	34		N		
18	O	37		O		
19	P	70		P		
20	Media	53				
21	Deviazione standard	22,7				
22						

Passo 2

Selezzionate **Inserisci > Funzione**. Poi **Statistica**, e infine **NORMALIZZA** per il nome della funzione.

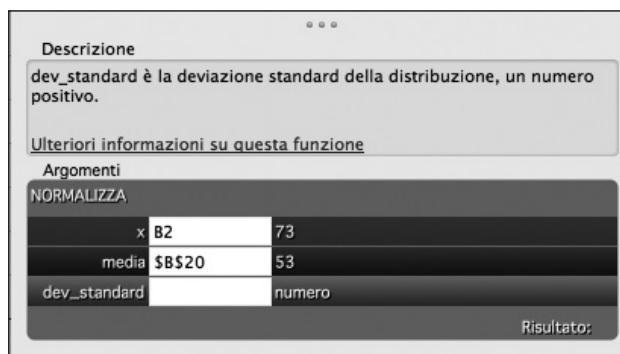
Passo 3

Selezzionate la cella B2.

	A	B	C	D	E	F
1		Storia			Valore standard	Valore di deviazione
2	Rui	73	Rui	A(B2)		
3	Yumi	61	Yumi			
4	A	14	A			
5	B	41	B			
6	C	49	C			
7	D	87	D			
8	E	69	E			
9	F	65	F			
10	G	36	G			
11	H	7	H			
12	I	53	I			
13	J	100	J			
14	K	57	K			
15	L	45	L			
16	M	56	M			
17	N	34	N			
18	O	37	O			
19	P	70	P			
20	Media	53				
21	Deviazione standard	22,7				
22						

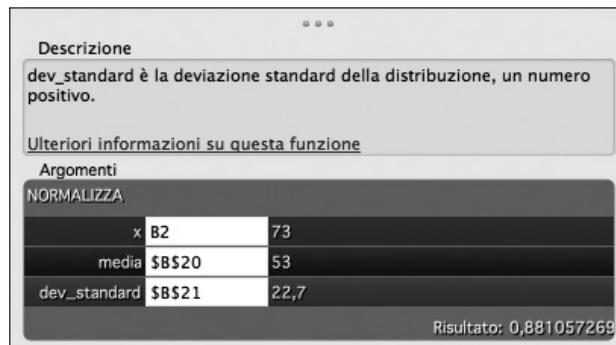
Step 4

Selezzionate B20 alla voce **MEDIA** e modificate B20 in \$B\$20.



Passo 5

Selezzionate B21 alla voce **DEV_STANDARD** e modificate B21 in \$B\$21, poi cliccate **OK**.



Passo 6

Avete calcolato il valore standard di Rui.

	A	B	C	D	E	F
1		Storia			Valore standard	Valore di deviazione
2	Rui	73	Rui		0,88	
3	Yumi	61	Yumi			
4	A	14	A			
5	B	41	B			
6	C	49	C			
7	D	87	D			
8	E	69	E			
9	F	65	F			
10	G	36	G			
11	H	7	H			
12	I	53	I			
13	J	100	J			
14	K	57	K			
15	L	45	L			
16	M	56	M			
17	N	34	N			
18	O	37	O			
19	P	70	P			
20	Media	53				
21	Deviazione standard	22,7				
22						

Passo 7

Avvicinate la punta del cursore al vertice in basso a destra della cella E2, verificate che diventi una crocetta nera, trascinatela in basso fino alla cella E19 tenendo premuto il tasto sinistro del mouse, e alla fine lasciate lo andare.

A	B	C	D	E	F
1	Storia			Valore standard	Valore di deviazione
2 Rui	73	Rui		0,88	
3 Yumi	61	Yumi			
4 A	14	A			
5 B	41	B			
6 C	49	C			
7 D	87	D			
8 E	69	E			
9 F	65	F			
10 G	36	G			
11 H	7	H			
12 I	53	I			
13 J	100	J			
14 K	57	K			
15 L	45	L			
16 M	56	M			
17 N	34	N			
18 O	37	O			
19 P	70	P			
20 Media	53				
21 Deviazione standard	22,7				
22					

Passo 8

Ora avete i valori standard di tutti!

A	B	C	D	E	F
1	Storia			Valore standard	Valore di deviazione
2 Rui	73	Rui		0,88	
3 Yumi	61	Yumi		0,35	
4 A	14	A		-1,72	
5 B	41	B		-0,53	
6 C	49	C		-0,18	
7 D	87	D		1,50	
8 E	69	E		0,70	
9 F	65	F		0,53	
10 G	36	G		-0,75	
11 H	7	H		-2,03	
12 I	53	I		0,00	
13 J	100	J		2,07	
14 K	57	K		0,18	
15 L	45	L		-0,35	
16 M	56	M		0,13	
17 N	34	N		-0,84	
18 O	37	O		-0,70	
19 P	70	P		0,75	
20 Media	53				
21 Deviazione standard	22,7				
22					

Passo 9

Selezzionate la cella F2 e scrivete $=E2*10+50$, poi premete **INVIO**.

	A	B	C	D	E	F
1		Storia			Valore standard	Valore di deviazione
2	Rui	73	Rui		0,88	$=E2*10+50$
3	Yumi	61	Yumi		0,35	
4	A	14	A		-1,72	
5	B	41	B		-0,53	
6	C	49	C		-0,18	
7	D	87	D		1,50	
8	E	69	E		0,70	
9	F	65	F		0,53	
10	G	36	G		-0,75	
11	H	7	H		-2,03	
12	I	53	I		0,00	
13	J	100	J		2,07	
14	K	57	K		0,18	
15	L	45	L		-0,35	
16	M	56	M		0,13	
17	N	34	N		-0,84	
18	O	37	O		-0,70	
19	P	70	P		0,75	
20	Media	53				
21	Deviazione standard	22,7				
22						

Passo 10

Trascinate il cursore fino alla cella F19, come al passo 7.

Passo 11

Ora avete i valori di deviazione di tutta la classe.

	A	B	C	D	E	F
1		Storia			Valore standard	Valore di deviazione
2	Rui	73	Rui		0,88	58,81
3	Yumi	61	Yumi		0,35	53,52
4	A	14	A		-1,72	32,82
5	B	41	B		-0,53	44,71
6	C	49	C		-0,18	48,24
7	D	87	D		1,50	64,98
8	E	69	E		0,70	57,05
9	F	65	F		0,53	55,29
10	G	36	G		-0,75	42,51
11	H	7	H		-2,03	29,74
12	I	53	I		0,00	50,00
13	J	100	J		2,07	70,70
14	K	57	K		0,18	51,76
15	L	45	L		-0,35	46,48
16	M	56	M		0,13	51,32
17	N	34	N		-0,84	41,63
18	O	37	O		-0,70	42,95
19	P	70	P		0,75	57,49
20	Media	53				
21	Deviazione standard	22,7				
22						

5. CALCOLARE LA PROBABILITÀ DELLA DISTRIBUZIONE NORMALE STANDARD



Per questo esempio useremo i dati visti a pagina 93.

Passo 1

Selezzionate la cella B2.

	A	B
1	z	1,96
2	metà	
3	Area(=Percentuale=Rapporto)	
4		

Passo 2

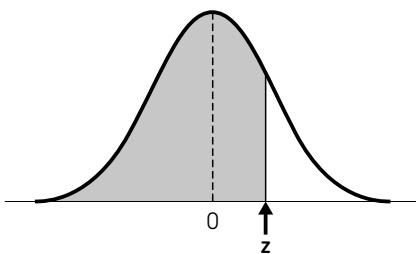
Selezzionate **Inserisci ▶ Funzione**, poi **Statistica**, e infine **DISTRIB.NORM.ST** per il nome della funzione.

Passo 3

Selezzionate la cella B1, e cliccate **OK**.

The screenshot shows a Microsoft Excel spreadsheet. The formula bar at the top has the formula =N.ST(B1) entered. The cell B1 contains the value 1,96. The cell B2 contains the text "metà". The cell B3 contains the formula "Area(=Percentuale=Rapporto)". A context menu is open over cell B3, with the option "N.ST(B1)" highlighted. The function dialog box for "DISTRIB.NORM.ST" is displayed, with the argument "B1" selected. The result of the function, 0.975002105, is shown in the bottom right corner of the dialog box.

In realtà, **Distrib.NORM.ST** è una funzione per il calcolo della probabilità, come si vede in questa figura



Passo 4

Scrivete $=B2-0,5$ nella cella B3.

	A	B
1	z	1,96
2	metà	0,975
3	Area(=Percentuale=Rapporto)	=B2-0,5
4		

Passo 5

Ora avete il valore dell'area.

	A	B
1	z	1,96
2	metà	0,975
3	Area(=Percentuale=Rapporto)	0,475
4		

6. CALCOLARE IL VALORE DELL'ASCISSA DELLA DISTRIBUZIONE CHI-QUADRO

I dati sono quelli a pagina 104.



Passo 1

Selezionate la cella B3.

	A	B
1	P	0,05
2	Gradi di libertà	1
3	Chi-quadro	
4		

Passo 2

Selezionate **Inserisci > Funzione**, poi la funzione **INV.CHI**.

Passo 3

Selezzionate le celle B1 e B2, poi cliccate **OK**.

Descrizione
grado_libertà è il numero di gradi di libertà, un numero compreso tra 1 e 100, escluso 100.

Ulteriori informazioni su questa funzione

Argomenti

INV.CHI

probabilità	B1	0,05
grado_libertà	B2	1

Risultato: 3,841458821

Passo 4

Ed ecco fatto.

	A	B
1	P	0,05
2	Gradi di libertà	1
3	Chi-quadro	3,84146
4		

7. CALCOLARE IL COEFFICIENTE DI CORRELAZIONE

Questi dati provengono dal sondaggio di *P-Girl Magazine* a pagina 116.

Passo 1

Selezzionate la cella B14.



	A	B	C
1		Spese per cosmetici (yen)	Spese per abbigliamento (yen)
2	Ms A	3000	7000
3	Ms B	5000	8000
4	Ms C	12000	25000
5	Ms D	2000	5000
6	Ms E	7000	12000
7	Ms F	15000	30000
8	Ms G	5000	10000
9	Ms H	6000	15000
10	Ms I	8000	20000
11	Ms J	10000	18000
12			
13			
14	Coefficiente di correlazione		
15			

Passo 2

Selezzionate **Inserisci > Funzione**, poi **Statistica**, e infine **CORRELAZIONE**.

Passo 3

Selezzionate l'area evidenziata in figura, poi cliccate **OK**.

A	B	C	D	E
1		Spese per cosmetici (yen)	Spese per abbigliamento (yen)	
2	Ms A	3000	7000	
3	Ms B	5000	8000	
4	Ms C	12000		
5	Ms D	2000		
6	Ms E	7000		
7	Ms F	15000		
8	Ms G	5000		
9	Ms H	6000		
10	Ms I	8000		
11	Ms J	10000		
12				
13				
14	Coefficiente di correlazione	:B2:B11;C2:C11)		
15				
16				
17				
18				

Passo 4

Ed ecco il coefficiente di correlazione.

A	B	C
1		Spese per cosmetici (yen)
2	Ms A	3000
3	Ms B	5000
4	Ms C	12000
5	Ms D	2000
6	Ms E	7000
7	Ms F	15000
8	Ms G	5000
9	Ms H	6000
10	Ms I	8000
11	Ms J	10000
12		
13		
14	Coefficiente di correlazione	0,968019613
15		

NOTA Nella maggior parte dei fogli elettronici non sono disponibili funzioni per il calcolo del rapporto di correlazione e del Coefficiente di Cramer.

8. TEST D'INDIPENDENZA



Questi dati provengono dal sondaggio visto a pagina 157.

Passo 1

Selezzionate la cella B8.

	A	B	C	D	E
1		Telefono	Email	Di persona	Somma
2	Femmine	34	61	53	148
3	Maschi	38	40	74	152
4	Somma	72	101	127	300
5					
6					
7		Telefono	Email	Di persona	
8	Femmine				
9	Maschi				
10					
11					
12	P-value				
13					
14					

Passo 2

Scrivete $=E2*B4/E4$ nella cella B8. Non premete INVIO.

	A	B	C	D	E
1		Telefono	Email	Di persona	Somma
2	Femmine	34	61	53	148
3	Maschi	38	40	74	152
4	Somma	72	101	127	300
5					
6					
7		Telefono	Email	Di persona	
8	Femmine	=E2*B4/E4			
9	Maschi				
10					
11					
12	P-value				
13					
14					

Passo 3

Selezzionate E2 nell'equazione che avete appena scritto e modificate E2 in \$E2. Non premete INVIO.

	A	B	C	D	E
1		Telefono	Email	Di persona	Somma
2	Femmine	34	61	53	148
3	Maschi	38	40	74	152
4	Somma	72	101	127	300
5					
6					
7		Telefono	Email	Di persona	
8	Femmine	=\$E2*B4/E4			
9	Maschi				
10					
11					
12	P-value				
13					
14					

Passo 4

Sempre nell'equazione nella cella B8, selezionate B4 e modificate lo in *B\$4*. Poi selezionate E4 e modificate lo in *\$E\$4*. Poi premete **INVIO**.

	A	B	C	D	E
1		Telefono	Email	Di persona	Somma
2	Femmine	34	61	53	148
3	Maschi	38	40	74	152
4	Somma	72	101	127	300
5					
6					
7		Telefono	Email	Di persona	
8	Femmine	=\\$E2*B\$4/\$E\$4			
9	Maschi				
10					
11					
12	P-value				
13					
14					

Passo 5

Selezzionate la cella B8, avvicinate la punta del cursore all'angolo in basso a destra della cella B8, verificate che il cursore sia diventato una crocetta nera, trascinatela in basso fino alla cella D8 tenendo premuto il tasto sinistro del mouse, e alla fine lasciatelo andare.

	A	B	C	D	E
1		Telefono	Email	Di persona	Somma
2	Femmine	34	61	53	148
3	Maschi	38	40	74	152
4	Somma	72	101	127	300
5					
6					
7		Telefono	Email	Di persona	
8	Femmine	35,52			
9	Maschi				
10					
11					
12	P-value				
13					

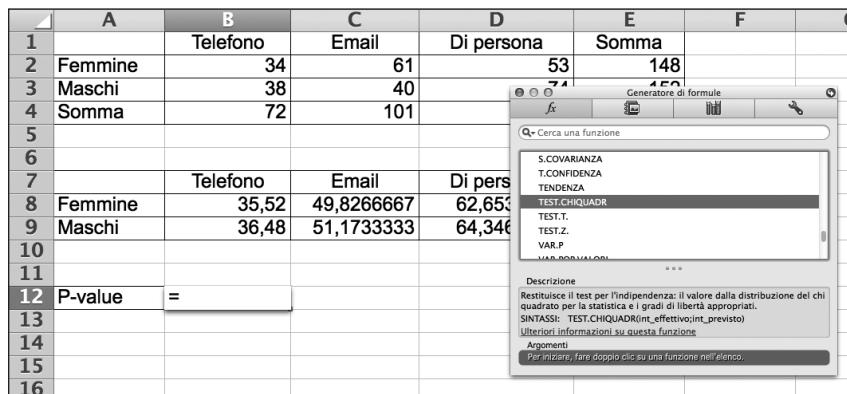
Passo 6

Selezionate l'area dalla cella B8 alla cella D8, avvicinate la punta del cursore all'angolo in basso a destra della cella D8, verificate che il cursore sia diventato una crocetta nera, trascinatela in basso fino alla cella D9 tenendo premuto il tasto sinistro del mouse, e alla fine lasciatelo andare.

A	B	C	D	E
1	Telefono	Email	Di persona	Somma
2 Femmine	34	61	53	148
3 Maschi	38	40	74	152
4 Somma	72	101	127	300
5				
6				
7	Telefono	Email	Di persona	
8 Femmine	35,52	49,8266667	62,65333333	
9 Maschi				
10				
11				
12 P-value				
13				

Passo 7

Selezionate la cella B12, poi **Inserisci > Funzione**, selezionate **Statistica**, e poi **TEST.CHIQUADR** per il nome della funzione.



The screenshot shows a Microsoft Excel spreadsheet with data in rows 1 through 12. Rows 1 through 4 contain the total counts for Femmine and Maschi across Telefonico, Email, and Di persona categories. Rows 5 through 11 are blank. Row 12 contains the formula =TEST.CHIQUADR(A8:D8,A12). A callout box from the formula bar displays the function's arguments: TEST.CHIQUADR(A8:D8,A12). The formula bar also shows the result of the function: 0.000123. The status bar at the bottom indicates "Per iniziare, fare doppio clic su una funzione nell'elenco."

Passo 8

Selezionate l'area evidenziata in figura e cliccate **OK**.

A	B	C	D	E	F	G
1		Telefono	Email	Di persona	Somma	
2	Femmine	34	61	53	148	
3	Maschi	38	40			
4	Somma	72	101			
5						
6						
7		Telefono	Email	Di pers		
8	Femmine	35,52	49,82666667	62,653		
9	Maschi	36,48	51,17333333	64,346		
10						
11						
12	P-value	D3:B8:D9)				
13						
14						
15						
16						
17						
18						
19						

Passo 9

Ecco fatto: ora potete verificare che il valore calcolato è uguale al p-value a pagina 177.

A	B	C	D	E
1	Telefono	Email	Di persona	Somma
2	Femmine	34	61	53
3	Maschi	38	40	74
4	Somma	72	101	127
5				300
6				
7	Telefono	Email	Di persona	
8	Femmine	35,52	49,82666667	62,65333333
9	Maschi	36,48	51,17333333	64,34666667
10				
11				
12	P-value	0,018233		
13				
14				

INDICE

A

- alfa, valore (a), 159, 163
- aritmetica, media, (v. media aritmetica)
- armonica, media (v. media armonica)
- assi orizzontali, 39, 102, 107, 109, 125
 - calcolare punti su, 107
- assi verticali, 39

C

- calcolo, fogli di. (v. fogli di calcolo)
- campione, 6, 7, 52, 57
- categorici, dati (v. dati categorici)
- chi-quadro, (v. distribuzione chi-quadro)
- chi-quadro, simbolo, 103
- chi-quadro, test
- d'indipendenza, 151–169
- classe, valore medio, 36–39, 54, 56
- classi,
 - ampiezza delle, 39, 54–57, 84
 - calcolo con la formula di Sturges, 55, 56, 58
 - varianza interclasse/ intraclasse, 117, 123, 124, 126
- Coefficiente di Cramer, 127–138
 - accuratezza del 147
 - calcolo, 130–135, 141
 - esempi di 127–136
 - fogli di calcolo e, 207
 - indici, 117, 129
 - ipotesi alternative, 186
 - ipotesi nulla, 168, 186
 - rapporti tra preferenze, 155
 - scelta informata sul, 147–148
 - standard informali del, 136
 - variazioni di popolazione, 145–150, 157, 186
- CONTA.SE funzione, 197–198
- contingenza, tabella di, 62–64, 128, 130, 135, 151, 153, 197–198
- correlazione, 115, 119
 - coefficiente di, 116–120,

D

- dati,
 - dispersione dei 49, 58, 69, 70, 80
 - non adatti al coefficiente di correlazione, 120
 - non misurabili (v. dati categorici)
 - numerici. (v. dati numerici)
 - raccolta dei 186
 - tipi di, 13–29, 117
 - dati categorici, 14–29
 - come risultato di un sondaggio, 60–64
 - creazione di una tabella, 60–64
 - definizione, 19
 - esempi di 20, 23–26
 - grafici a cilindro, 114
 - grafici a dispersione, 114
 - indici, 117
 - panoramica, 14–19
 - rapporto di correlazione, 121
 - dati numerici, 14–29
 - definizione, 19
 - deviazione standard, 48–53, 70–79
 - esempi di 21–23, 26
 - grafico a dispersione, 114
 - indici, 117
 - istogrammi, 38–39, 54, 58
 - media, 40–43
 - mediana, 44–47
 - panoramica, 31–58
 - rapporto di correlazione, 121
 - statistiche descrittive, 57–58
 - tabelle di distribuzione di frequenza, 32–39, 54–56, 58
 - teoria della stima, 57–58
 - deviazione standard, 48–53, 70–79
 - calcolo, 195–196
 - dati numerici, 48–53, 70–79

E

- distribuzione normale e, 87–91
- distribuzione normale standard e, 89–90
- popolazione, 52
- deviazione, valore di, 74–80, 199–203
- dispersione dei dati, 49, 58, 69, 70, 80
- DISTRIB.CHI, funzione, 107
- DISTRIB.F, funzione, 107
- DISTRIB.NORM, funzione, 107
- DISTRIB.NORM,ST, funzione , 107, 204
- DISTRIB.T, funzione, 107
- distribuzione,
 - F, 106–107
 - fogli di calcolo e, 107–109
 - normale, 86–91
 - standard normale, 89–98, 204–205
 - t, 106
- distribuzione chi-quadro, 99–105
 - calcolo 130–133
 - descrizione, 99
 - esempi di 99–105, 152
 - gradi di libertà, 99–108
 - punti sull'asse orizzontale 205–206
- distribuzione normale standard, 89–98, 204–205

F

- F distribuzione, (v. distribuzione F,)
- fogli di calcolo, 191–211
 - coefficiente di correlazione 206–207
 - deviazione standard, 195–196
 - distribuzione chi-quadro, 205–206
 - distribuzione normale standard, 204–205
 - distribuzioni e, 107–109
 - media, 195–196
 - mediana, 195–196
 - tabella delle frequenze, 192–195

- tabelle di contingenza, 197–198
- test d'indipendenza, 208–211
- valore standard, 199–202
- valori di deviazioni, 74–80, 199–203
- fogli di calcolo, funzioni,
 - CONTA.SE, 197–198
 - CORRELAZIONE, 207
 - DISTRIB.CHI, 107
 - DISTRIB.F, 107
 - DISTRIB.NORM, 107
 - DISTRIB.NORM.ST, 107, 204
 - DISTRIB.T, 107
 - FREQUENZA, 193–194
 - INV.CHI, 107, 205–206
 - INV.F, 107
 - INV.NORM, 107
 - INV.NORM.ST 107
 - MEDIA, 196
 - NORMALIZZA, 199–201
 - TEST.CHIQUADR, 210–211
- FREQUENZA, funzione, 193–194
 - frequenze,
 - attese, 130, 131
 - descrizione, 36
 - osservate, 130, 131
 - relative, 36–37, 39
 - tabelle di distribuzione di, 32–39
- frequenze, tabelle di,
 - ampiezze di classe, 54–56
 - con fogli di calcolo, 192–195
- funzione di densità di probabilità, 82–85, 99, 107, 109

- G**
- geometrica, media (v. media geometrica)
- gradi
 - di libertà, 99–108
 - di relazione, 115, 116–120
- grafici,
 - a cilindro, 114
 - conversione, 33–39
 - conversione di sondaggi in, 62–64
 - conversione di tabelle di prezzi in, 33–39
 - forma dei, 100–101
 - gradi di relazione e, 115
 - pendenza dei, 101
 - rapporto di correlazione, 126
 - spese, 116–120
- grafici a dispersione,
 - esempi di 114, 116
 - rapporto di correlazione, 122, 126
 - spese mensili, 116–120

- I**
- indici,
 - coefficiente di correlazione, 120
 - Coefficiente di Cramer, 117, 129
 - dati numerici, 117
 - intraclass. varianza, 117, 123, 124, 126
 - INV.CHI, funzione, 107, 205–206
 - INV.F, funzione, 107
 - INV.NORM funzione, 107
 - INV.NORM.ST funzione, 107
 - INV.T funzione, 107
 - ipotesi alternative,
 - accuratezza delle, 166
 - Coefficiente di Cramer, 186
 - considerazioni, 174
 - esempi di 161, 171–173
 - panoramica, 170–174
 - P-value e, 175–179
 - test tra rapporti di popolazioni, 173
 - ipotesi nulle,
 - Coefficiente di Cramer, 168, 186
 - considerazioni, 174
 - difficoltà di dimostrazione, 174
 - esempi di 167–174
 - impossibilità di scartare, 150, 167, 178, 179, 187
 - in test di correlazione, 172
 - in test di differenze tra rapporti di popolazioni, 173
 - in test d'indipendenza, 172
 - in test di rapporto di correlazione, 172
 - panoramica, 170–174

- P**
- P-value e, 175–179
- scartare, 158, 159, 178
- ipotesi, test di, (v. test di ipotesi statistica)
- istogrammi,
 - ampiezze di classe e, 84, 85
 - esempi di, 39, 83, 84, 154
 - funzione di densità di probabilità, 83–84
 - panoramica, 38–39
 - vantaggi dei 83
 - variabili, 39

- M**
- media,
 - aritmetica, 43, 73, 74
 - armonica, 43
 - con fogli di calcolo, 195–196
 - definizione, 43
 - distribuzione normale e, 87–89
 - distribuzione normale standard e, 89–90
 - esempi, 40–44
 - geometrica, 43
- MEDIA (funzione), 196
- mediana
 - con fogli di calcolo, 195–196
 - definizione, 45
 - esempi 45–47
 - utilizzi, 44

- N**
- Nepero, costante di, 86
- nessuna correlazione, 119
- NORMALIZZA, funzione, 199–201
- normalizzazione 71–72

- O**
- omogeneità, test di (v. test d'ipotesi statistica, test di omogeneità)

- P**
- Pearson, statistica
- chi-quadro di, 132, 152–155, 158
- pendenza, 101
- percentuale, 5, 37, 62, 64

popolazione,
- Coefficiente di Cramer, 145–150, 157, 186
- definizione, 6
- deviazione standard, 52
- rapporti tra, 149, 171, 173
- test d'ipotesi e, 149, 186
- variazioni nella, 145–150, 157, 186
previsioni meteorologiche, 82
prezzi, 33–39
probabilità 81–109
- associate, 104
- definizione, 82
- distribuzione chi-quadro, 99–105, 205–206
- distribuzione e fogli di calcolo, 107–109
- distribuzione F, 106–107
- distribuzione normale, 86–89
- distribuzione normale standard, 89–98, 204–205
- distribuzione t, 106
- gradi di libertà, 99–108
P-value
- ipotesi alternative e, 175–179
- ipotesi nulle 175–179
- test d'indipendenza, 175
- test d'ipotesi, 163, 175–179, 189

Q
qualitativi, dati, (v. dati categorici)
quantitativi, dati, (v. dati numerici)
questionari, 15–19

R
regione critica, 159, 165–167, 187
relazioni,
- frequenza relativa, 36–37, 39
- gradi di, 115, 116–120
- lineari, 120
- non lineari, 120
- rapporto di correlazione, 117, 121–127
- variabili, 112–115
risparmio medio, 46–47

risposte multiple, 28

S

significatività, livello di (a), 159, 163
sondaggi, 4–7
- conversione in grafici, 62–64
- dati categorici, 60–64
- limiti dei 4–7
- test d'indipendenza, 137, 208–211
standard, deviazione, (v. deviazione standard)
standardizzazione, 71–72, 80
statistica descrittiva, 57–58
statisticamente significativo, 187
statistiche,
- definizione, 4
- descrittive, 57–58
- teoria della stima, 4–7
STEP test, 23–25
stima, teoria della (v. teoria della stima)
Sturges, formula di, 55, 56, 58

T

t, distribuzione, (v. distribuzione t)
tabelle,
- dati categorici 60–64
- di contingenza, 128, 130, 135, 151, 153
- di frequenze, (v. frequenze, tabella di)
- distribuzione chi-quadro, 102–105, 205–206
- distribuzione normale, 107
- distribuzione normale standard, 92–93, 104, 108
teoria della stima, 57–58
TEST.CHIQUADR, funzione, 210–211
test d'indipendenza, 208–211. (v. anche test d'ipotesi statistica)
- chi-quadro, 151–169
- esempi di, 149, 171, 184–186
- P-value, 175
- test di omogeneità e, 186

- utilizzi dei, 137, 149
test d'ipotesi statistica 143–189. (v. anche test d'indipendenza)
- chi-quadro, 151–169
- conclusioni, 187
- considerazioni sulla popolazione, 149, 186
- definizione, 149
- esempi di, 149, 168–174
- panoramica, 144–150
- procedure per 150, 175–179
- P-value, 163, 175–179, 189
- test di correlazione, 149, 171, 172
- test di omogeneità, 184–186
- test di differenze tra medie di popolazioni, 149, 171, 173
- test di rapporti di correlazione, 149, 171, 172
- test di rapporti tra popolazioni, 149, 171, 173
- tipi di 149, 171

V

valore alfa (a), 159, 163
valore standard, 65–80, 73, 199–202
valore z, (v. valore standard)
valori,
- deviazione, 74–80
- standard, 65–80, 73, 199–202
- valutazione, 71
variabili, 111–142
- coefficiente di correlazione, 116–120
- Coefficiente di Cramer, 127–138, 141, 142
- gradi di relazione, 115, 116–120
- istogrammi, 39
- rapporto di correlazione, 121–127
- relazioni, 112–115

Z

z, valore. (v. valore standard)

UN'AFFASCINANTE GUIDA ALLA STATISTICA. A FUMETTI!



SE LA STATISTICA NON VI FA DORMIRE SONNI TRANQUILLI O SEMPLICEMENTE VOLETE APPLICARLA ALLA VITA DI TUTTI I GIORNI, "**I MANGA DELLE SCIENZE - STATISTICA**" VI AIUTERÀ A RISOLVERE I VOSTRI PROBLEMI! QUESTA GUIDA A FUMETTI VI METTERÀ IN MEN CHE NON SI DICA SULLA STRADA GIUSTA PER ARRIVARE AL CUORE DELLA MATERIA. E CHE LIBRO DI STATISTICA SAREBBE SENZA ESERCIZI? ALL'INTERNO TROVERETE TUTTI GLI STRUMENTI PER VERIFICARE I VOSTRI PROGRESSI PASSO DOPO PASSO.

SEGUIRETE RUI NEL SUO PERCORSO DI AVVICINAMENTO ALLA STATISTICA GRAZIE ALL'AUTO DEL PAZIENTISSIMO MR YAMAMOTO CHE LE INSEGNERÀ A:

- » CALCOLARE MEDIA, MEDIANA E VALORI STANDARD DEI PUNTEGGI DI BOWLING;
- » TRACCIARE GLI ISTOGRAMMI DEI PREZZI DEL RAMEN;
- » DETERMINARE LA PROBABILITÀ DI PRENDERE IL MASSIMO DEI VOTI IN UN COMPITO DI MATEMATICA;
- » CALCOLARE IL COEFFICIENTE DI CRAMER PER STABILIRE IN CHE MODO RAGAZZI E RAGAZZE PREFERISCONO ESSERE INVITATI AD USCIRE;
- » CAPIRE COME VIENE USATO IL VALORE STANDARD PER NORMALIZZARE I RISULTATI DI UN TEST.

LASCIAVEI GUIDARE DA RUI E DA MR YAMAMOTO E, GRAZIE A BIZZARRI ESEMPI TRATTI DAL MONDO REALE, IN POCO TEMPO PADRONEGGERETE ARGOMENTI CHE TUTTI GLI ALTRI TROVANO DIFFICILI!



Ohmsha

no starch
press

la Repubblica Le Scienze



Pubblicazione settimanale da vendersi esclusivamente
in abbinamento a la Repubblica oppure a Le Scienze.
Supplemento al numero in edicola.
9,90 euro + il prezzo di Repubblica oppure de Le Scienze.