UNIVERSITA' degli STUDI di ROMA
TOR VERGATA

STATISTICA MEDICA

(Prof.ssa CARLA ROSSI)

II-III lezione

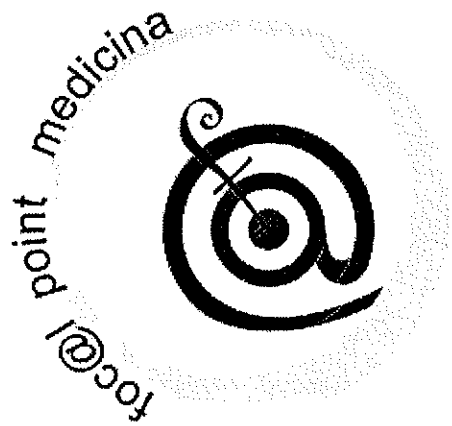


TABELLE DI CONTINGENZA MULTIPLE 2x2

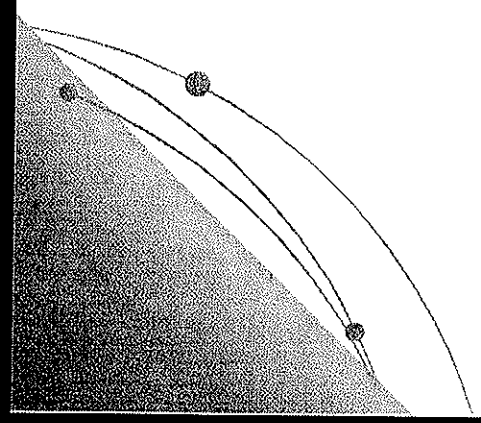
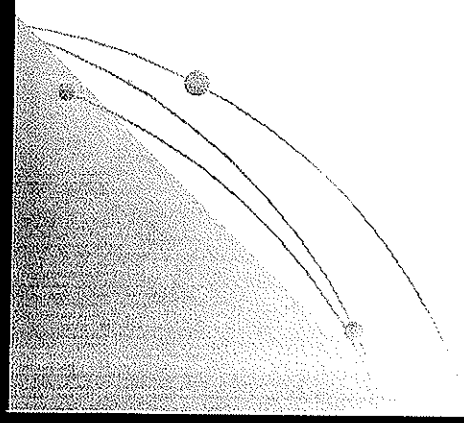


Tabelle di contingenza e odds ratio

Come valutare i fattori di rischio



Attraverso le tabelle di contingenza è stato condotto uno studio sulla relazione tra fumo e stenosi aortica (restringimento dell'aorta che ostacola il flusso sanguigno).

Poiché il sesso è associato ad entrambe le variabili, si prendono in considerazione due tabelle di contingenza 2x2 (**tabelle di contingenza multiple**) che sono il risultato di un singolo studio che è stato STRATIFICATO, o disaggregato, in relazione ad una determinata variabile che si ritiene possa influenzare il risultato (in questo caso il sesso)

FEMMINE

Stenosi Aortica	Fumatrici		totale
	SI	NO	
SI	14	29	43
NO	19	47	66
Totale	33	76	109

MASCHI

Stenosi Aortica	Fumatore		totale
	SI	NO	
SI	37	25	62
NO	24	20	44
Totale	61	45	106

ODDS RATIO

Misura utilizzata per confrontare il rischio di malattia in due differenti situazioni o gruppi

Confrontiamo le frequenze condizionate di riga

FEMMINE

Stenosi Aortica	Fumatrici		totale
	SI	NO	
SI	14	29	43
NO	19	47	66
Totale	33	76	109

Il rapporto delle frequenze condizionate: odd

$$ODD = \frac{F(malattia \mid esposto)}{1 - F(malattia \mid esposto)}$$

FEMMINE

Stenosi Aortica	Fumatrici		
	SI	NO	
ODD	14/19	29/47	43/66
=	0,74	0,62	0,65

L'odd misura quanto più frequente sia la presenza della malattia rispetto alla sua assenza, condizionatamente all'altra variabile. Nell'ultima colonna troviamo il rapporto che misura la frequenza della malattia rispetto alla sua assenza se si trascura l'informazione relativa all'abitudine al fumo.

Come si presentano gli odds

- Si osserva subito che il rapporto per le donne fumatrici è molto maggiore che per le non fumatrici e il rapporto sull'ultima colonna si situa in una posizione intermedia.
- In altre parole lo stato di fumatrice fa aumentare la frequenza di malattia (e diminuire la frequenza della sua assenza).
- Possiamo sintetizzare il risultato in un unico indice calcolando il rapporto tra i due rapporti condizionati.

ODDS RATIO

L'odds ratio è definito come l'odd della malattia tra soggetti esposti diviso l'odd della malattia tra soggetti non esposti:

Si calcola cioè l'odd per gli esposti e si divide per lo stesso per i non esposti.

In simboli:

$$OR = \frac{F(malattia \mid esposto)/[1 - F(malattia \mid esposto)]}{F(malattia \mid non\ esposto)/[1 - F(malattia \mid non\ esposto)]}$$

ODDS RATIO

Supponiamo che i nostri dati – rappresentati da un campione di *n* soggetti – siano organizzati nella forma di una tabella di contingenza 2x2.

	esposti	non esposti	Totale
malattia	*a*	*b*	*a + b*
non malattia	*c*	*d*	*c + d*
totale	*a + c*	*b + d*	*n*

ODDS RATIO

In questo caso si ha che:

F(malattia | esposti) = a / (a + c)

e

F(malattia | non esposti) = b / (b + d)

per cui

1 - F(malattia | esposti) = 1- a / (a + c) = c / (a + c)

e

1 - F(malattia | non esposti) = 1- b / (b + d) = d / (b + d)

ODDS RATIO

Utilizzando questi risultati possiamo esprimere l'odds ratio come:

$$OR = \frac{[a\ /(a+c)]\ /[c\ /(a+c)]}{[b\ /(b+d)]\ /[d\ /(b+d)]} = \frac{a\ /\ c}{b\ /\ d} = \frac{ad}{bc}$$

Questo è il rapporto del prodotto crociato dei valori nella tabella 2x2

OR maschi

MASCHI

Stenosi Aortica	Fumatore		totale
	SI	NO	
SI	37	25	62
NO	24	20	44
Totale	61	45	106

$$OR_M = \frac{(37)(20)}{(25)(24)} = 1{,}23$$

OR femmine

FEMMINE

Stenosi Aortica	Fumatrici		totale
	SI	NO	
SI	14	29	43
NO	19	47	66
Totale	33	76	109

$$OR_F = \frac{(14)(47)}{(29)(19)} = 1{,}19$$

Da questi dati si evince lo stesso andamento in entrambi i gruppi della popolazione:

Sia per i maschi che per le femmine l'odds di sviluppare una stenosi aortica è maggiore tra i fumatori rispetto ai non fumatori

Possiamo tentare di combinare le informazioni provenienti dalle due tabelle per giungere ad una unica conclusione che verifichi la correlazione tra fumo e stenosi aortica

Stenosi Aortica	Fumatore		totale
	SI	NO	
SI	51	54	105
NO	43	67	110
Totale	94	121	215

$$OR = \frac{(51)(67)}{(54)(43)} = 1{,}47$$

PARADOSSO DI SIMPSON

PARADOSSO DI SIMPSON

Ignorando l'influenza del sesso, la forza dell'associazione tra fumo e stenosi aortica appare maggiore rispetto a quella ottenuta separatamente per i maschi e per le femmine

IL PARADOSSO DI SIMPSON SI VERIFICA QUANDO IL RAPPORTO TRA DUE VARIABILI E' INFLUENZATO DALLA PRESENZA DI UNA TERZA VARIABILE DI CONFONDIMENTO (in questo caso il sesso) NELLA RELAZIONE TRA ESPOSIZIONE E MALATTIA

Se la tabella non è solo 2x2?

- Nel caso la tabella sia più complessa possiamo definire gli odd relazionando le diverse modalità della patologia tra loro nei vari modi possibili.
- Osserviamo, nel caso della tabella sulla gravità degli attacchi cardiaci e presenza dell'anticorpo, che questo provoca sempre odds ratio maggiori dell'unità. La presenza dell'anticorpo raddoppia, per esempio, il rischio di avere un attacco grave piuttosto che lieve...e così via

ODDS (grave/lieve)	
0,57	0,28
ODDS RATIO	2,05
ODDS (grave/medio)	
0,68	0,42
ODDS RATIO	1,62
ODDS (medio/lieve)	
0,83	0,66
ODDS RATIO	1,27

Esempio di utilizzo della formula di Bayes

Consideriamo un caso reale, anche se molto semplificato. Supponiamo che un paziente che accusa disturbi gastrici si presenti in un centro clinico per avere una diagnosi ed una conseguente terapia. Supponiamo che, per quel genere di disturbi, ci siano tre possibili cause, indagabili con un primo test clinico, che denotiamo con T: A="il paziente soffre di ulcera", B="il paziente soffre di gastrite", C="il paziente non ha né ulcera, né gastrite". Supponiamo anche che tali eventi si presentino con frequenze note tra gli individui che soffrono degli stessi sintomi. Possiamo, pertanto, utilizzare tali frequenze per valutare le probabilità a priori di A, B e C. Siano P(A)=0.25, P(B)=0.35 e P(C)=0.40.

Osserviamo che, per come è formulato il problema, lo scopo iniziale del medico è, probabilmente, di escludere una delle malattie "Ulcera" o "Gastrite", mediante il test T, passando, eventualmente, ad analisi più approfondite nel caso la diagnosi sia C, per considerare altri stati patologici diversi dai primi due. L'evento C considerato è definito in modo tale che l'insieme degli eventi A, B e C siano incompatibili ed esaustivi. In tutti i casi analoghi occorre procedere similmente in modo che sia completamente definito l'insieme dei risultati considerati possibili e di interesse per il problema e che i risultati possibili siano tra loro incompatibili ed esaustivi.

Supponiamo che il test T dia risultato positivo "+" nel 90% dei casi di ulcera, nel 40% dei casi di gastrite e nel 10% di tutti gli altri casi. Questo permette di scrivere: P(+/A)=0.9, P(+/B)=0.4, P(+/C)=0.1, da cui si ricavano immediatamente anche le probabilità del risultato negativo, che risultano: P(-/A)=0.1, P(-/B)=0.6, P(-/C)=0.9.

Supponiamo ora che il risultato effettivo del test sia positivo, ci chiediamo qual è la probabilità a posteriori di A, B e C.

Per ottenere la risposta dobbiamo solo utilizzare per gli eventi di interesse la formula di Bayes riportata sopra che, per esempio, per l'evento A si scrive:

$$P(A / ?) ? \frac{P(? / A)P(A)}{P(? / A)P(A) ? P(? / B)P(B) ? P(? / C)P(C)}$$

Effettuando i calcoli si ottiene: P(A/+)=0.555, P(B/+)=0.346, P(C/+)=0.099. Come si vede il test T avvalora l'ipotesi che i sintomi siano causati da ulcera o gastrite, ma non permette di ben discriminare tra queste due, occorrono altre analisi per precisare la diagnosi. Il medico, comunque, ha ottenuto un'informazione che permette di concentrarsi su due possibili cause escludendone altre.

Per completezza controlliamo che cosa succede se il risultato del test è negativo. Applicando di nuovo la formula di Bayes, otteniamo: P(A/-)=0.042, P(B/-)=0.353, P(C/-)=0.605. In questo caso il risultato del test permette di escludere praticamente l'ulcera, ma occorrono altre indagini per precisare la diagnosi. C'è ancora da rilevare che le probabilità a posteriori della gastrite (B) sono comunque quasi uguali alla probabilità a priori. Questo dipende dal fatto che la probabilità di risultato del test positivo in caso di gastrite (0.6) non è molto diversa da quella di risultato negativo (0.4), in altre parole il test non è adatto ad identificare la gastrite.

Elementi di Calcolo delle Probabilità

- Modelli per i fenomeni aleatori (incerti)

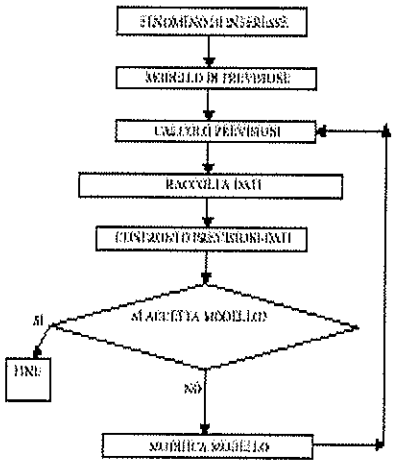
Carla Rossi
rossi@mat.uniroma2.it

I modelli

- Il cammino della scienza percorre incessantemente 3 passi:
- Osservazioni
- Costruzione di un modello che descriva o spieghi le osservazioni
- Utilizzo del modello per prevedere osservazioni future

Se le osservazioni future non sono in accordo con le previsioni, il modello non è adeguato e deve essere sostituito o modificato.

Schema generale



- Lo stesso schema viene seguito da un clinico nell'effettuare una diagnosi medica sulla base dei sintomi (osservazioni), modello (ipotesi sulla patologia compatibile), raccolta dati (indagini cliniche sul soggetto) e confronto con l'ipotesi fatta ed eventuale modifica dell'ipotesi.

Il procedimento avviene in generale in condizioni di incertezza spesso anche nelle conclusioni, la diagnosi è accettata fino a prova contraria e lo stesso avviene per i modelli che si applicano in campo scientifico, per esempio nei fenomeni fisici. Sono molto rare le situazioni in cui si opera con certezza.

- *Un uomo stava guardando un ritratto. Qualcuno gli chiese: "di chi è il ritratto che stai guardando? ". L'uomo rispose: "Fratelli e sorelle io non ne ho, ma il padre di quest'uomo è figlio di mio padre".*

- *E' possibile* ***decidere con certezza*** *se è vero o falso l'evento A= "l'uomo sta guardando il suo ritratto" e l'evento B=" l'uomo sta guardando il ritratto di suo figlio"?.*

•*Da quanto detto* ***si deduce con certezza*** *che, dato che l'uomo che risponde non ha né fratelli, né sorelle, il figlio di suo padre è solo lui. Ne segue che, se il padre dell'uomo del ritratto è figlio del padre dell'uomo che guarda, il ritratto non può che rappresentare il figlio dell'uomo che guarda. L'evento* ***A*** *è, pertanto,* ***certamente falso*** *e l'evento* ***B*** *è* ***certamente vero****.*

Decisioni cliniche

- Le **decisioni cliniche**, non solo sono inevitabili, ma devono anche essere prese **in condizioni di incertezza**
- L'incertezza deriva da molti fattori:
- Errori nei dati clinici
- Ambiguità dei dati e variazioni di interpretazione
- Incertezza nelle relazioni tra informazione clinica e presenza della malattia
- Incertezza circa gli effetti del trattamento

Rischio, incertezza, probabilità

- Valutare il rischio che consegue una qualunque decisione assunta in condizioni di incertezza implica la valutazione e il confronto di probabilità.
- Alcune ricerche dimostrano che le persone tendono a sovrastimare le probabilità di situazioni poco familiari, catastrofiche o molto "chiacchierate", e a sottostimare le probabilità di eventi più usuali e meno spettacolari.

'Non è sindrome dei Balcani ma del veterano'

- "Non è sindrome dei Balcani": è questo il verdetto di uno studio condotto dal Ministero della difesa belga su 16mila soldati che hanno prestato servizio nella ex Jugoslavia.
- Non essendo stati riscontrati più casi di cancro fra i militari mandati nei Balcani che fra quelli rimasti a casa, risulta impossibile stabilire un legame con l'uranio impoverito.
- Un terzo degli intervistati soffre, invece, di fatica cronica e disturbi della concentrazione: la cosiddetta "sindrome del veterano".

Lo spazio campione

- Per affrontare un problema di decisione in situazioni incerte, un passo fondamentale è lo studio di tutti i possibili esiti che, in alcuni casi, possono anche essere esplicitamente enumerati semplificando notevolmente il problema.
- Il caso più semplice si ha quando gli esiti sono solo 2: tumore si, tumore no; Balcani si, Balcani no....

Gli eventi

- Nel caso in cui i possibili risultati da considerare siano solo 2 si parla di evento.
- Un evento è una proposizione che può essere vera o falsa e di cui non è noto il valore logico al momento di assumere la decisione.
- E' un evento la proposizione: il sig. X ha prestato servizio nel Balcani (prima dell'intervista).
- Dopo l'intervista si potrà affermare con certezza se l'evento è vero o falso.

Evento certo e impossibile

- Con riferimento ad un certo *stato di informazione*, *per un certo individuo*, cioè all'insieme dei fatti che si suppongono noti (dati), un evento è ***certo***, e viene solitamente denotato con Ω, *quando l'individuo ritiene che, come conseguenza necessaria dei fatti, sia* ***VERO*** *(l'uomo sta guardando il ritratto di suo figlio)*
- ***impossibile***, denotato con $\varnothing$, quando ritiene, come conseguenza necessaria dei fatti, che sia ***FALSO*** *(l'uomo sta guardando il suo ritratto).*

Il caso generale

- *Un evento in generale è* ***possibile*** *quando nessuno dei due esiti può essere stabilito a priori (e quindi nessuno può essere escluso).*
- ***La distinzione in certo, impossibile e possibile ha carattere soggettivo e temporaneo. Agli eventi si attribuisce una misura di probabilità.***

La probabilità è una misura delle aspettative nel verificarsi di un evento

- ***il valore della probabilità*** *è la* ***misura*** *(un numero) che esprime l'****opinione*** *del soggetto (decisore) in merito al* ***verificarsi di un ben determinato evento A****,* ovvero esprime ***il suo grado di fiducia nel verificarsi dell'evento****.*
- Dato un evento A la sua probabilità si indica con P(A).

Prime regole di calcolo

- Si ha:

$$P(\Omega)=1 \text{ e } P(\varnothing)=0$$

- Per ogni evento possibile A:

$$0<P(A)<1$$

Come misurare l'incertezza

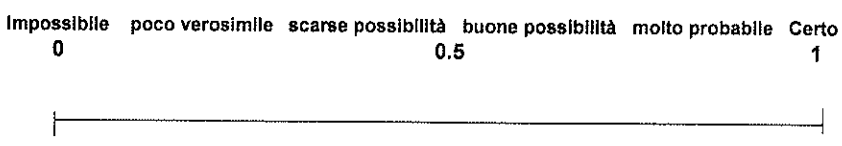
- Supponiamo di considerare il seguente evento:

A="domani a Roma pioverà".

- Possiamo in molti modi comunicare la misura della nostra aspettativa rispetto al verificarsi di questo evento; possiamo, per esempio, dire che:

- è certo;
- è molto probabile;
- ci sono buone possibilità;
- ci sono scarse possibilità;
- è poco verosimile;
- è impossibile.

Scala dei giudizi relativi alle aspettative nel verificarsi di un evento



Operazioni sugli eventi

- E' possibile combinare gli eventi:
- L'evento negazione di A $\bar{A}$ è vero quando A è falso e viceversa.
- Se A=il paziente che sto per visitare ha un tumore:
- $\bar{A}$=il paziente che sto per visitare non ha un tumore.
- $P(\bar{A})=1-P(A)$.

Intersezione

- Si dirà *intersezione degli eventi A e B* e si indicherà con $C=A \cap B$, l'evento C che è vero se sono veri entrambi gli eventi A e B, mentre risulta falso se almeno uno dei due è falso.

$C=A \cap B$

- A=il prossimo paziente che visiterò ha più di 50 anni
- B=il prossimo paziente che visiterò è una donna
- $C=A \cap B$ =il prossimo paziente che visiterò è una donna di almeno 50 anni.

Unione

- Dati due eventi A e B, l'*unione degli eventi* $D=A \cup B$, è l'evento D che è vero se uno almeno degli eventi A o B è vero.

D= A∪B

- A=il prossimo paziente che visiterò ha più di 50 anni
- B=il prossimo paziente che visiterò è una donna
- D= A∪B=il prossimo paziente che visiterò è una donna o ha almeno 50 anni.

Eventi incompatibili ed esaustivi

- Si dirà che due eventi *A e B sono incompatibili* se il verificarsi di uno esclude il verificarsi dell'altro.
- Due eventi *A e B si diranno esaustivi* se è inevitabile che almeno uno si verifichi.
- Non cambia nulla se gli eventi da prendere in considerazione sono più di 2.
- Quando si elencano i possibili esiti di uno studio si suddividono le varie possibilità in eventi incompatibili ed esaustivi.

Come costruire lo spazio campione degli eventi incompatibili ed esaustivi

- Per costruire una tabella che illustri i risultati di uno studio sull'efficacia dei caschi protettivi per bicicletta nella prevenzione dei traumi cranici, si possono considerare varie costruzioni, la più semplice considera solo gli eventi:
- A=il ciclista usava il casco protettivo
- B=nella caduta ha riportato un trauma cranico
- Però si potrebbe considerare il tipo di casco e la gravità del trauma...e così via.

Il caso più semplice

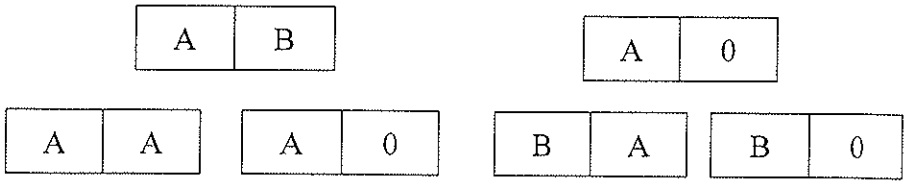
Trauma Cranico	Casco Protettivo		totale
	SI	NO	
SI	17	218	235
NO	130	428	558
Totale	147	646	**793**

Valutazione classica

- La ***valutazione*** di una probabilità si basa talvolta su ***semplici giudizi qualitativi di equiprobabilità***.
- Se i risultati elementari (incompatibili ed esaustivi) di uno spazio campione sono giudicati da un individuo ugualmente probabili, la valutazione della probabilità di ciascuno è p=1/n, e un risultato (evento E) che sia costituito di k degli n eventi equiprobabili ha probabilità p=k/n.
- ***La probabilità di un evento E è data dal rapporto tra il numero dei casi favorevoli al verificarsi di E (k) e quello dei casi possibili (n).***

Le applicazioni genetiche

- Proviamo a prevedere il gruppo sanguigno di un individuo con genitori con gruppi sanguigni AB e A0.
- In base alle leggi di Mendel è noto che il figlio eredita una sola lettera da ognuno dei genitori con uguale probabilità;



Dal genotipo al fenotipo

- combinazioni genotipiche possibili per il figlio:
- AA, A0, AB, B0, ognuna con probabilità p=1/4.
- La determinazione standard del gruppo sanguigno (fenotipica) non fa distinzione tra i gruppi AA e A0.
- Pertanto l'individuo in questione apparterrà al gruppo sanguigno A=AA$\cup$A0 con probabilità p=1/2 (2 casi favorevoli su 4 possibili), al gruppo sanguigno AB con probabilità p=1/4, al gruppo sanguigno B con probabilità p=1/4.

Regole di calcolo

- ***Se A e B sono incompatibili:***

$$P(A \cup B)=P(A)+P(B)$$

- ***Altrimenti:***

$$P(A \cup B)=P(A)+P(B)-P(A \cap B)$$

Esperimenti statistici in genetica

- Mendel, sperimentando sugli incroci fra piselli odorosi con cotiledoni di due colori, trovò nella discendenza degli ibridi 6022 semi gialli e 2001 semi verdi, cioè due numeri che stanno fra loro nella proporzione di

  3,01:1.

- Ripetendo l'esperimento nel corso degli anni altri ottennero proporzioni leggermente diverse, ma molto simili tra loro.

I cotiledoni

- ***I cotiledoni*** *sono foglie embrionali, con struttura semplificata e funzione di nutrimento dell'embrione dall'inizio della germinazione al momento in cui si sviluppano la radice e le prime foglie. Hanno funzione di riserva.*

Valutazione di probabilità geniche

- Poiché dovrebbe trattarsi di differenze dovute alla variabilità campionaria, quei rapporti vanno considerati tutti uguali e forniscono la valutazione statistica della probabilità del seme "giallo" e del seme "verde":

$$P(G)=3/4$$

$$P(V)=1/4$$

La regola di valutazione statistica o frequentista

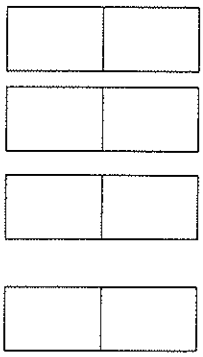
- Se si considera una successione di eventi E_1, E_2,..,E_n,.., in qualche modo analoghi, e si indica con f_n la frequenza relativa del numero di volte che si è osservato il verificarsi dell'evento sui primi n, si è generalmente portati a prevedere un andamento simile per la frequenza di analoghi eventi futuri. In altre parole si valuta che la probabilità del verificarsi di ognuno degli analoghi eventi futuri sia $p=f_n$.

Regola di valutazione frequentista

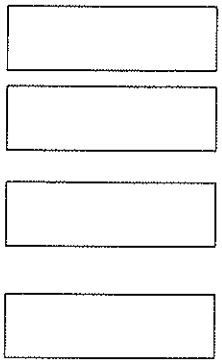
- ***per valutare la probabilità di un evento si calcola la frequenza relativa osservata su un numero (possibilmente grande) (*) di eventi quanto più possibile analoghi a quello che interessa e si utilizza tale valore come probabilità.***

Il modello teorico: giallo dominante

Genotipo



Fenotipo



Dal modello teorico al calcolo

- Una volta stabilito il modello teorico, la probabilità si calcola anche considerando gli eventi elementari equiprobabili e calcolando casi favorevoli e casi possibili.
- Il risultato per il caso dei semi gialli e verdi è lo stesso:

$$P(G)=3/4$$
$$P(V)=1/4$$

Non c'è un unico modello

- In alcuni casi è possibile valutare la probabilità di un certo evento A sia su base classica che su base frequentista. Non è detto, in tal caso, che le due valutazioni coincidano.
- Supponiamo di voler valutare la probabilità che un nascituro sia di sesso maschile (evento A).
- Possiamo ragionare in diversi modi.

Equiprobabilità

- Possiamo immaginare che ci sia simmetria tra i due casi possibili, perché a priori non c'è motivo di ritenere il contrario, e valutare P(A)=1/2.

La genetica

- Un individuo è di sesso maschile se possiede una coppia di cromosomi XY,
- di sesso femminile se possiede una coppia di cromosomi XX;
- ogni nuovo concepito eredita a caso in modo simmetrico un cromosoma della coppia del padre e uno della madre.

Come si determina il sesso geneticamente

- Da una coppia di genitori XY-XX sono possibili le seguenti combinazioni per i cromosomi del nascituro:
- XX, XX, YX, YX,
- tutte con probabilità 1/4.
- Due combinazioni danno luogo ad un individuo di sesso maschile e due ad uno di sesso femminile:
- P(A)=1/2.

Valutazione statistica

- Supponiamo, poi, di voler ragionare statisticamente. Prendiamo le statistiche ufficiali sulle nascite e consultiamo la tabella in cui si riportano i dati dei nati suddivisi per sesso.
- Scopriamo, allora, che le frequenze di nascite maschili e femminili sono regolari nel tempo e sempre diverse da 1/2, con un'eccedenza non trascurabile di nascite maschili.

Nascono più maschi

- Se assumiamo come valutazione di P(A) la frequenza relativa delle nascite maschili osservate nel periodo considerato, otteniamo, su base statistica, P(A)=0,515.
- Con diverse impostazioni o, meglio, utilizzando modelli e informazioni diverse, si arriva, dunque, a valutazioni diverse.

Modificare un modello

- E' possibile che la discrepanza sia dovuta a una eccessiva semplicità del modello genetico per la trasmissione dei cromosomi X e Y.
- La probabilità di trasmettere Y è lievemente più elevata della probabilità di trasmettere X.
- Non c'e' perfetta simmetria come per gli altri cromosomi.

Come influisce un'informazione

- La valutazione della probabilità di un evento dipende dalle situazioni in cui viene effettuata e dalle informazioni che si possiedono sull'evento stesso e su altri eventi in qualche modo collegati.
- Se cambia la situazione o si acquisiscono nuove informazioni sugli eventi di interesse, cambia anche la valutazione di probabilità.

Story Name: Calcium and Blood Pressure

- Does increasing calcium intake reduce blood pressure? Observational studies suggest that there is a link, and that it is strongest in African-American men. Twenty-one African-American men participated in an experiment to test this hypothesis. Ten of the men took a calcium supplement for 12 weeks while the remaining 11 men received a placebo. Researchers measured the blood pressure of each subject before and after the 12-week period. The experiment was double-blind.

I dati

Treatment	Begin	End	Treatment	Begin	End
Calcium	107	100	Placebo	123	124
Calcium	110	114	Placebo	109	97
Calcium	123	105	Placebo	112	113
Calcium	129	112	Placebo	102	105
Calcium	112	115	Placebo	98	95
Calcium	111	116	Placebo	114	119
Calcium	107	106	Placebo	119	114
Calcium	112	102	Placebo	112	114
Calcium	136	125	Placebo	110	121
Calcium	102	104	Placebo	117	118
			Placebo	130	133

Lo schema teorico

- In questo caso possiamo dire che l'evento A è :
  A=assunzione di calcio
- e l'evento B è:
- B=diminuzione della pressione
- I dati possono essere riassunti in una tabella di contingenza, dopo aver contato i "successi" (diminuzioni di pressione) per i due trattamenti.

Tabella di contingenza

	Diminuzione		
Trattamento	**si**	**no**	**Totale**
si	6	4	10
no	3	8	11
Totale	9	12	21

Confrontiamo le probabilità di successo

- Se valutiamo statisticamente le probabilità, otteniamo:
- P(B)=9/21=0,43, per il totale
- P(B)=6/10=0,60 per chi ha avuto il trattamento con il calcio (evento A)=P(B/A)
- P(B)=3/11=0,27 per chi non ha avuto il trattamento con il calcio (evento $\bar{A}$)
- =P(B/ $\bar{A}$)

Osserviamo comunque che il numero di casi è molto piccolo e non sarebbe giustificata la valutazione statistica della probabilità

Normalizzando per riga

Osserviamo che i valori delle probabilità sono stati ottenuti normalizzando la tabella per riga (distribuzioni condizionate di riga).

	Diminuzione		
Trattamento	si	no	Totale
si	6/10	4/10	1
no	3/11	8/11	1
Totale	9/21	12/21	1

Probabilità condizionate

- P(B/A) si legge probabilità di B dato A e indica la valutazione della probabilità dell'evento B quando si è venuti a conoscenza del verificarsi dell'evento A.
- P(B/A) è la probabilità condizionata di B dato A.
- Lo stesso vale per l'altra probabilità condizionata.

E se il trattamento non fosse influente?

- Se l'informazione sul verificarsi di A non alterasse la valutazione della probabilità di B, i due eventi A e B si direbbero indipendenti, la tabella coinciderebbe con la tabella attesa e la probabilità condizionata sarebbe uguale alla probabilità calcolata sul totale dei casi:

$$P(B/A)=P(B/\bar{A})=P(B)=9/21=0,43$$

Regola del prodotto

- Per gli eventi indipendenti vale la regola:

$$P(A\cap B)=P(A)P(B)$$

- che possiamo utilizzare per valutare la diversa probabilità di trasmissione del cromosoma Y.
- In generale invece si ha:

$$P(A\cap B)=P(A)P(B/A)=P(A/B)P(B)$$

Quanto vale P(Y)?

- Le 4 combinazioni possibili sono:

  XX, XX, YX, YX
- Si ha statisticamente:
- P(XY)=0,515=P(X dalla madre $\cap$ Y dal padre)= 1*P(Y dal padre),
- per cui:
- P(Y dal padre)=0,515
- che coincide con la probabilità di nascituro maschio calcolata statisticamente.

Eventi correlati

- Se tra eventi c'è dipendenza, può risultare P(B/A)>P(B) oppure P(B/A)<P(B), nella prima situazione si parla di *eventi correlati positivamente*: il verificarsi di A rende più probabile il verificarsi anche di B; nel secondo caso si parla di *eventi correlati negativamente:* il verificarsi di A ostacola il verificarsi anche di B.

Calcium story

- Dalla valutazione risulta che:
- il trattamento con calcio è correlato positivamente con l'abbassamento della pressione: $P(B/A)>P(B)$;
- il trattamento con il placebo è correlato negativamente con l'abbassamento della pressione: $P(B/\bar{A})<P(B)$.

Le distribuzioni condizionate di colonna

Se normalizziamo la tabella per colonna, otteniamo le probabilità condizionate del trattamento data l'informazione sull'esito (P(A/B))

	Diminuzione		
Trattamento	si	no	Totale
si	6/9	4/12	10/21
no	3/9	8/12	11/21
Totale	1	1	1

Le probabilità condizionate di colonna rispondono alla domanda: con che probabilità un soggetto ha ricevuto il trattamento con il calcio dato che si è osservata una diminuzione di pressione?

Formula della probabilità totale

- L'evento B si può decomporre nell'unione dei due eventi incompatibili:
- $B=(B \cap A) \cup (B \cap \bar{A})$,
- calcolando la probabilità si ottiene la formula della probabilità totale:
- $P(B)=P(B/A)P(A)+P(B/\bar{A})P(\bar{A})$.
- I due addendi che compaiono nella somma costituiscono le probabilità composte.

Formula di Bayes

•Se esplicitiamo P(A/B) dalla relazione:

$$P(A)P(B/A)=P(A/B)P(B)$$

• e sostituiamo a P(B) al denominatore la probabilità composta, otteniamo la formula di Bayes:

$$P(A/B)=\frac{P(B/A)P(A)}{P(B)}=\frac{P(B/A)P(A)}{P(B/A)P(A)+P(B/\bar{A})P(\bar{A})}$$

Campi di applicazione

- La formula di Bayes trova applicazione nel campo della diagnostica clinica, della valutazione di colpevolezza di un imputato in un certo processo, delle cause per il riconoscimento di paternità, delle procedure di screening di popolazione e in molte altre situazioni.

Un problema diagnostico

- Dobbiamo effettuare una diagnosi di sieropositivà al virus HCV (è il virus che provoca l'epatite C) su un individuo proveniente da una popolazione in cui la proporzione di HCV positivi è p=0,006.
- Indichiamo con M l'evento:
  M=il soggetto è sieropositivo $\Longrightarrow$ P(M)=0,006.
- P(M) è la probabilità a priori di M (prevalenza di HCV nella popolazione).

Accuratezza del test

- Il test utilizzato fornisce una risposta positiva (+) nel 99% dei casi in cui l'individuo è malato (M) e una risposta negativa (-) nel 99% dei casi in cui l'individuo è sano (S):
- P(+/M)=0,99; P(-/S)=0,99;
- P(-/M)=0,01; P(+/S)=0,01.
- P(+/M) è la sensibilità del test
- P(-/S) è la specificità del test

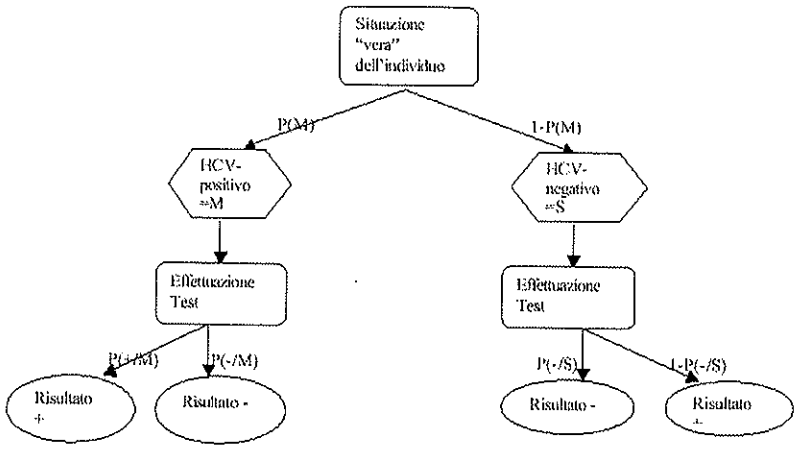
Probabilità a posteriori

- Calcoliamo, mediante la formula di Bayes, la probabilità che un individuo, scelto a caso dalla popolazione di partenza, sia M, essendo risultato positivo al test:

$$P(M/+)=\frac{P(+/M)P(M)}{P(+/M)P(M)+P(+/\overline{M})P(\overline{M})}=\frac{(0,99)(0,006)}{(0,99)(0,006)+(0,01)(0,994)}$$

- P(M/+)=0,374. P(M/+) è la probabilità a posteriori di M.

Rappresentazione grafica



Calcolo lungo il grafo

- Scorrendo i rami dell'albero, si calcola il prodotto delle probabilità dei diversi cammini e si ottengono le probabilità che compaiono al numeratore della formula di Bayes.
- Sommando le probabilità dei diversi rami che conducono ad uno stesso risultato (+ o -), si ottiene la probabilità che compare al denominatore.

Strategia diagnostica

- Dato il margine di errore del test e la "bassa" prevalenza dell'infezione nella popolazione solo il 37% circa degli individui risultati positivi al test è effettivamente infetto, gli altri sono falsi positivi.
- Possiamo "azzardare" una diagnosi o è meglio raccogliere altre informazioni?
- E' evidente che è molto più ragionevole acquisire altre informazioni. In casi come quello considerato, infatti, si richiede generalmente di ripetere il medesimo test a distanza di qualche tempo.

Iterazione del procedimento

- Se si ripete il procedimento effettuando un nuovo test, si riutilizza la formula di Bayes sostituendo alla probabilità a priori per il primo test P(M), la probabilità a posteriori P(M/+), che diventa la probabilità a priori per il secondo test.
- Infatti, nel momento in cui si effettua tale secondo test, il risultato del primo è ormai noto.
- Effettuando il calcolo, nell'ipotesi che si abbia un secondo risultato positivo per il test, si ottiene P(M/++)=0,983 che consente la diagnosi.

Se il secondo test è negativo

- Calcoliamo per completezza anche la probabilità P(M/+-), relativa all'altro risultato possibile per il secondo test.
- Si ottiene: P(M/+-)=0,006.
- Il risultato negativo del secondo test ha, come ci si attendeva, annullato l'influenza del primo riproducendo il valore della probabilità a priori.
- Questo risultato dipende dal fatto che specificità e sensibilità del test sono uguali. Con valori diversi si ottengono risultati diversi.

I MODELLI STATISTICI

http://www.mat.uniroma2.it/~rossi/dida.html

1

VARIABILI ALEATORIE

- In un certo numero di fenomeni, il risultato di una osservazione è "grandezza" rappresentata da un numero (variabili quantitative).
- Ciascuna di queste grandezze può assumere valori diversi, e non si può stabilire in anticipo quale valore la grandezza assumerà in una singola osservazione; sarà solo possibile valutare a priori con che probabilità essa può assumere ciascuno dei valori possibili.

2

VARIABILI ALEATORIE

- grandezze rappresentate da numeri, non noti a priori, sono:
  - ✓ il numero di chiamate effettuate da abbonati ai telefoni in un certo intervallo definito di tempo, in un certo luogo;
  - ✓ la percentuale di risposte "*sì*" a una domanda binaria di un questionario;
  - ✓ il tempo di disintegrazione di un atomo;
  - ✓ la percentuale di pezzi difettosi fabbricati da una macchina in un certo intervallo di tempo fissato;
  - ✓ il numero di battiti cardiaci per minuto di un certo individuo dopo una prova da sforzo.

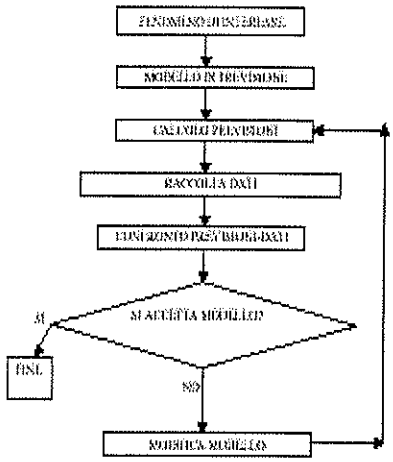
3

VARIABILI ALEATORIE

- Ogni volta che associamo ad un'unità non ancora osservata del collettivo in esame una caratteristica che lo contraddistingue, espressa attraverso un valore numerico e non nota a priori, otteniamo una ***variabile aleatoria***.
- La variabile aleatoria è il modello teorico della variabile statistica.
- Le variabili aleatorie sono discrete (numeri interi) o continue (valori su una scala continua).

4

Modellizzazione



- Lo schema viene seguito, per esempio, per fissare i limiti di normalità per le risposte delle analisi di laboratorio su misure biologiche come il numero di globuli rossi o bianchi (variabile discreta) o il tasso glicemico a digiuno o il colesterolo totale (variabili continue)....

5

Lanciamo i dadi

- Supponiamo di avere due dadi regolari distinguibili le cui facce sono numerate da 1 a 6.
- Se i dadi vengono lanciati e i numeri sulle facce superiori vengono registrati come *coppia ordinata*, l'insieme dei possibili risultati elementari è rappresentato da uno spazio campione contenente i 36 elementi, rappresentati nella seguente tabella.

6

Spazio campione

Secondo dado

Primo dado						
	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

7

Probabilità degli eventi elementari

- Ogni faccia di uno specifico dado ha probabilità di verificarsi 1/6,
- Ogni coppia ha probabilità 1/36, data l'indipendenza tra i dadi:

$$P(D1=i \cap D2=j)=(1/6)(1/6)$$

per la regola del prodotto.

8

La variabile somma delle facce

primo dado	secondo dado					
	1	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

9

La distribuzione di probabilità

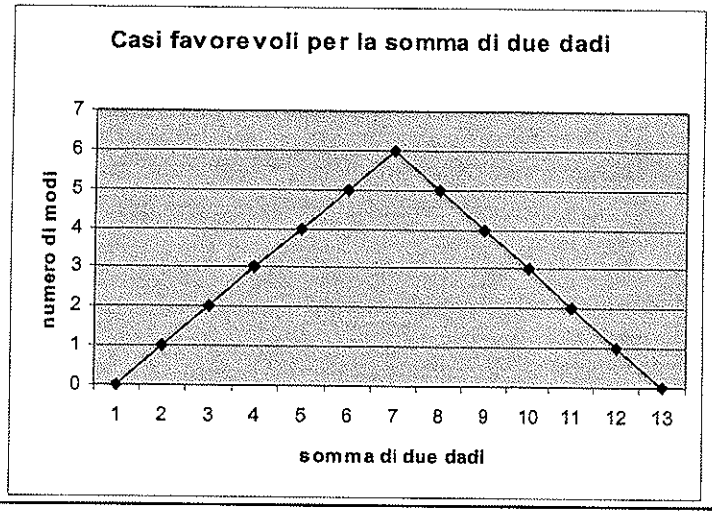
- Possiamo sintetizzare il risultato del calcolo delle probabilità dei singoli punteggi nella tabellina seguente in cui nella prima riga sono riportati i valori possibili e nella seconda le rispettive probabilità.

X	2	3	4	5	6	7	8	9	10	11	12
p	1/ 36	2/ 36	3/ 36	4/ 36	5/ 36	6/ 36	5/ 36	4/ 36	3/ 36	2/ 36	1/ 36

10

Distribuzione triangolare

Se si riportano su un grafico cartesiano, in ascissa i valori di X e in ordinata i corrispondenti numeratori di p, si ottiene un grafico di tipo triangolare che rappresenta l'andamento della distribuzione



11

Distribuzione di probabilità

- Più in generale, se X può assumere solo n valori distinti x_i (i = 1, 2, ..., n) e se indichiamo con p_i la sua probabilità (p_i=P(X= x_i)), si ha:
- $p_1 + p_2 + p_3 + \ldots + p_n = 1;$
- cioè: *la somma delle probabilità di tutti i valori assumibili è uguale a 1.*
- Questa probabilità totale *si distribuisce*, in un certo modo, tra i valori diversi di X e, precisamente, la proporzione p_i corrisponde al valore x_i (i = 1, 2, ..., n).

12

Il modo più semplice per fornire la distribuzione di probabilità di una variabile X, con un numero finito di valori possibili, è una tabella, in cui sono enumerati tutti i valori possibili di X e le corrispondenti probabilità,

X	x_1	x_2	...	x_i	...	x_n
p	p_1	p_2	...	p_i	...	p_n

13

La funzione di ripartizione

- Anche per una variabile aleatoria X, in analogia al caso statistico, si definisce la ***funzione di ripartizione*** *che ad ogni valore x, assumibile da X, associa la probabilità che X sia non superiore a x*:

$$F(x) = P(X \leq x)$$

14

Calcolo

Nel caso della somma delle facce di due dadi la funzione di ripartizione risulta:

X	2	3	4	5	6	7	8	9	10	11	12
F(X)	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	1

15

Descrizione sintetica

- Spesso è sufficiente indicare soltanto alcuni parametri numerici che caratterizzano i tratti essenziali della distribuzione di interesse: per esempio, un ***indice di posizione*** *attorno al quale si raggruppano i valori possibili della variabile aleatoria; un indice che caratterizza la* ***dispersione*** *di questi valori attorno al* ***valore medio****; ecc.*
- ***Sono questi valori caratteristici, o sintetici, della distribuzione che ne forniscono un'immagine riassuntiva, sufficiente per alcuni scopi.***

16

La media di una v.a.discreta

- Il valore atteso o media indica il valore centrale e rappresenta un'utile sintesi della distribuzione in un unico valore.
- Si indica con E(X) e si calcola come media ponderata analogamente al caso di una tabella statistica:

$$E(X)= x_1p_1 + x_2 p_2 + x_3 p_3 + \ldots + x_n p_n$$

$$E(X) = \sum_{i=1}^{n} x_i p_i$$

- Nel caso della somma dei dadi la media è 7.

17

La media di una v.a.

3. ***La media della somma di due variabili aleatorie è uguale alla somma delle medie delle singole variabili.*** *Cioè:*

$$E(X + Y) = E(X) + E(Y).$$

18

La variabile aleatoria scarto

- La differenza tra X e la sua media:

$$X - E(X).$$

si chiama ***deviazione,*** o ***variabile aleatoria scarto.***

- Si dimostra che:

$$E(X - E(X)) = 0.$$

⇒***La media della variabile aleatoria scarto è nulla.***

19

La varianza di una v.a.discreta

- Si definisce ***varianza di X***, e si indica con $\sigma^2(X)$, *la media del quadrato della variabile aleatoria scarto*:

$$\sigma^2(X) = \sum_{i=1}^{n} (x_i - E(X))^2 p_i$$

20

Deviazione standard di una v.a.discreta

- La ***dispersione*** *di una distribuzione di probabilità è sintetizzata dallo* ***scarto quadratico medio***, **o** ***scarto standard*** **o** ***deviazione standard*** **di X**, data dalla *radice quadrata (aritmetica) della varianza*:

$$\sigma(X) = \sqrt{[\sigma^2(X)]},$$

21

Coefficiente di variazione di una v.a.

- Analogamente a quanto visto in campo statistico si definisce il ***coefficiente di variazione***:

$$CV(X) = 100\sigma(X)/E(X).$$

Analogamente al caso statistico si definiscono anche la mediana, la moda e i quantili di una distribuzione di probabilità.

22

Variabili standardizzate

- La variabile Z :

$$Z = \frac{X - E(X)}{\sigma(X)}$$

- ha media 0 e varianza e scarto standard 1.
- Tale variabile si dice standardizzata
- La stessa operazione si può effettuare sulle variabili statistiche

23

I parametri

- In generale si cerca di determinare delle forme facili per i modelli statistici, cioè di esprimere la distribuzione di probabilità di una variabile aleatoria attraverso formule matematiche il più possibile semplici che dipendono dalla variabile e da un certo numero di costanti che sono denotate parametri della distribuzione.

24

Modelli particolari

- In realta', per dati biomedici, ci sono certe situazioni standard che ricorrono spesso, per quanto riguarda il modello probabilistico.
- Queste situazioni vengono riassunte in un numero ristretto di distribuzioni di probabilita' con nomi particolari.
- Quale distribuzione sia adatta a una certa situazione, dipende dalla natura delle osservazioni.

25

Tipi di dati

- Si possono distinguere osservazioni
- - discrete: normalmente a valori interi, tipicamente risultato di un conteggio (numero di guariti in un gruppo di pazienti di entita' nota, numero di batteri/ml in una cultura diluita, ecc)
- - continue: tipicamente osservazioni di variabili fisiologiche, ma anche durate temporali e qualunque cosa sia concettualmente possibile osservare e rappresentare in modo decimale.

26

Problemi con variabili conteggio

- Molti problemi con variabili discrete riguardano situazioni di conteggio.
- I modelli sono diversi e dipendono dalle ipotesi e dalle informazioni disponibili.
- Possiamo conoscere la probabilità di una risposta positiva ad un tipo di osservazione o sperimentazione, o l'incidenza di una caratteristica per unità di tempo, di spazio...

27

Il modello binomiale

- Di tutte le ostriche di un allevamento il 23% contiene perle.
- Qual'è la probabilità che, scegliendo a caso e indipendentemente, 12 ostriche più di 5 contengano perle?
- Le chiamate ad un pronto soccorso riguardano per il 32% incidenti domestici.
- Qual è la probabilità che sulle prossime 20 chiamate ci siano più di 10 casi di incidenti domestici?
- A questi quesiti si risponde utilizzando la distribuzione binomiale.

28

Il modello di Poisson

- Il numero medio di piantine per metro quadrato in un appezzamento è di 3. Qual è la probabilità che in un'area di 1 metro quadrato vi siano meno di 5 piantine?
- Il numero medio di chiamate ad un pronto soccorso è di 10 l'ora.
- Qual è la probabilità che nelle prossime due ore ci siano più di 30 chiamate?
- A questi quesiti si risponde utilizzando la distribuzione di Poisson.

29

Distribuzione binomiale

- E' il modello adatto ad una situazione che si può schematizzare mediante una serie di n esperimenti indipendenti tra loro,
- ogni esperimento può dar luogo a due esiti, "successo" (S) e "insuccesso" (I)
- in ogni esperimento si considera costante uguale a p la probabilità di S.

30

Probabilità di una sequenza

- Per l'ipotesi di indipendenza, la probabilità di una specificata sequenza di S(uccesso) e I(nsuccesso) (SSSIISISS) è il prodotto di tanti fattori p quanti sono i valori di S nella sequenza e di tanti fattori (1-p) quanti sono i valori I nella sequenza:

$$P(SSSIISISS)=p^6(1-p)^3$$

31

Distribuzione binomiale

- La variabile aleatoria X = "numero di S nella sequenza di n prove" può assumere i valori {0, 1, 2, ..., n} e tutte le sequenze con uno stesso numero di S sono tra loro equivalenti (casi favorevoli).
- Dato k={0, 1, 2, , ..., n} si ha allora:
- $P(X = k) = \binom{n}{k} p^k(1-p)^{n-k}$, dove $\binom{n}{k}$ significa : $\frac{n!}{k!(n-k)!}$
- E conta le sequenze con un numero fissato di S in qualunque ordine (coefficiente binomiale).

32

Contiamo le sequenze equivalenti

- Per imparare a contare le sequenze possiamo considerare un numero qualsiasi di lanci di una moneta e la vincita di un euro per ogni testa ottenuta, calcolando i diversi valori ottenibili e il corrispondente numero di elementi dello spazio campione (casi favorevoli).
- Il valore della vincita conta il numero di teste che possiamo identificare con il numero di Successi.

33

Lanciamo la moneta

- 1. In un solo lancio si hanno solo due elementi dello spazio campione: T o C, cui corrispondono valori della vincita 1 o 0, ognuno ottenibile in un solo modo.
- 2. A due lanci corrispondono i seguenti elementi per lo spazio campione: TT, TC, CT, CC, cui corrispondono valori della vincita: 2, 1, 0, ottenibili rispettivamente in un solo modo, in due modi possibili, in un solo modo.

34

Tre lanci

Possiamo associare ad ognuno degli elementi dello spazio campionario la vincita associata e contare i modi:

TTT: 3, TTC: 2, TCT: 2, CTT: 2, TCC: 1, CTC: 1, CCT: 1, CCC: 0

Valori	0	1	2	3
Numero di modi per ottenerli	1	3	3	1

35

Quattro lanci

- E' evidente che il valore della vincita rimane invariato se il quarto lancio ha come risultato *C* e aumenta di 1 se il risultato di tale lancio è T.
- Possiamo quindi contare il numero di possibilità a partire dai risultati dei primi tre lanci.
- Generalizzando possiamo rappresentare i vari casi con una tabella triangolare (triangolo di Tartaglia).

36

Valori (1 lancio)	0	1				
Numero di modi	1	1				
Valori (2 lanci)	0	1	2			
Numero di modi	1	2	1			
Valori (3 lanci)	0	1	2	3		
Numero di modi	1	3	3	1		
Valori (4 lanci)	0	1	2	3	4	
Numero di modi	1	4	6	4	1	
Valori (5 lanci)	0	1	2	3	4	5
Numero di modi	1	5	10	10	5	1

37

Il triangolo di Tartaglia

1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1

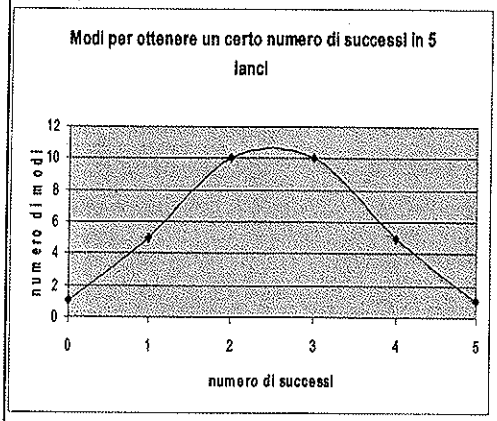
- In sintesi si scrive in forma triangolare e, passando da una riga alla successiva, all'interno della cornice di valori unitari (0 successi o n successi su n prove), i valori sono la somma dei due valori della riga sovrastante:

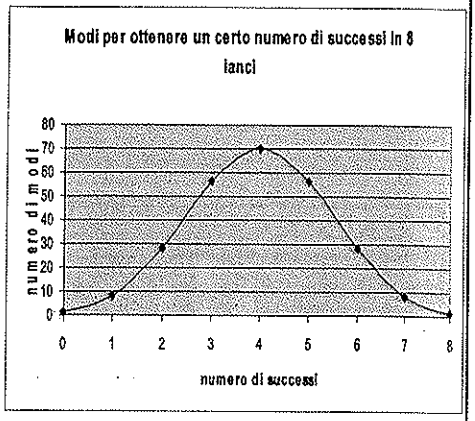
$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

38

Confrontiamo l'andamento del grafico

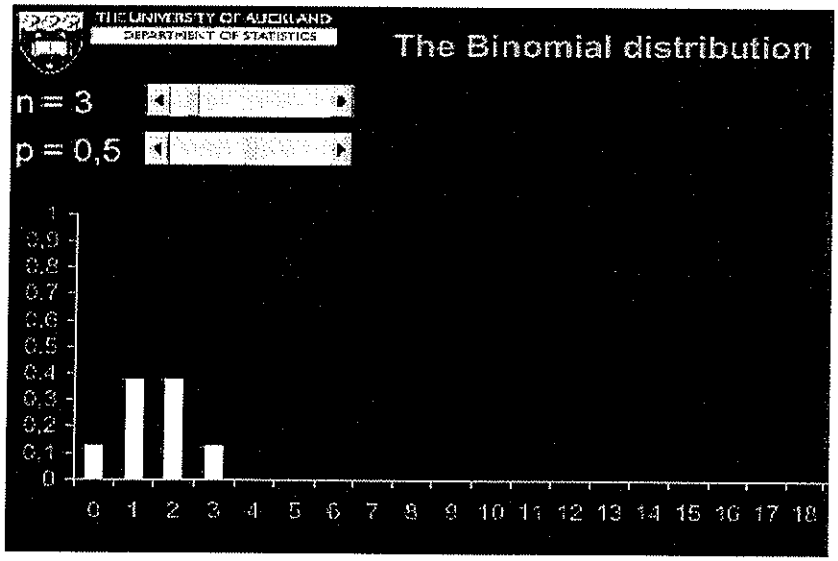
All'aumentare dei lanci l'andamento del grafico (sempre simmetrico) che rappresenta il numero di possibilità per ogni valore del numero di successi assume forma sempre più campaniforme.





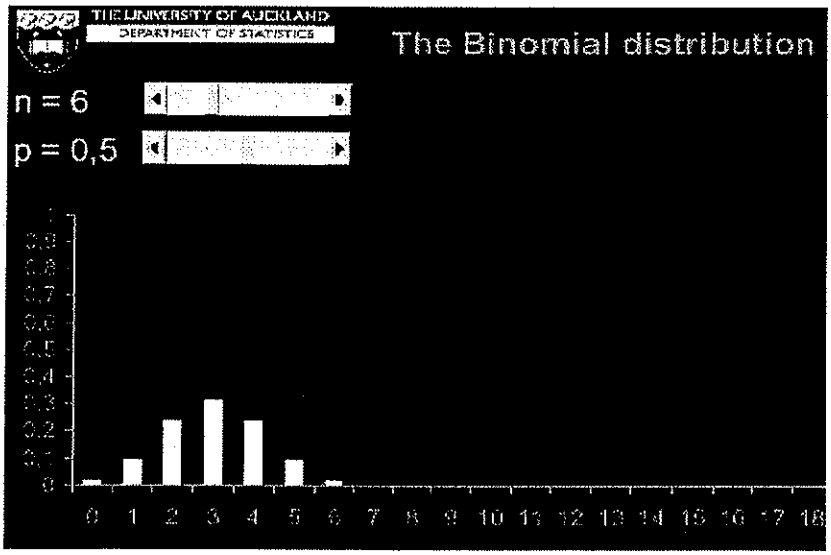
39

Esempio 1



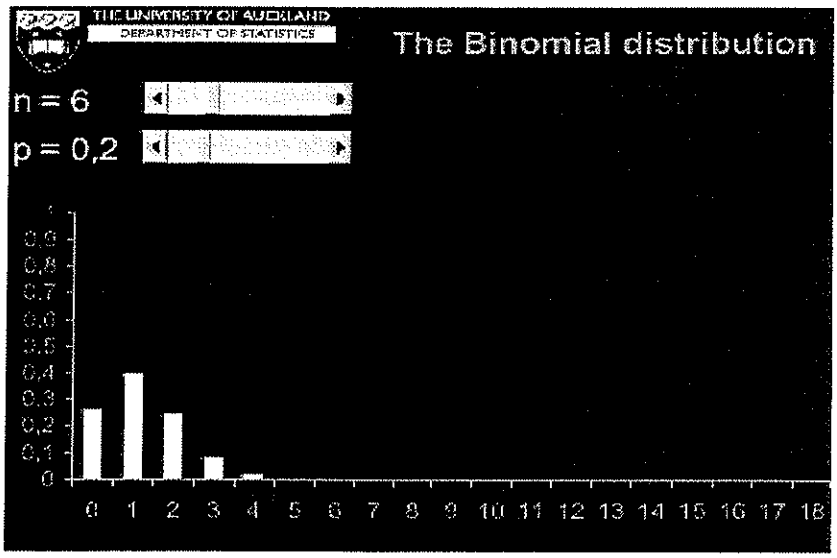
40

Esempio 2



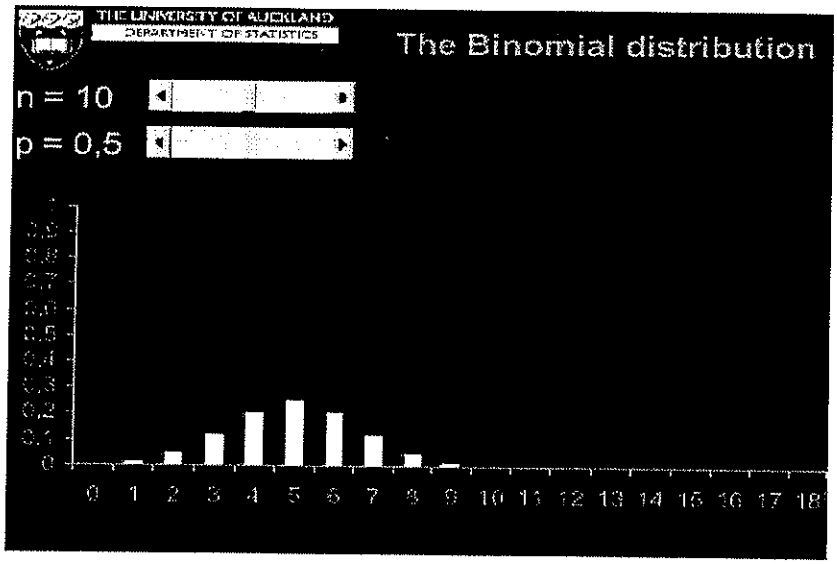
41

Esempio 3



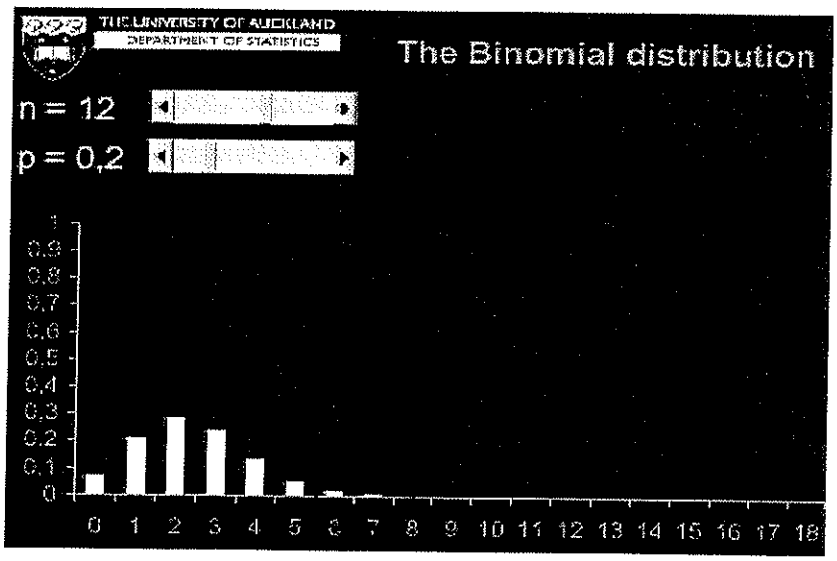
42

Esempio 4



43

Ostriche



44

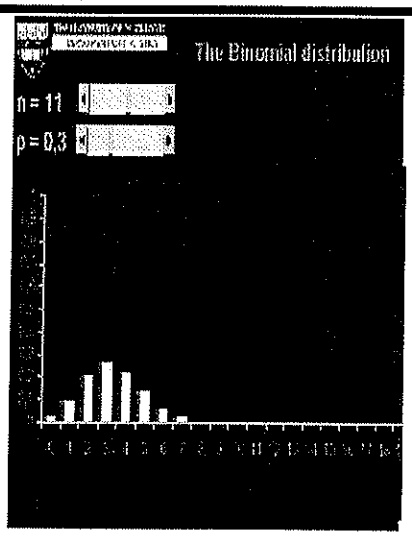
Media e varianza

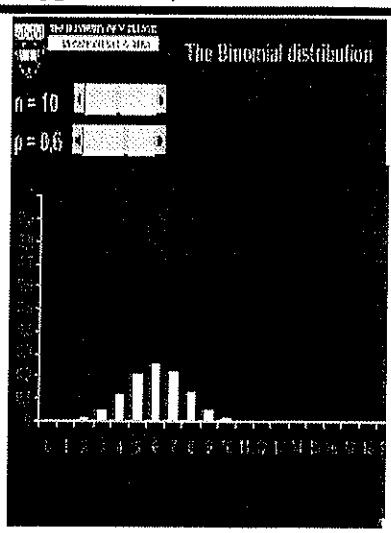
- Per una distribuzione binomiale si ha:
  - $E(X)=np$
  - $\sigma^2(X)=np(1-p)$
  - $\sigma(X)=\sqrt{[np(1-p)]}$

45

Calcium story

La figura a sinistra rappresenta il modello teorico per i soggetti sottoposti a placebo, quello sulla destra il modello per i soggetti sottoposti a trattamento





6

Distribuzione di Poisson

- Ci sono situazioni in cui abbiamo a disposizione l'informazione sul numero di unità coinvolte in una specifica situazione di interesse per unità di tempo, o di spazio. Sappiamo cioè quante osservazioni si hanno in media per unità di tempo o di spazio.
- Tale informazione è sintetizzata in un parametro λ, e il modello di cui si può fare uso è la *distribuzione di Poisson di parametro* λ.

47

Ipotesi di base

- Per poter utilizzare il modello di Poisson devono essere soddisfatte alcune ipotesi:
- La probabilità che si verifichi un evento in un intervallo molto piccolo è proporzionale alla lunghezza dell'intervallo;
- gli eventi sono tra loro indipendenti;
- intervalli disgiunti sono tra loro indipendenti.

48

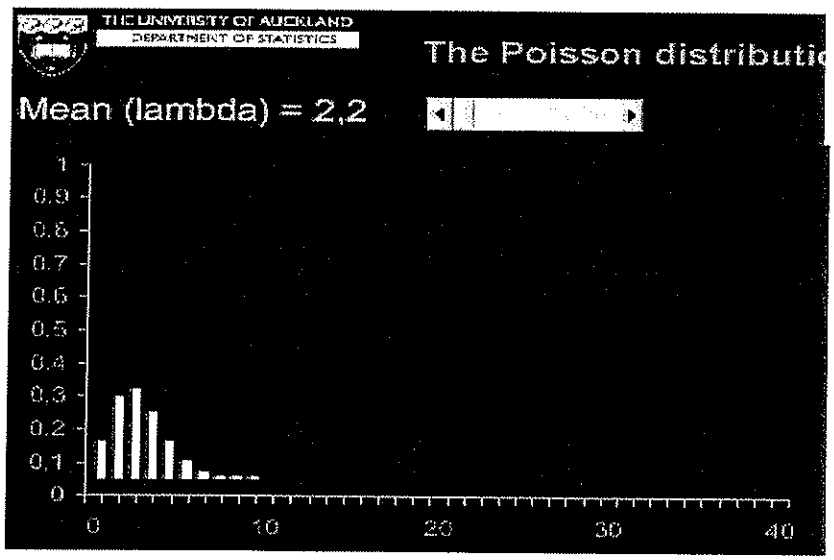
Media e varianza

- Tale distribuzione assegna ad ogni k={0, 1, 2, , ...} la probabilità:

- $P(X = k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$
-
- si ha:
- $E(X) = \lambda$;
- $\sigma^2(X) = \lambda$.

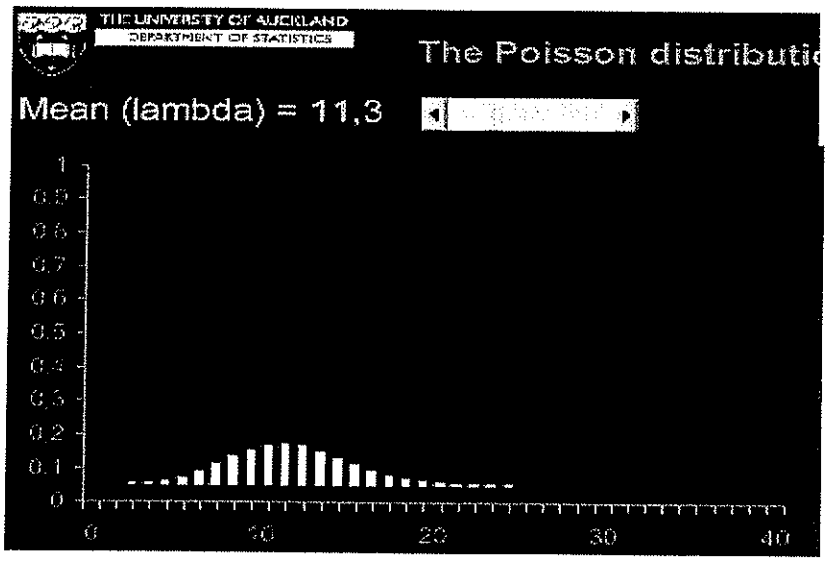
49

Esempio 1



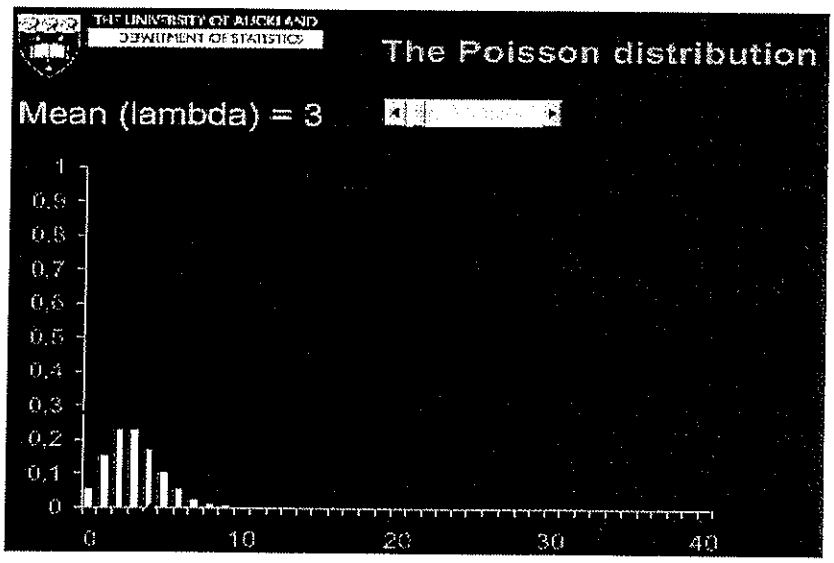
50

Esempio 2



51

Piantine



52

Morti per calcio di cavallo

- The classic Poisson example is the data set of von Bortkiewicz (1898), for the chance of a Prussian cavalryman being killed by the kick of a horse. Ten army corps were observed over 20 years, giving a total of 200 observations of one corps for a one year period. The period of observation is thus one year. The total deaths from horse kicks were 122, and the average deaths per year per corps was thus 122/200 = 0,61.

53

Quanto vale λ?

- This is a rate of less than 1.
- In any given year, we expect to observe, well, not exactly 0,61 deaths in one corps but sometimes none, sometimes one, occasionally two, perhaps once in a while three, and (we might intuitively expect) very rarely any more.
- Here, then, is the classic Poisson situation: a rare event, whose average rate is known, and is small, with observations made over many small intervals of time.

54

Calcoliamo le probabilità

- Let us see if our formula gives a close fit for the actual Prussian data, where λ equals the average number expected per year for the whole sample, and the successive terms of the Poisson formula are the successive probabilities. The first term of the formula, $e^{-\lambda}$ or rather, in this case, $e^{-0,61}$, should give the probability that 0 deaths occurred in a given year. This works out to 0,543.

55

Calcoliamo le previsioni

- This probability, multiplied by 200 (the total years observed) gives the number of zero-death years to be expected in our sample. The result is 108,6 or, to the nearest observable number, 109 years. It turns out that 109 is exactly the number of years in which the Prussian data recorded no deaths from horse kicks. The match between expected and actual values is not merely good, it is perfect.

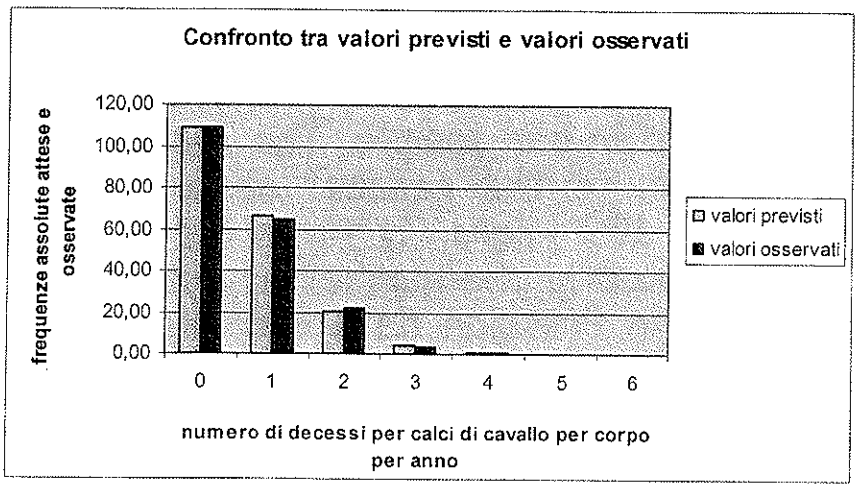
56

E=previsioni; A=osservazioni

Deaths	E	A
0	108,67	109
1	66,29	65
2	20,22	22
3	4,11	3
4	0,63	1
5	0,08	0
6	0,01	0

57

Confrontiamo con i dati



58

Confrontiamo

- The Poisson frequency profile gives the expectation (E) when the events in question are indeed random.
- Comparing that expectation with our actual results (A), we judge that the Prussian data set appears to be the result of random events.
- There is no reason to suspect any systematic cause, or any connection between separate events.

59

Morti casuali?

- These deaths, then, just happened. (If bad training in one corps were responsible, for instance, the pattern of deaths should be more clustered). It is the ability of the Poisson Distribution to give a model for stuff that "just happens," that accounts for its power in statistics.

60

Esempio 2: visite mediche

La distribuzione di frequenza del numero di visite mediche nel periodo di un anno di 24 pazienti dimessi dopo un intervento è quella riportata nella tabella.

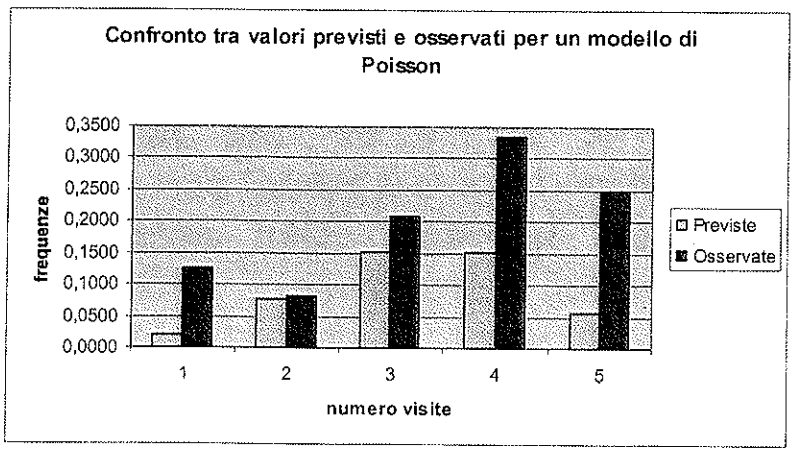
E' possibile approssimarla con un modello di Poisson, essendo la media empirica pari a 3,92?

Numero di visite	Frequenza
0	3
1	2
2	5
5	8
7	6
Totale	24

61

Confrontiamo

- La popolazione reale non si può considerare omogenea. Indizio importante "bimodalità".



62

Disomogeneità

- Il gruppo di soggetti è presumibilmente costituito da almeno due gruppi omogenei:
- I pazienti che non hanno avuto complicazioni (o non sono ansiosi) e effettuano poche visite;
- I pazienti che hanno avuto complicazioni (o sono ansiosi) ed effettuano molte visite.

63

Funzione di densità di una v.a. continua

- *Per descrivere la distribuzione di una variabile aleatoria continua, non si può più assegnare una probabilità positiva ad ogni valore possibile.* Si assume allora di poter specificare una funzione, detta **funzione densità di probabilità, f(x)**, *definita sull'insieme D⊆R di valori possibili, che sia non negativa dappertutto e per la quale si abbia*

$$\int_D f(x)dx = 1$$

64

Funzione di densità di una v.a. continua

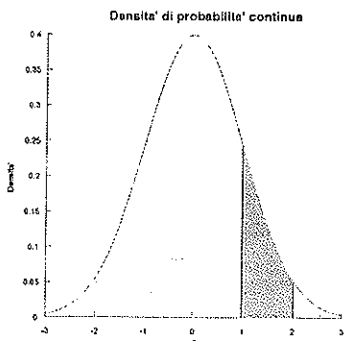
- Tale funzione esprime, in analogia con la funzione densità di materia per un corpo materiale, il modo di distribuirsi della probabilità totale sull'insieme dei valori assumibili da X. Per calcolare la probabilità $P(X \in A)$, per un dato sottoinsieme A di D, occorre procedere per integrazione:

$$P(X \in A) = \int_A f(x)dx$$

65

Probabilità di un evento (intervallo)

- Nel grafico e' rappresentata una distribuzione continua, con valori possibili tra -3 e 3. L'evento {risultato tra 1 e 2} ha probabilita' rappresentata dall'area segnata;
- e' difficile vedere immediatamente che quest'area rappresenta circa 13.6% dell'area totale sotto la curva.



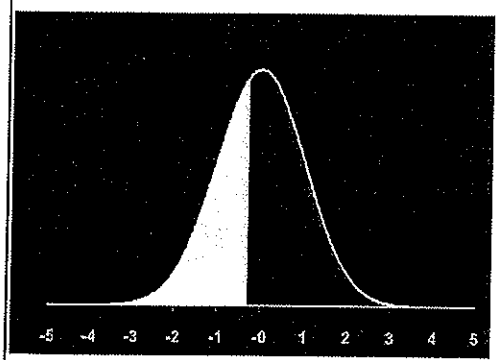
66

Calcolo con la funzione di ripartizione

- Per fare calcoli di aree, come quella sopra, ci si serve spesso della funzione di ripartizione corrispondente alla densita' di probabilita'.
- Si tratta di una formula o di una tabella che per ogni valore x dell'ascissa indica l'area sovrastante l'intervallo tra l'estrema sinistra del grafico e il valore x.
- Per ogni x, il valore indicato corrisponde dunque alla probabilita' dell'evento A={il risultato e' minore o uguale a x}.

67

Calcolo della funzione di ripartizione



- La figura a fianco mostra in giallo l'area corrispondente al valore della funzione di ripartizione F(x) calcolata nel punto
- x=-0,3.

68

Funzione di ripartizione di una v.a. continua

- E' possibile definire formalmente la funzione di ripartizione:

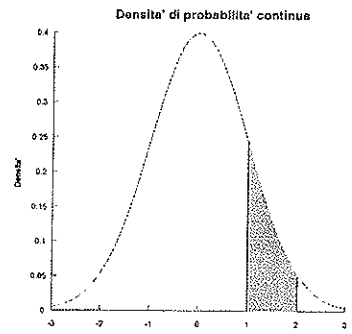
★ $F(x)=P(X\leq x)=\int_{-\infty}^{x} f(y)dy$

Per qualunque intervallo [a,b] si ha, allora:

$$P(a\leq X\leq b)=\int_{a}^{b} f(y)dy = F(a)-F(b)$$

69

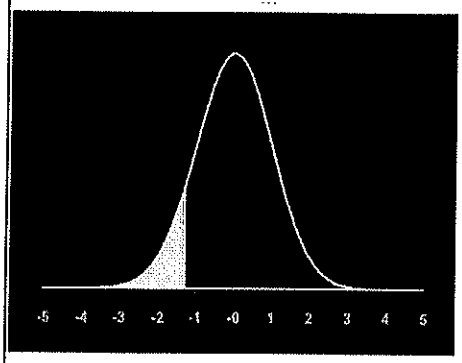
Calcoliamo



- Nel nostro caso, avremmo visto che la funzione di ripartizione dava il valore 0,977 per x = 2 e il valore 0,841 per x =1. La differenza tra questi due valori e' proprio l'area compresa tra x =1 e x = 2.

70

Calcolo dei decili



- Se si pone F(x)=0,1 e si determina di conseguenza il valore di x, si ottiene x=-1,28 che rappresenta il primo decile, cioè il valore che suddivide la distribuzione in 2 parti, un decimo a sinistra di x e 9 decimi a destra.

71

Percentili

- I quartili e percentili si ricavano analogamente dalla funzione di ripartizione.
- Per esempio, la tabella seguente riporta tutti I decili della distribuzione considerata.

72

Area, partendo da sinistra	Ascissa = decile
0.05	-1.64
0.10	-1.28
0.15	-1.04
0.20	-0.84
0.25	-0.67
0.30	-0.53
0.35	-0.40
0.40	-0.26
0.45	-0.13
0.50	0.00
0.55	0.13
0.60	0.26
0.65	0.40
0.70	0.53
0.75	0.67
0.80	0.84
0.85	1.04
0.90	1.28
0.95	1.64

73

Media di una v.a. continua

- nel caso continuo il **valor medio** si definisce:

$$E(X) = \int_D x f(x) dx$$

- La media soddisfa le stesse proprietà della media nel caso discreto.

74

Varianza di una v.a. continua

- La ***varianza***, che abbiamo definito come $E[(X - E(X))^2]$, risulta data da:

$$\sigma^2(X) = \int_D (x - E(X))^2 f(x)dx$$

la deviazione standard e il coefficiente di variazione si calcolano di conseguenza.

75

Variabili standardizzate

- Anche per le variabili continue, si verifica che, come nel caso statistico, la variabile Z così definita:

$$Z = \frac{X - E(X)}{\sigma(X)}$$

- ha media pari a 0 e varianza e scarto standard pari a 1.
- Tale variabile si dice standardizzata.

76

La distribuzione normale

- La distribuzione normale (riconoscibile dalla curva a forma di campana) è la più usata tra tutte le distribuzioni, perché molte misure si distribuiscono in modo approssimativamente gaussiano. La sua derivazione matematica fu presentata per la prima volta da De Moivre nel 1733, ma è spesso riportata come la distribuzione Gaussiana, dal nome di Carl Gauss (1777-1855), che ricavò anche la sua equazione da uno studio degli errori nelle misure ripetute della stessa quantità (Fisica).

77

La distribuzione normale

- Nel 1844 Quetelet annunciò che la *legge degli astronomi* (legge normale) era applicabile anche alla distribuzione di caratteristiche umane, come l'altezza e la circonferenza della vita, e così estese l'uso della funzione esponenziale a campana anche al campo della biometria, non relegandola più esclusivamente allo studio degli errori.

78

La distribuzione normale

- Esaminando le liste dei dati somatici di migliaia di coscritti francesi, Quetelet notò che le misure dell'altezza e del peso si distribuiscono, non solo simmetricamente da una parte e dall'altra della loro media, ma secondo la curva degli errori di osservazione di Gauss.
- con considerazioni teoriche e riscontri empirici, verificò che la lunghezza delle ordinate della curva di frequenza è all'incirca proporzionale ai vari termini di sviluppo del binomio $(\frac{1}{2} + \frac{1}{2})^n$, dove n è uguale al numero delle classi di frequenza meno uno (distribuzione binomiale con 0=0,5).

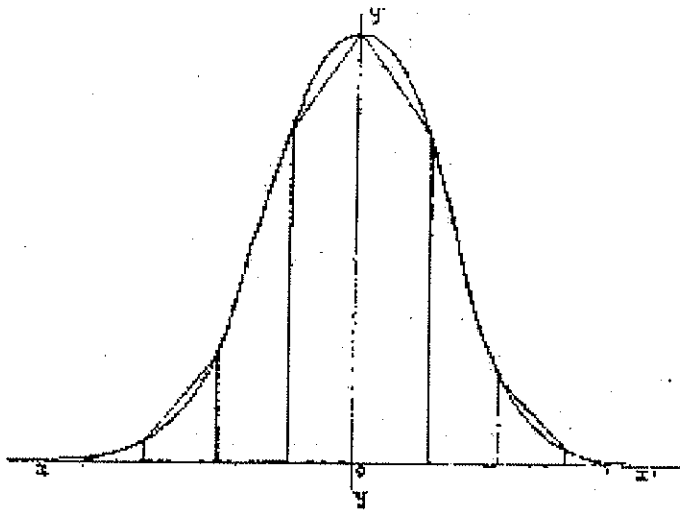
79

La distribuzione normale

- Inoltre Quetelet fece un accostamento ancor più significativo di quello fra coscritti ed errori di osservazione, rilevando l'analogia fra gli scarti dei fenomeni fisici rispetto alle proprie leggi e le variazioni biometriche rispetto alle proprie medie.
- Come in fisica gli scarti non escludono l'esistenza di una legge, così le variazioni biometriche non escludono l'esistenza di una legge di natura (andamento medio).

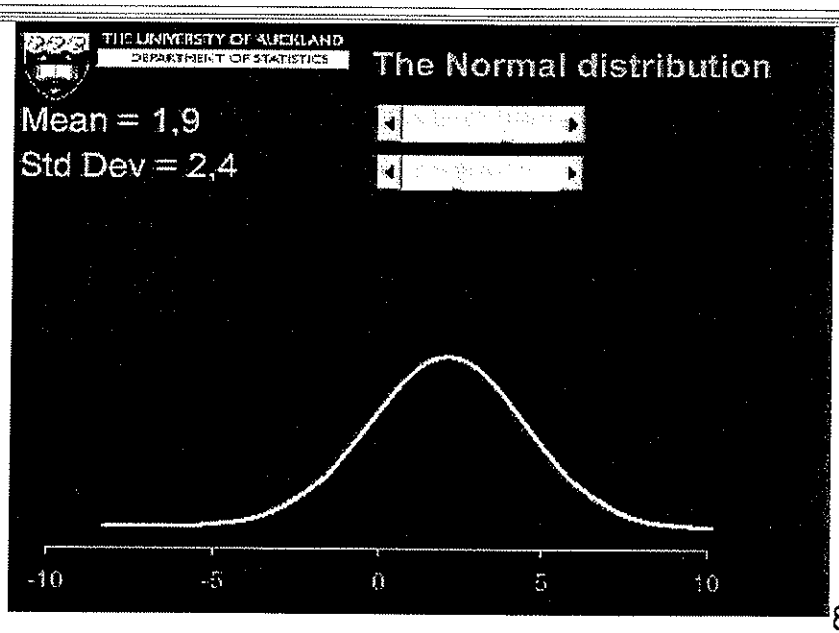
80

La distribuzione normale



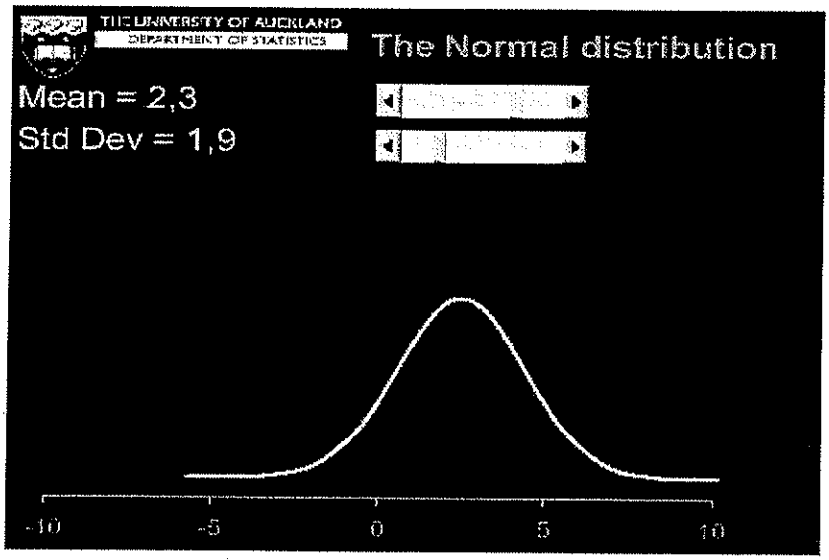
81

Esempio 1



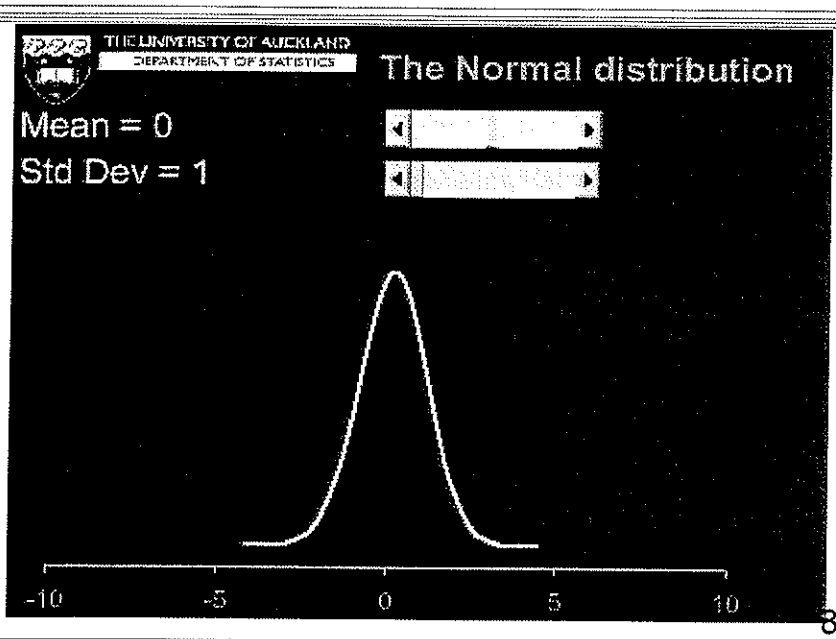
82

Esempio 2



83

La curva normale standardizzata



84

La densità normale

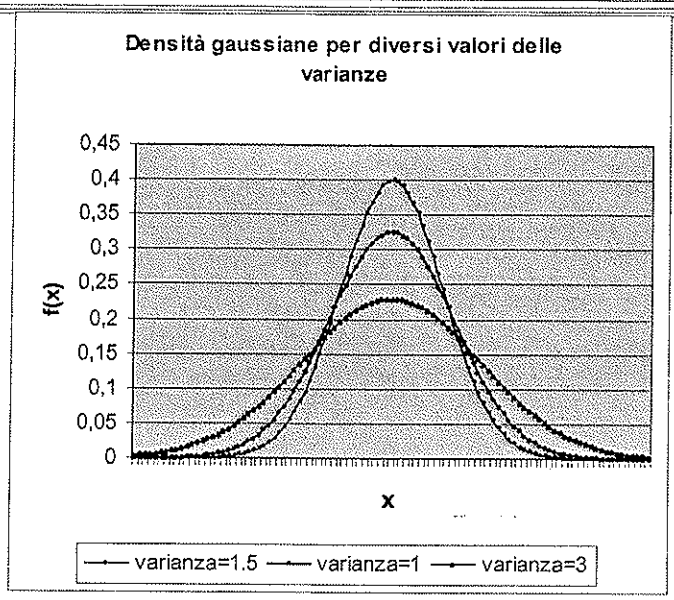
- L'espressione della funzione densità della curva normale è:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}.$$

- $E(X)=\mu$
- $\sigma(X)=\sigma$

85

Varianza



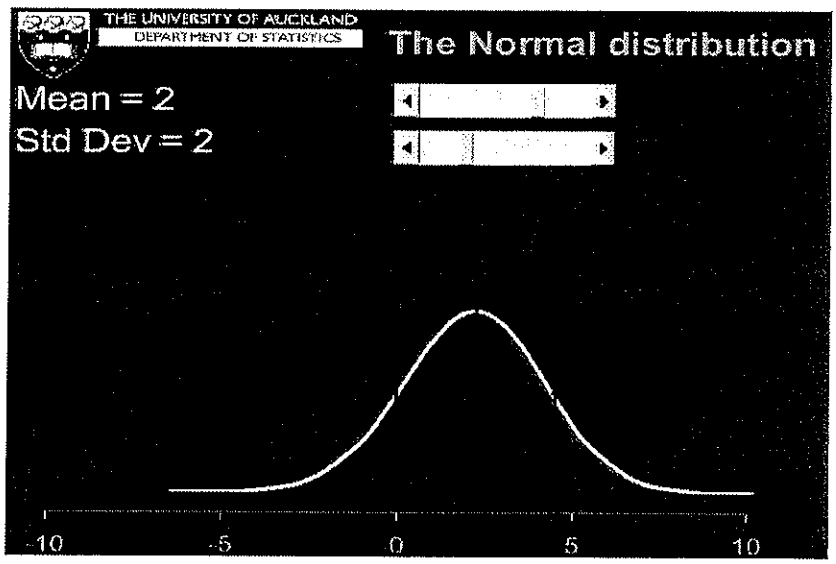
86

I flessi

- Per la distribuzione normale, le posizioni dei flessi sono simmetricamente a distanza σ dalla media e l'area sotto la curva compresa tra i due flessi corrisponde a una probabilità pari a 0,68.

87

$\mu - \sigma$, μ, $\mu + \sigma$



88

La standardizzazione

Anche per la distribuzione gaussiana vale la standardizzazione con la stessa formula e le stesse proprietà:

$$Z = \frac{X - E(X)}{\sigma(X)}$$

La densità si scrive:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

89

Proprietà della distribuzione normale

- Nel caso standardizzato, si può dimostrare che :
- (I)il 50% dei valori è inferiore a 0
- (II) il 50% dei valori è maggiore di 0
- (III) approssimativamente il 68% è tra - 1.0 e + 1.0
- (IV) approssimativamente il 95% è tra - 2.0 e + 2.0
- (V) esattamente il 95% è tra - 1.96 e + 1.96
- (VI) esattamente il 90% è tra - 1.645 e + 1.645
- (VII) esattamente il 99% è tra - 2.576 e + 2.576

90

Generalizzando

- Questi risultati consentono di applicare le seguenti affermazioni ad ogni variabile con una distribuzione normale o quasi normale.
- (a) approssimativamente il 95% di tutti i valori dovrebbe essere compreso entro due deviazioni standard della media.
- (b) praticamente tutti i valori dovrebbero essere entro 3 D.S. della media.

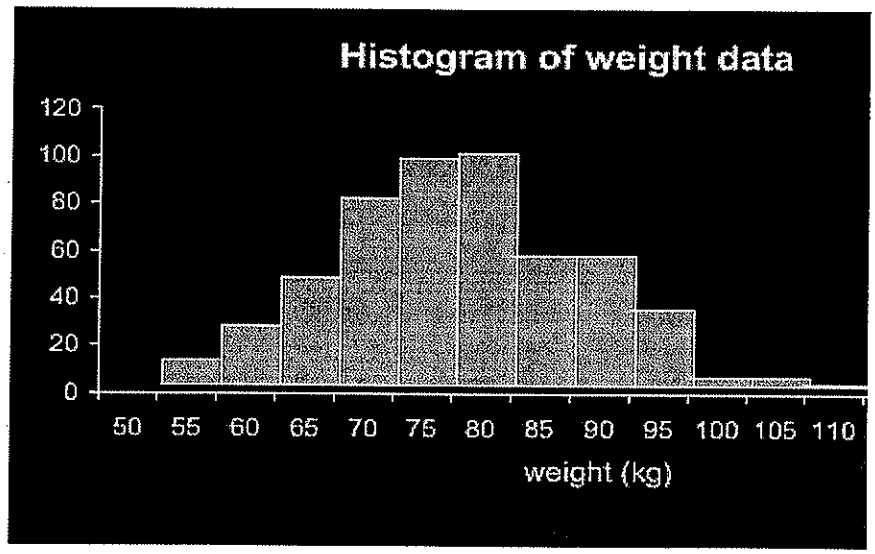
91

Esempio 1: Peso in Kg

78	66	58	58	87	74	75	79	84
93	77	76	58	73	69	65	74	69
83	64	68	85	58	73	59	66	80
83	59	83	53	70	76	65	67	79
93	70	92	62	77	64	88	60	70
64	90	82	92	84	77	71	74	92
57	93	78	78	73	69	67	69	69
89	81	79	75	91	68	77	56	71
65	67	76	63	87	81	60	74	87
83	74	70	109	89	83	67	79	59
74	58	72	87	84	74	73	81	71
86	69	57	80	73	84	54	74	70
69	73	76	83	77	67	95	51	60
75	86	68	60	77	93	92	65	93
72	68	97	64	62	60	54	89	80
66	65	84	77	93	63	76	87	85
87	67	66	66	111	126	87	88	62
62	78	80	72	62	65	64	72	78
73	83	81	75	78	93	64	74	85
53	71	69	92	88	75	76	61	74

92

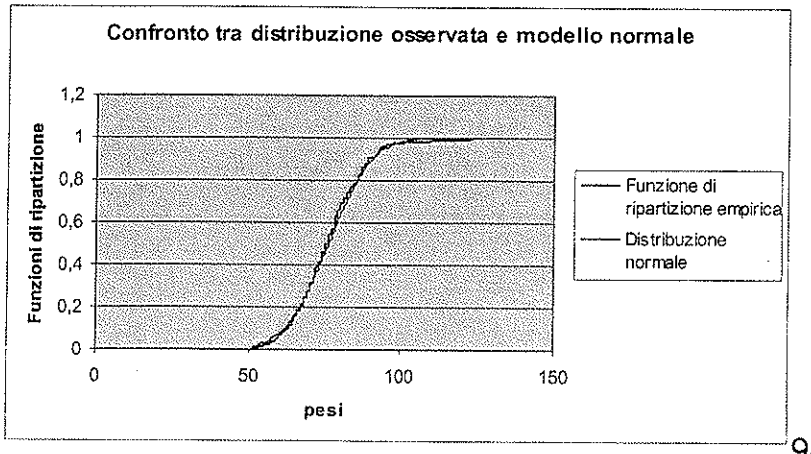
Esempio 1



93

Esempio 1

- Il modello normale con stessa media (76,17) e stessa deviazione standard (11,08) approssima bene la funzione di ripartizione empirica.



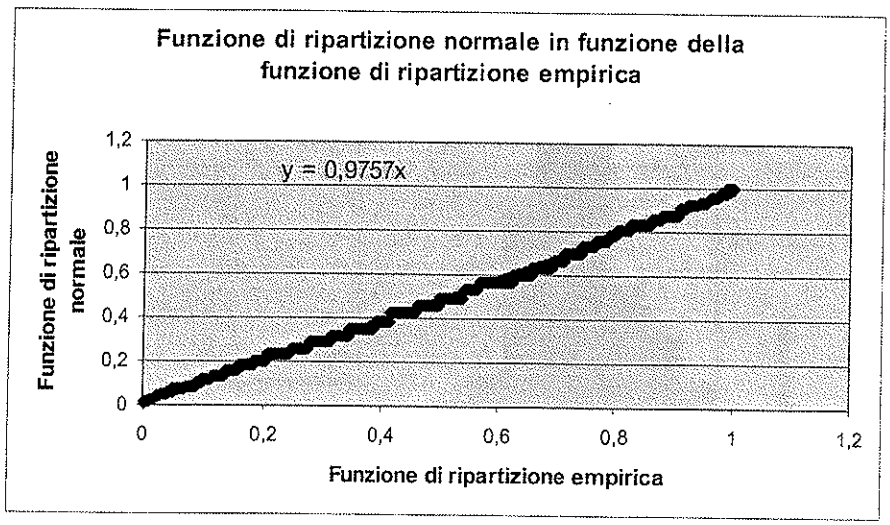
94

Confrontiamo: il P-plot

- Per meglio confrontare le due funzioni di ripartizione riportiamo i loro corrispondenti valori su un piano cartesiano:
  - ✓ Per ogni valore di x osservato, consideriamo la funzione di ripartizione empirica F*(x) e la funzione di ripartizione teorica (normale) F(x) e rappresentiamo nel piano il punto che ha per ascissa F*(x) e per ordinata F(x).
  - ✓ Se il modello approssima bene la distribuzione empirica i punti si addensano attorno alla diagonale del primo quadrante e sono bene interpolati dalla bisettrice con equazione y=x.

95

P-Plot (pesi)



96

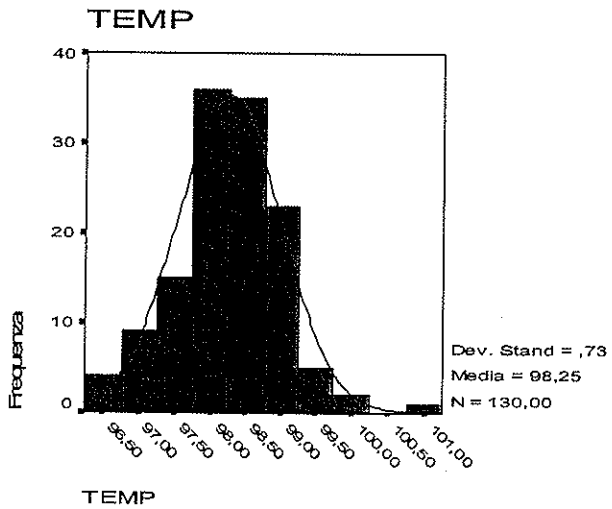
Temperatura corporea e battiti

- Numerosità campionaria: n=130
- Metà maschi e metà femmine

Temp	Sex	Battiti
96,30	1	70
96,70	1	71
96,90	1	74
97,00	1	80
97,10	1	73
97,10	1	75
97,10	1	82
99,10	2	74
99,20	2	77
99,20	2	66
99,30	2	68
99,40	2	77
99,90	2	79
100,00	2	78
100,80	2	77

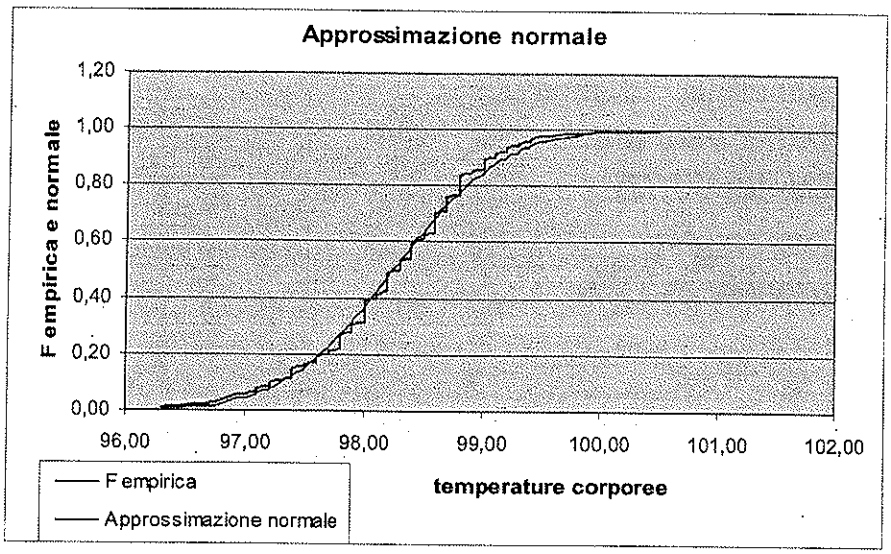
97

Approssimazione normale



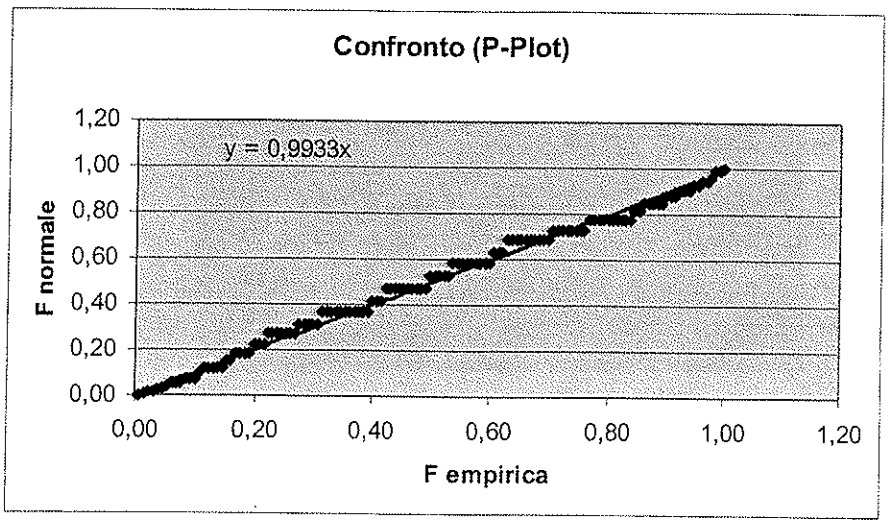
98

Confronto tra le funzioni di ripartizione



99

P-plot (temperature corporee)



100

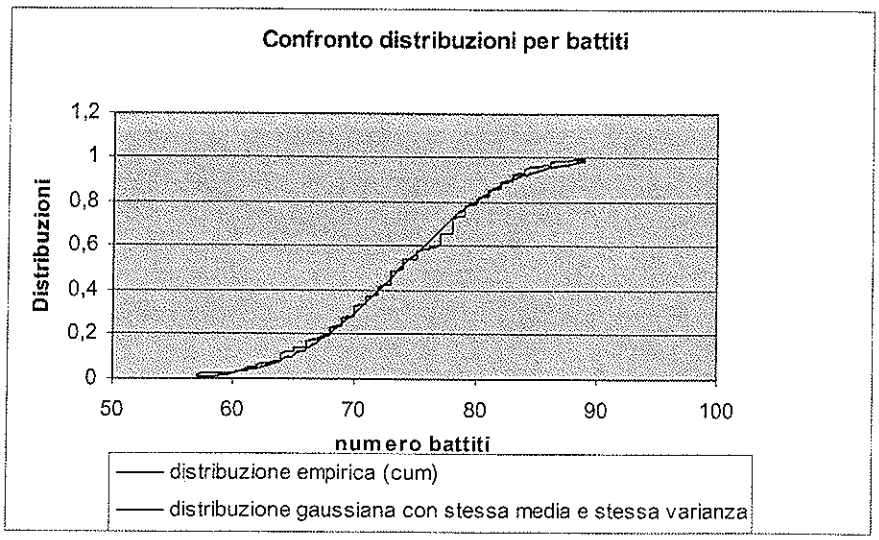
Esempio: battiti

- 130 soggetti
- Variabile discreta: battiti cardiaci.
- Media=73,76.
- Deviazione standard=7,06.

BattM	BattF	BattN
70	66	84
71	71	65
74	77	73
80	78	69
73	83	67
75	66	79
82	70	76
64	62	80
69	73	79
70	76	81
66	76	73
72	61	74
76	78	84
70	80	83
75	76	82
74	79	86
69	81	68
73	71	77
77	83	72
66	63	79
73	70	69
65	76	84
74	69	66
76	62	82
72	76	84
76	66	70
71	66	83
74	67	69
67	61	69
84	84	73
76	61	84
73	77	76
67	62	79
69	71	81
64	66	80
71	69	74
72	70	77
66	76	66
72	67	68
66	76	77
70	73	79
62	69	76
84	61	77
	73	

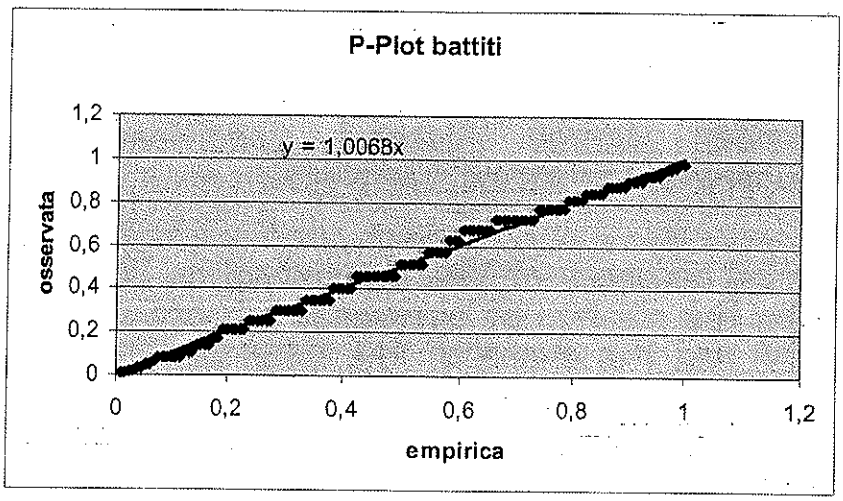
101

Confrontiamo



102

Confrontiamo



103

Importance of the normal distribution

- The normal distribution is one which appears in a variety of statistical applications. One reason for this is the central limit theorem. This theorem tells us that sums of random variables are approximately normally distributed if the number of observations is large. For example, if we toss a coin, the total number of heads approaches normality if we toss the coin a lot of times. Even when a distribution may not be exactly normal, it may still be convenient to assume that a normal distribution is a good approximation.

104

Il teorema centrale

- The central limit theorem states that the sum of a large number of independent observations from the same distribution has, under certain general conditions, an approximate normal distribution.
- The accuracy of the approximation depends on the number of summands (choose a large N) and the shape of the intial distribution. The approximation works best for symmetric distributions.
-

105

Il teorema centrale

- *L'applicazione del teorema centrale, in una delle sue formulazioni meno restrittive, permette di considerare accettabile l'approssimazione gaussiana per la distribuzione delle misure effettivamente osservate.*
- In molti casi sperimentali, le distribuzioni statistiche risultano avere un andamento ben descritto da una curva normale.

106

Il teorema centrale

- Si può senz'altro affermare che le numerose conferme sperimentali del buon adattamento della distribuzione normale a diversi fenomeni sostengono la validità del teorema centrale, d'altra parte, svariati risultati sperimentali trovano nel teorema la loro giustificazione teorica.

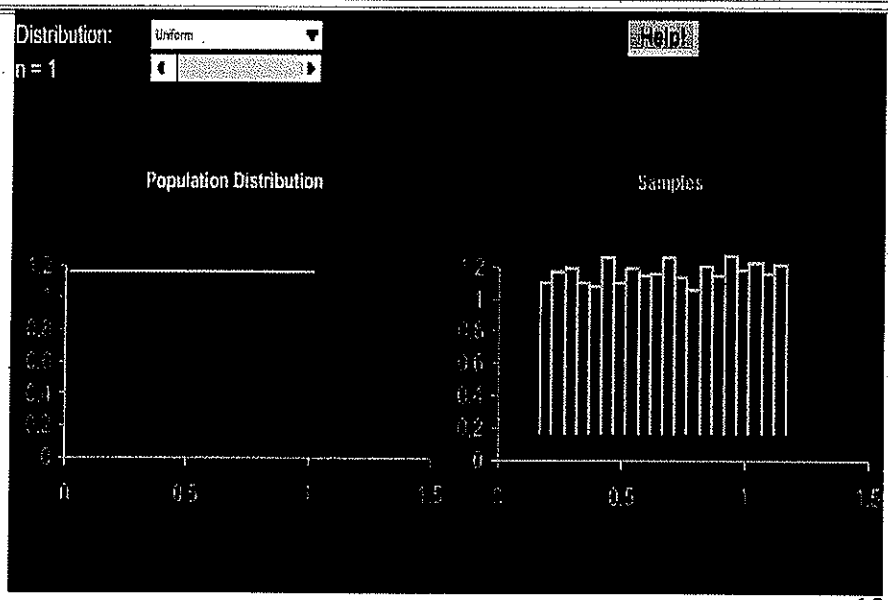
107

Il teorema centrale

- Pur partendo da una popolazione con distribuzione asimmetrica, l'andamento della distribuzione della media statistica risulta più regolare e simmetrico, fin da numerosità campionarie basse.
- La distribuzione normale si e' rivelata un modello adatto per molti esperimenti dove si misurano variabili fisiologiche (altezza, pressione del sangue, ecc.) o dove la variabilita' provviene da errori casuali negli strumenti di misura.

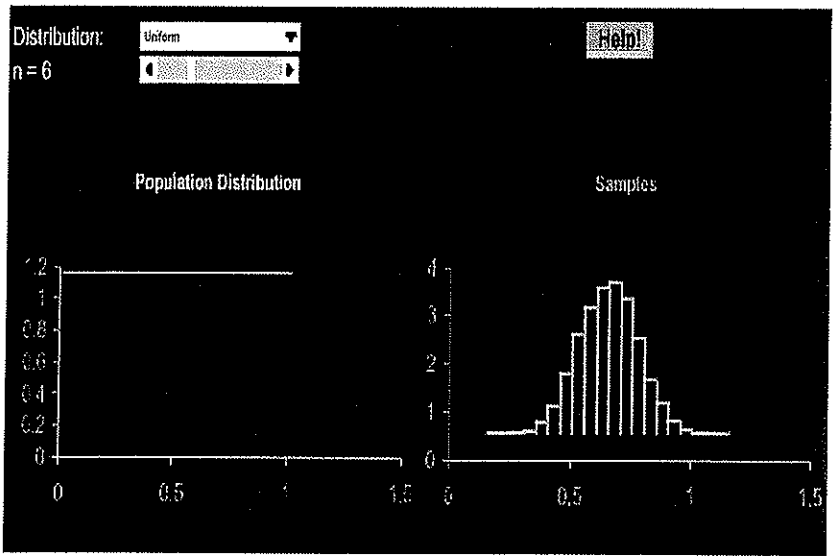
108

Esempio 1



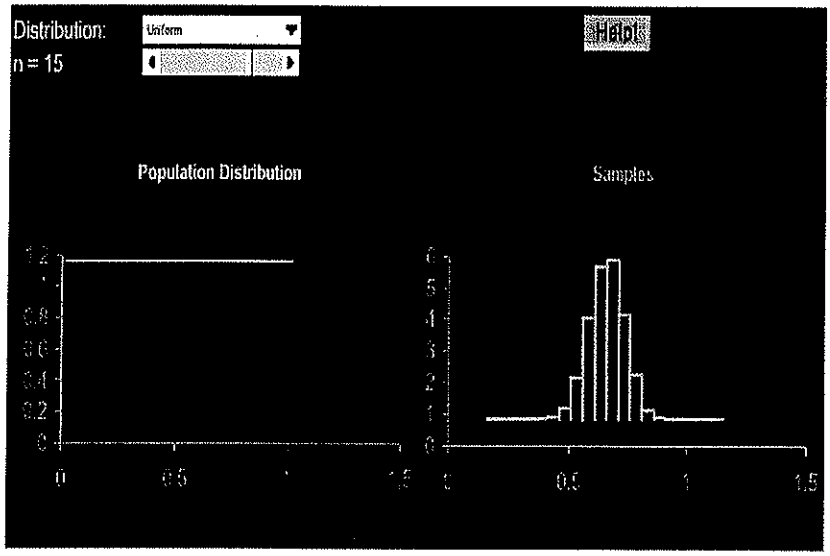
109

Esempio 1



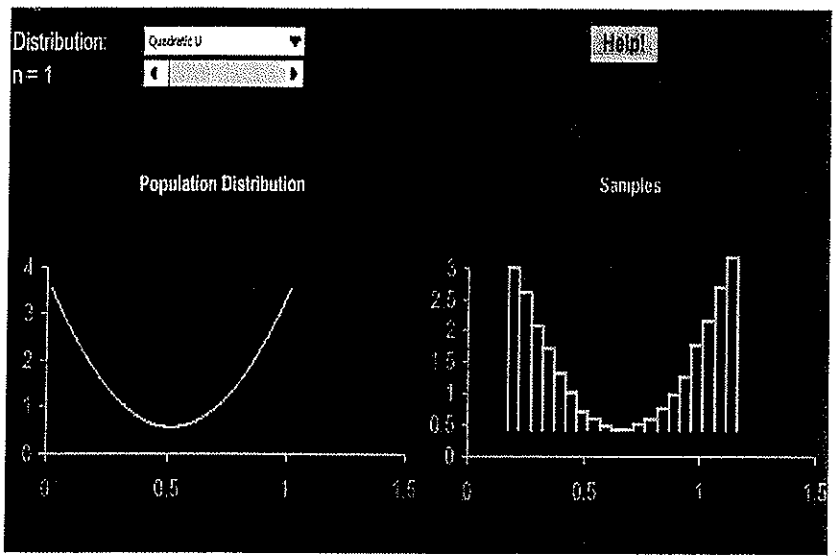
110

Esempio 1



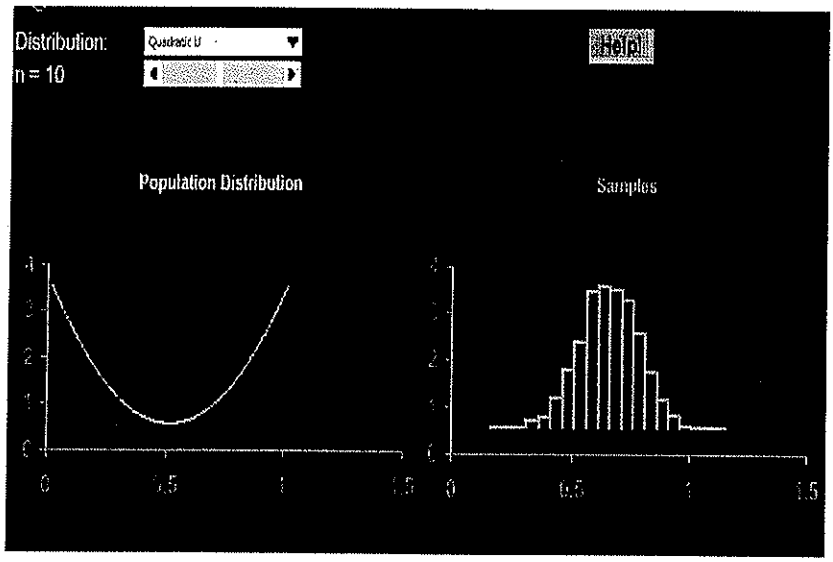
111

Esempio 2



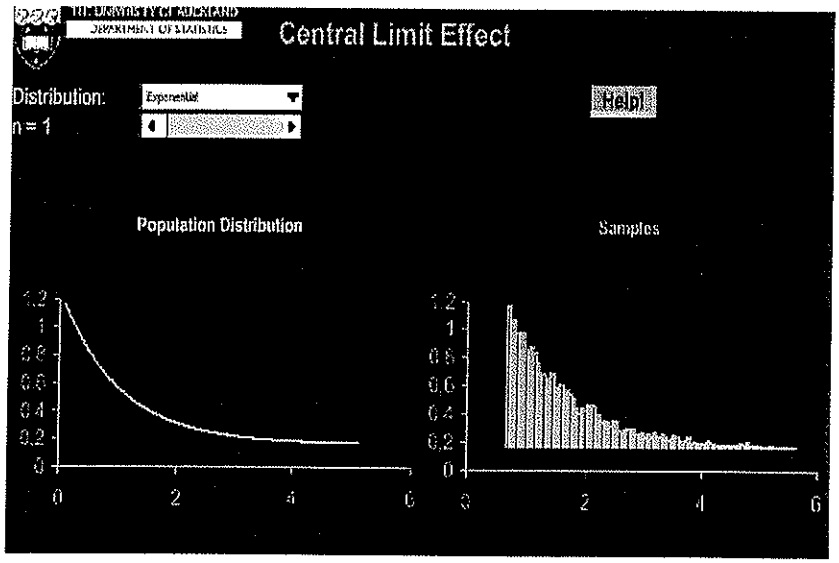
112

Esempio 2



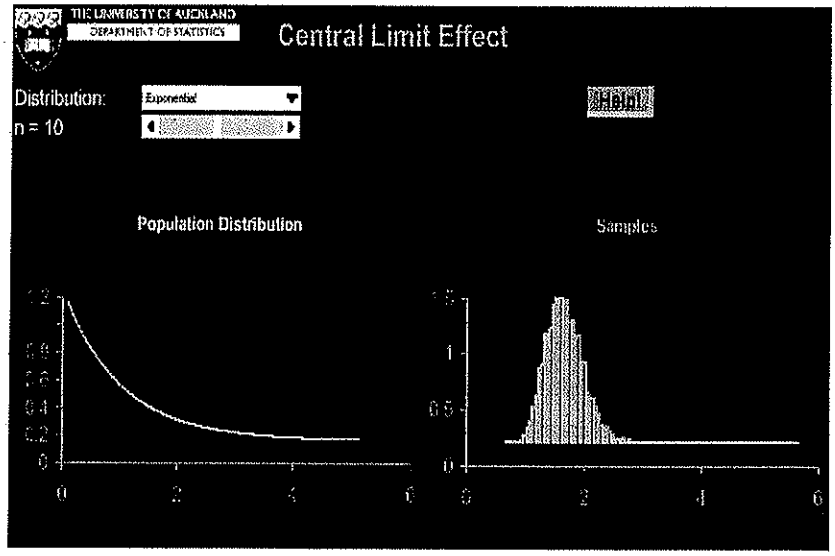
113

Esempio 3



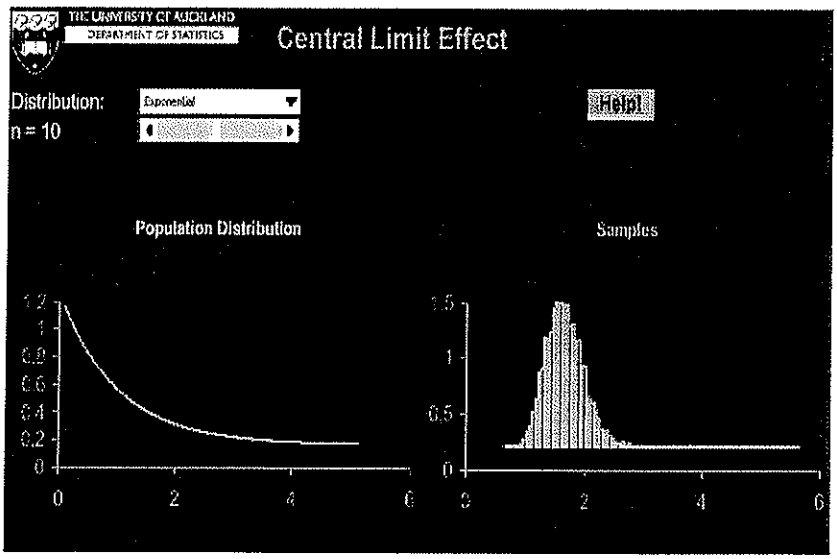
114

Esempio 3



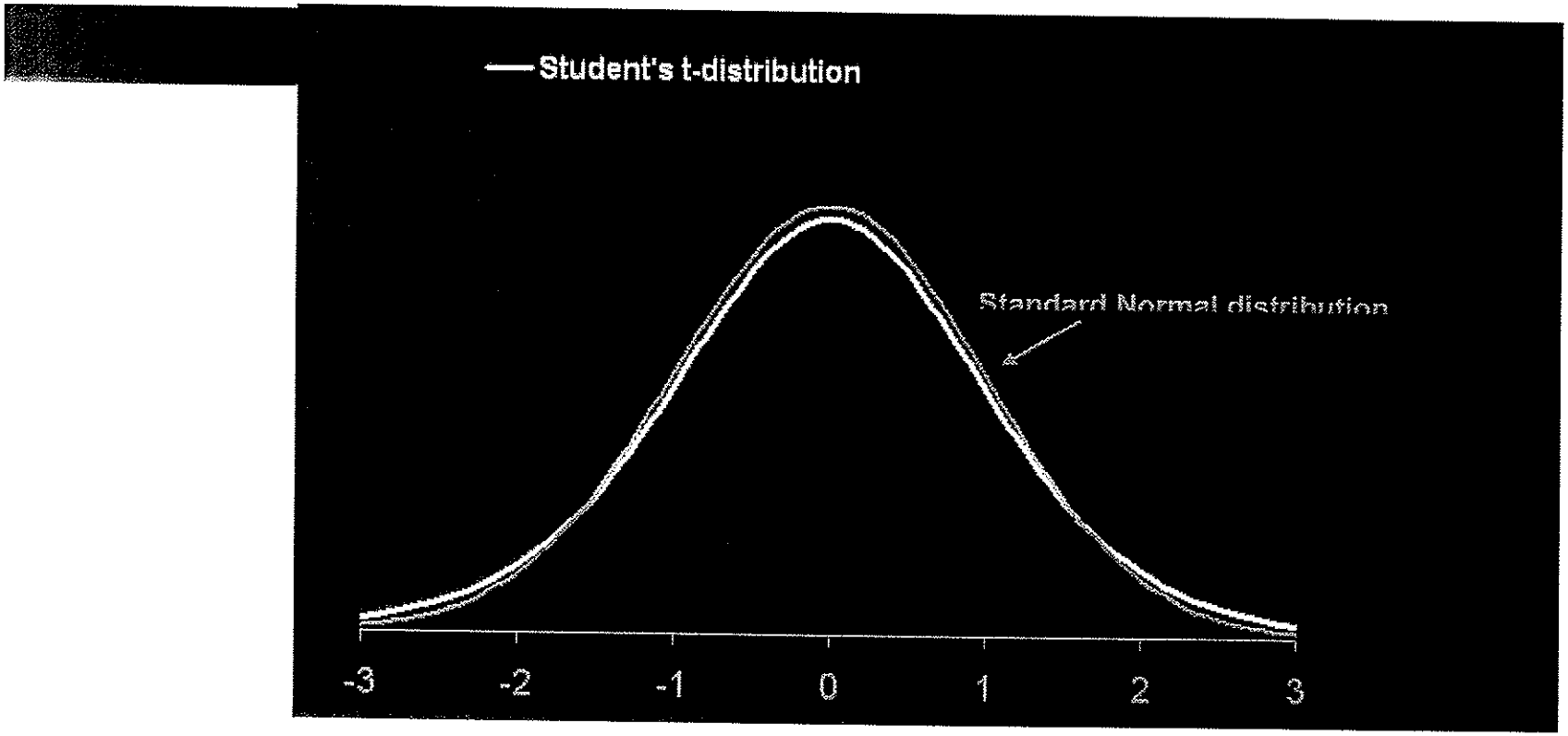
115

Esempio 3



116

The Student's t-distribution



REGRESSIONE LINEARE
CON EXCEL

Lo scopo è quello di valutare la relazione esistente tra due variabili e rappresentare tale relazione in forma grafica. La funzione più immediata che ponga in relazione due variabili non può che essere una retta, appunto **la retta di regressione.** Questo concetto considera solo caratteri **quantitativi** e vuole graficare la retta che meglio si adatta ad un insieme di punti, quelli assunti appunto dalle due variabili (**analisi lineare**). In prima istanza, calcoliamo il **coefficiente di correlazione** r $(-1 \leq r \leq 1)$ il quale misura l'intensità dell'associazione lineare tra le due variabili. Una volta constatato che esiste correlazione (tanto più r si avvicina ad 1) l'analisi prosegue con il relativo grafico di dispersione tra le due variabili, il calcolo della retta di regressione e il **coefficiente di determinazione** R^2 . Quest'ultimo ci dice se il modello ha le caratteristiche per essere adeguato e lo è tanto più questi si avvicina ad 1. Infine mediante lo studio dei **residui** è possibile affermare che il modello è adeguato o meno.

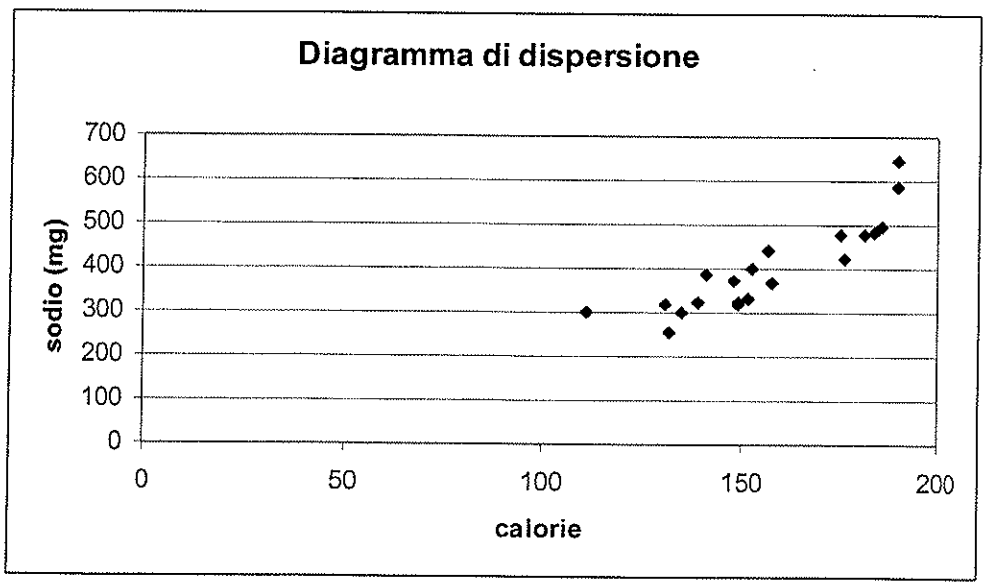
ESEMPIO

Si vuole studiare l'esistenza di una relazione lineare tra le misure di sodio e calorie presenti negli hot dogs . I dati sono riportati nel primo foglio del file **hot_dogs.xls**. e suddivisi in tre gruppi che corrispondono alla tipologie di carne che li compongono. Vengono così classificati, a seconda che la carne sia di manzo, di pollami vari o una mistura di questi con carne di maiale. I dati riferiti a queste tre tipologie vengono riportati in fogli differenti e analizzati, mediante l'analisi di regressione, in modo che le calorie (variabile esplicativa o indipendente) è sita nella prima colonna di ogni foglio elettronico mentre i valori corrispondenti del sodio (variabile di risposta o dipendente) nella seconda colonna.

La prima analisi è riferita alla tipologia: Manzo.

Calcoliamo r: opzioniamo **Inserisci → funzione → statistiche → CORRELAZIONE,** la schermata richiede le variabili, inseriamo le due colonne e dando invio compare il coefficiente che, in questo esempio, vale 0,887. Quindi c'è un evidente correlazione lineare. Passiamo, dunque, a disegnare il relativo grafico di dispersione.

Selezioniamo **Inserisci, grafico** e più in basso **Dispers.(XY),** premiamo il tasto Avanti e Excel ci chiede l'intervallo dei dati. Chiaramente bisognerà selezionare le due colonne e come per magia apparirà il grafico sottostante. Premendo Avanti sarà possibile inserire il titolo del grafico e le etichette da assegnare agli assi coordinati.

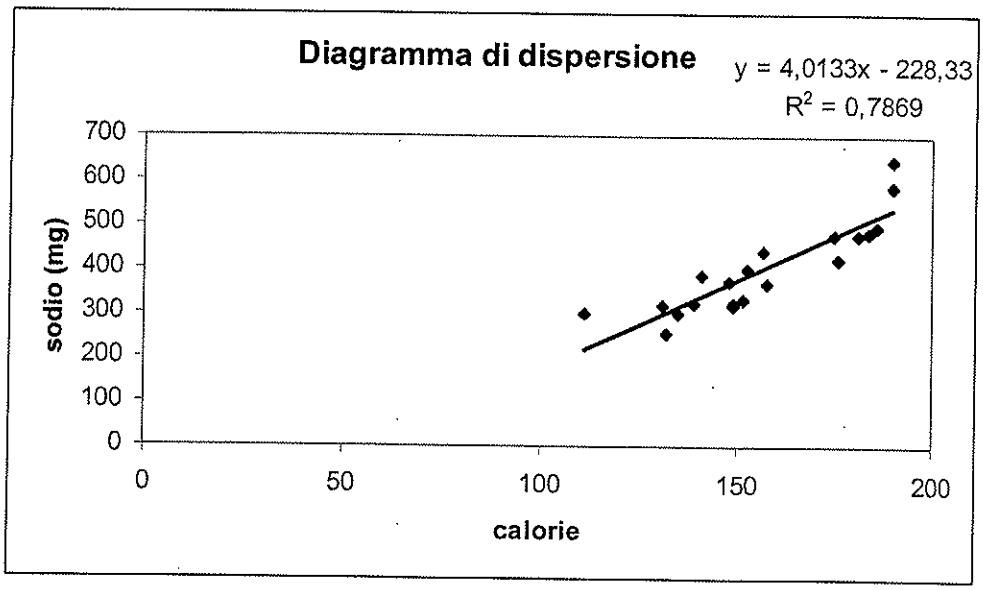


Come preannunciato, la nuvola dei punti si dispone in forma ovale la quale, anche in maniera intuitiva, presagisce ad una retta come approssimazione migliore.

Inseriamo, in maniera grafica, la retta:

selezioniamo i punti sul grafico, tasto destro del mouse, l'opzione **Aggiungi linea di tendenza**.... ed appare la retta fittante.

Si può inoltre visualizzare l'equazione della retta di regressione ed il valore del coefficiente di determinazione posizionandosi sulla retta, tasto destro, selezionare **Formato linea di tendenza** quindi **Opzioni** e successivamente i campi **Visualizza l'equazione sul grafico** e **Visualizza il valore di** R^2 **sul grafico:**



Per valutare l'adeguatezza del modello, ovvero quanto bene la regressione è riuscita a rappresentare (spiegare) i valori della variabile Y (sodio), Excel calcola il **coefficiente di determinazione** R^2. Per definizione è il quadrato del coefficiente di correlazione r e se $R^2 = 1$ tutta la variabilità di Y è spiegata dalla relazione lineare con X.

Allora si dirà che la retta di regressione ha i presupposti per essere un modello adeguato se R^2 tende a 1. In questo caso vale 0,787, il grado di adeguatezza può considerarsi buono.

Un ulteriore criterio per valutare la bontà del modello di regressione è l'**analisi dei residui.** Il residuo è definito come la differenza tra il valore osservato Y e il valore atteso $\hat{Y}$ (quello che deriva dalla stima mediante retta)

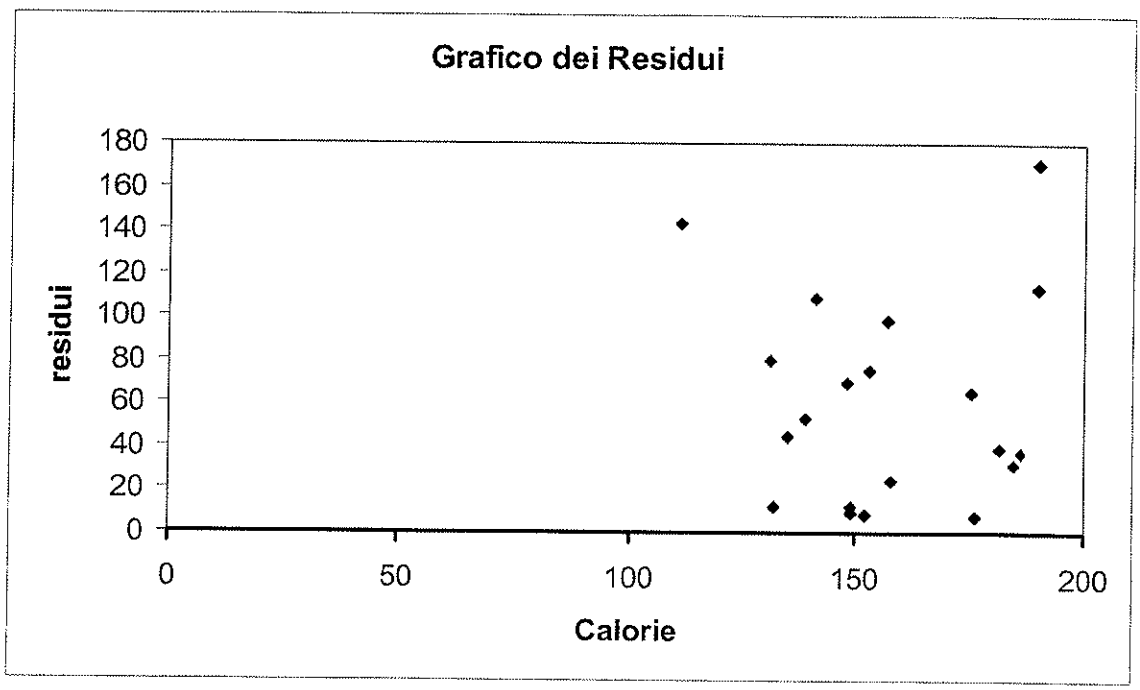
per ogni valore osservato di X. La capacità di adattamento ai dati della retta di regressione può essere valutata rappresentando in un grafico i residui (in ordinata) con i rispettivi valori della X. Il modello sarà appropriato se nel grafico dei residui **non** si riscontrerà alcun andamento particolare, ovvero i punti del grafico si disporranno in maniera del tutto casuale privi di una qualsiasi struttura funzionale.

Vediamo come si effettua l'analisi dei residui in forma grafica:

ogni punto del grafico di dispersione dei dati originali è trasferito in un altro grafico, il **grafico dei residui**, nel modo seguente: il valore di x rimane invariato, mentre sull'asse verticale, invece di y, viene riportato il residuo corrispondente. Per un buon adattamento vogliamo che essi siano distribuiti casualmente attorno alla retta orizzontale $y = 0$ ($\Leftrightarrow Y_{oss}=\hat{Y}$).

Il diagramma dei residui si ottiene cliccando su:

Strumenti → Analisi dati → Regression(X,Y) e selezionando **Residual plot**, il diagramma dei residui di calorie sodio è riportato nella seguente figura:



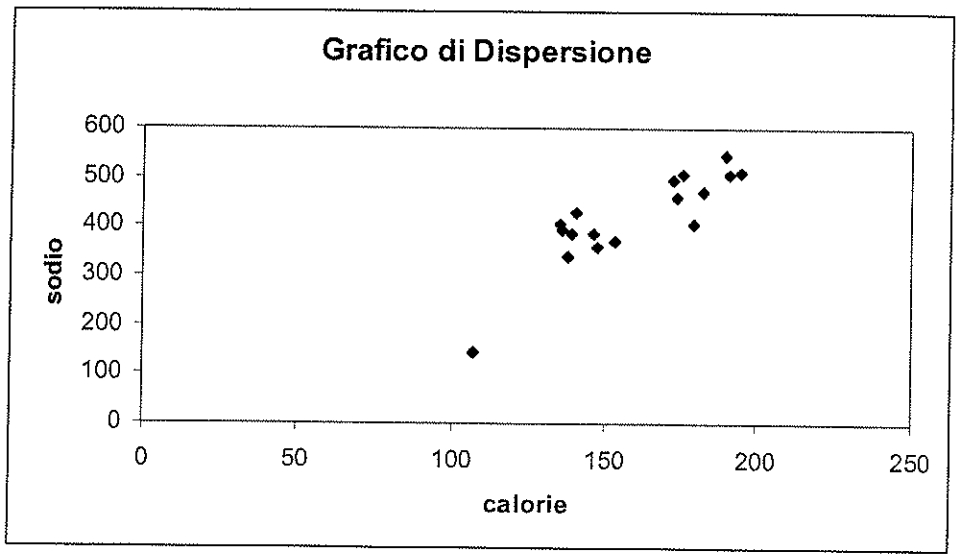
Come si può notare i residui appaiono distribuiti piuttosto casualmente attorno alla retta orizzontale.

Quindi possiamo concludere che il modello di regressione lineare, per il nostro esempio, è un modello adeguato.

In Excel lo strumento **Analisi dei dati** non è sempre disponibile (non è presente in tutte le versioni di WINDOWS), quindi per effettuare l'analisi dei residui sarà necessario costruirsi la colonna dei residui inserendo la formula dei residui.

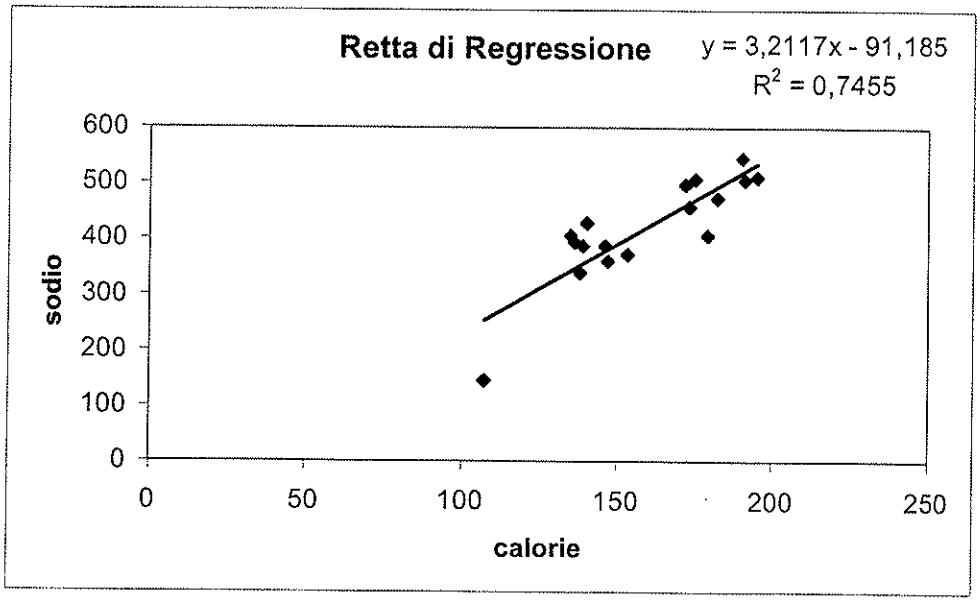
Le stesse analisi possono essere riproposte per la seconda tipologia di carne degli hot dogs che, come abbiamo detto, è data da una mistura di manzo pollame e maiale.

Il coefficiente di correlazione vale r =0,863 e il relativo grafico di dispersione:

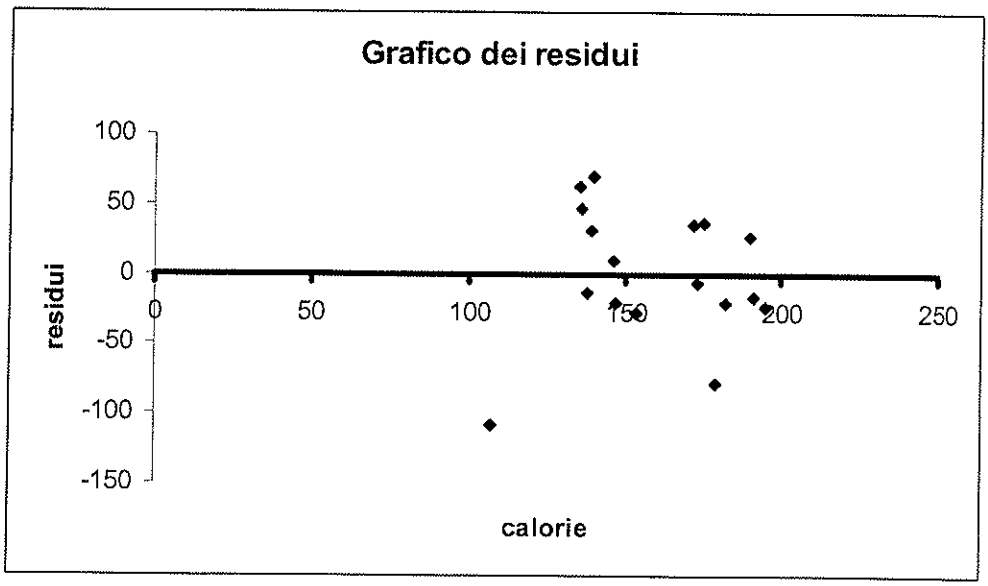


La nuvola dei punti, in tale grafico, fa intuire che una retta possa fittare i dati in maniera adeguata.

Inseriamo la retta di regressione che ha equazione y=3,217 x - 91,185 e coefficiente di determinazione R^2=0,7455:



Anche in questo caso il coefficiente R^2 è molto prossimo ad 1 il che ci suggerisce che il modello è adeguato. Vediamo il grafico dei residui:

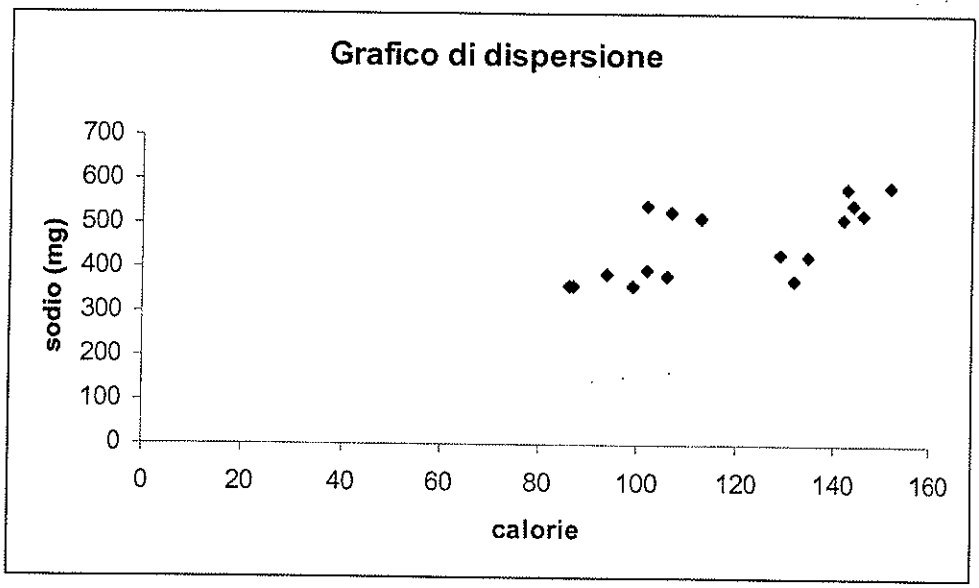


I punti si distribuiscono in maniera casuale intorno a y=0.

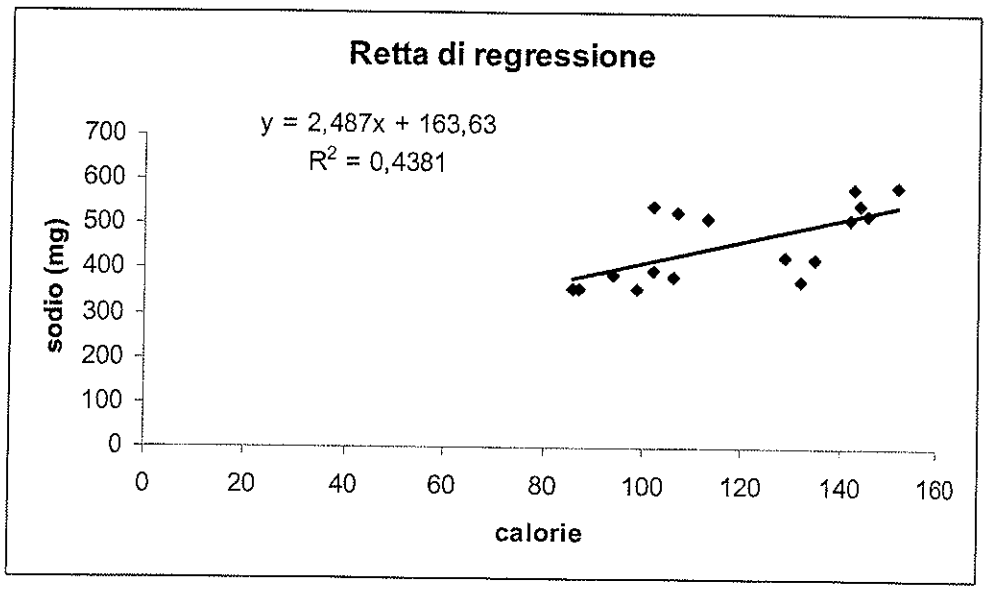
Il modello di regressione lineare è adeguato.

Infine vediamo l'analisi di regressione per la terza tipologia: pollame.

Il coefficiente di correlazione r=0,662 e il grafico di dispersione:

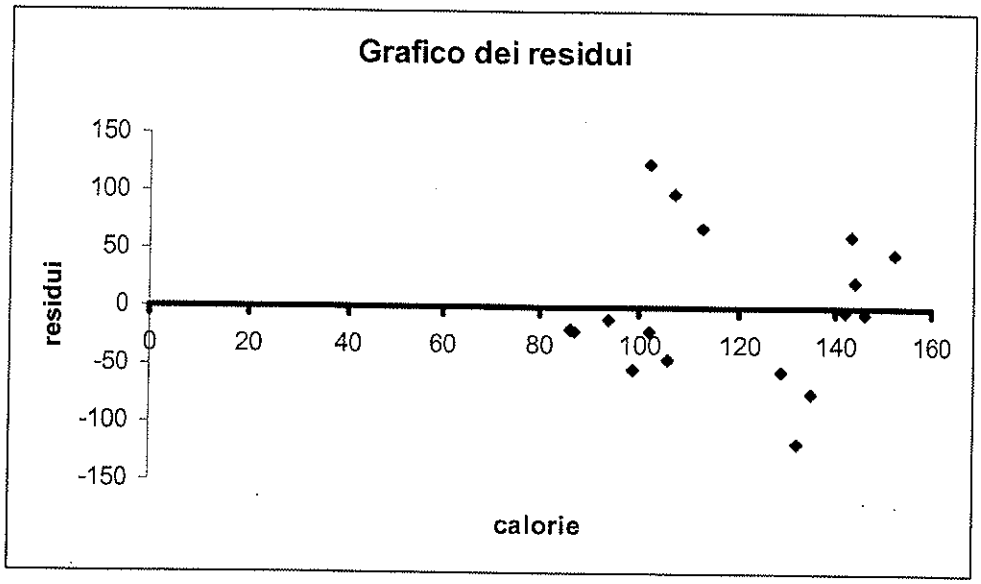


La nuvola dei punti è disposta in modo che una retta difficilmente fitterà bene i dati ed infatti:



Il coefficiente di determinazione R^2 non è prossimo ad 1.

Tracciamo il relativo grafico dei residui:



Osserviamo con più accuratezza tale grafico, spiccano tra gli altri due punti di coordinate P_1=(102;124,696) e P_2=(132;-116,914) caratterizzati per aver il massimo residuo (distanza dalla retta y=0). Tali punti, in gergo statistico, si dicono **outlier** e in potrebbero corrispondere ad errori avvenuti nella rilevazione campionaria.

Questi corrispondono (nel foglio relativo alla dispersione) rispettivamente ai records individuati da una quantità di sodio pari 542 e 375, vediamo come gli outlier influenzano la retta di regressione e le sue statistiche.

Si ripropone tutta l'analisi eliminando gli outlier dalle osservazioni e si osserva che R^2 aumenta fino a 0.67 e quindi si ottiene un notevole miglioramento tale che si può affermare che il modello ad eccezione degli outlier è adeguato.

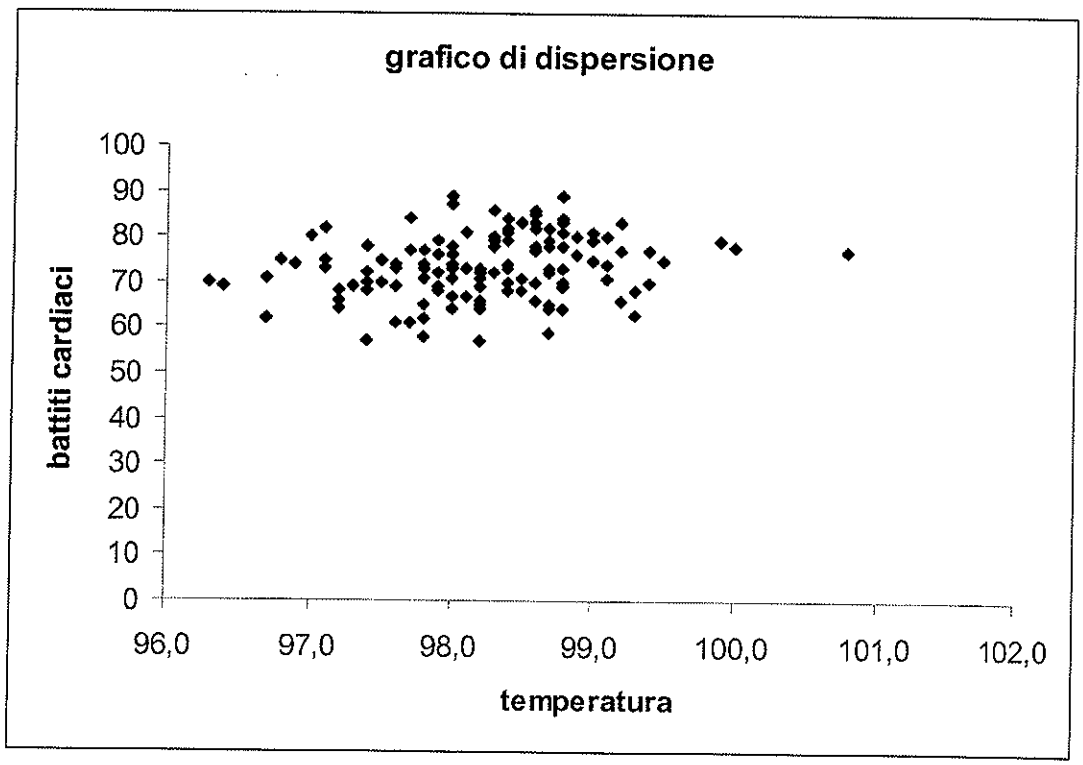
CONTROESEMPIO

Vediamo come sia possibile effettuare ugualmente un'analisi di regressione anche quando questa non ha senso.

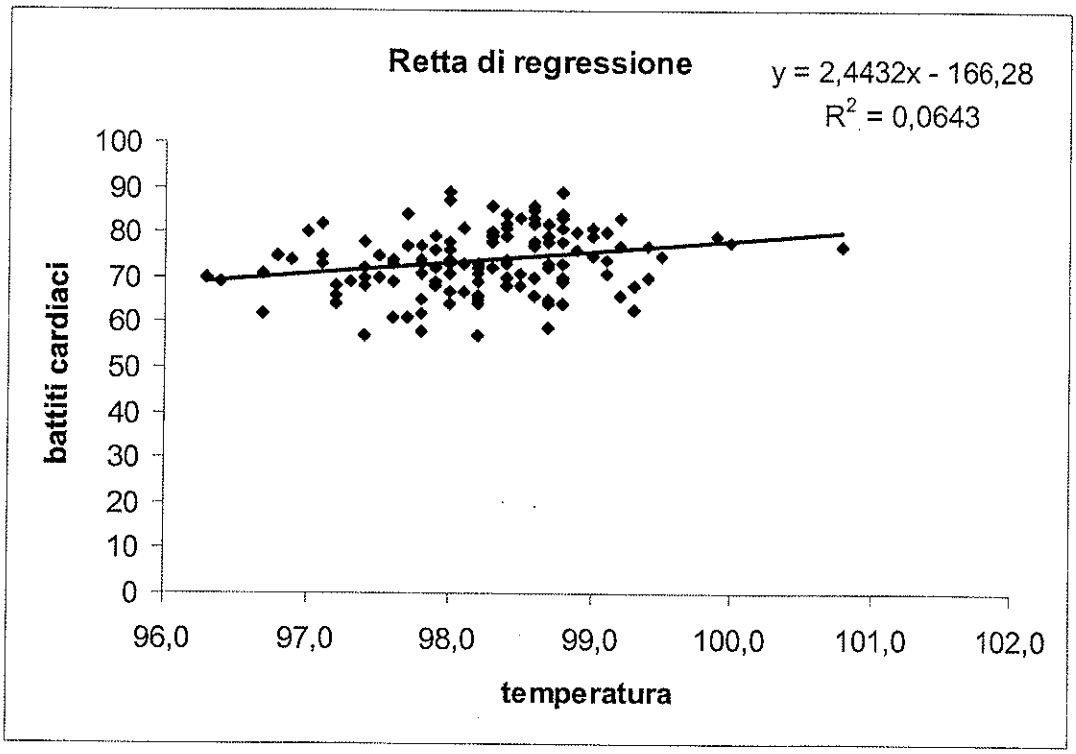
Apriamo il file **Temperature.xls**, in tale file sono presenti 130 osservazioni su uomini e donne relative a temperature corporee (espresse in °F) e il corrispettivo numero di battiti cardiaci al minuto.

Si vuole sapere se esiste una relazione lineare tra questi due set di osservazioni.

Vediamo il grafico di dispersione:



Come si può ben vedere la nuvola dei punti non ha una forma ovale ben definita e quindi una retta non li interpolerà in modo proficuo. Calcoliamo ugualmente la retta e il coefficiente di determinazione:

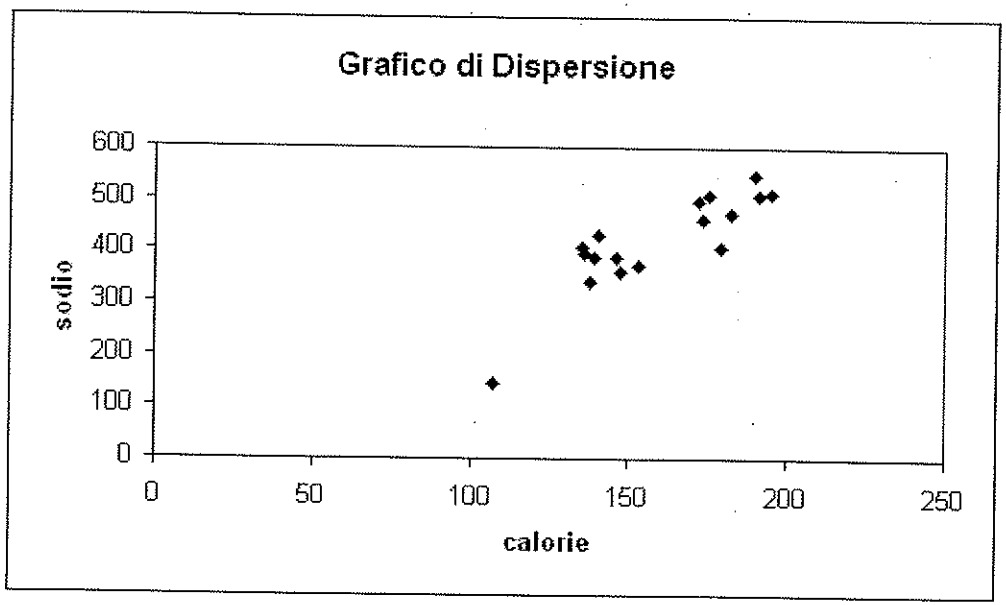


Come si può notare il coefficiente di determinazione è ben lungi da essere vicino a 1; quindi possiamo affermare che l'analisi di regressione non è un modello adeguato per questo set di dati. ($r = 0{,}25$ non c'è correlazione!!)

Calorie	Sodio (mg)
173	458
191	506
182	473
190	545
172	496
147	360
146	387
139	386
175	507
136	393
179	405
153	372
107	144
195	511
135	405
140	428
138	339

r

0,863403



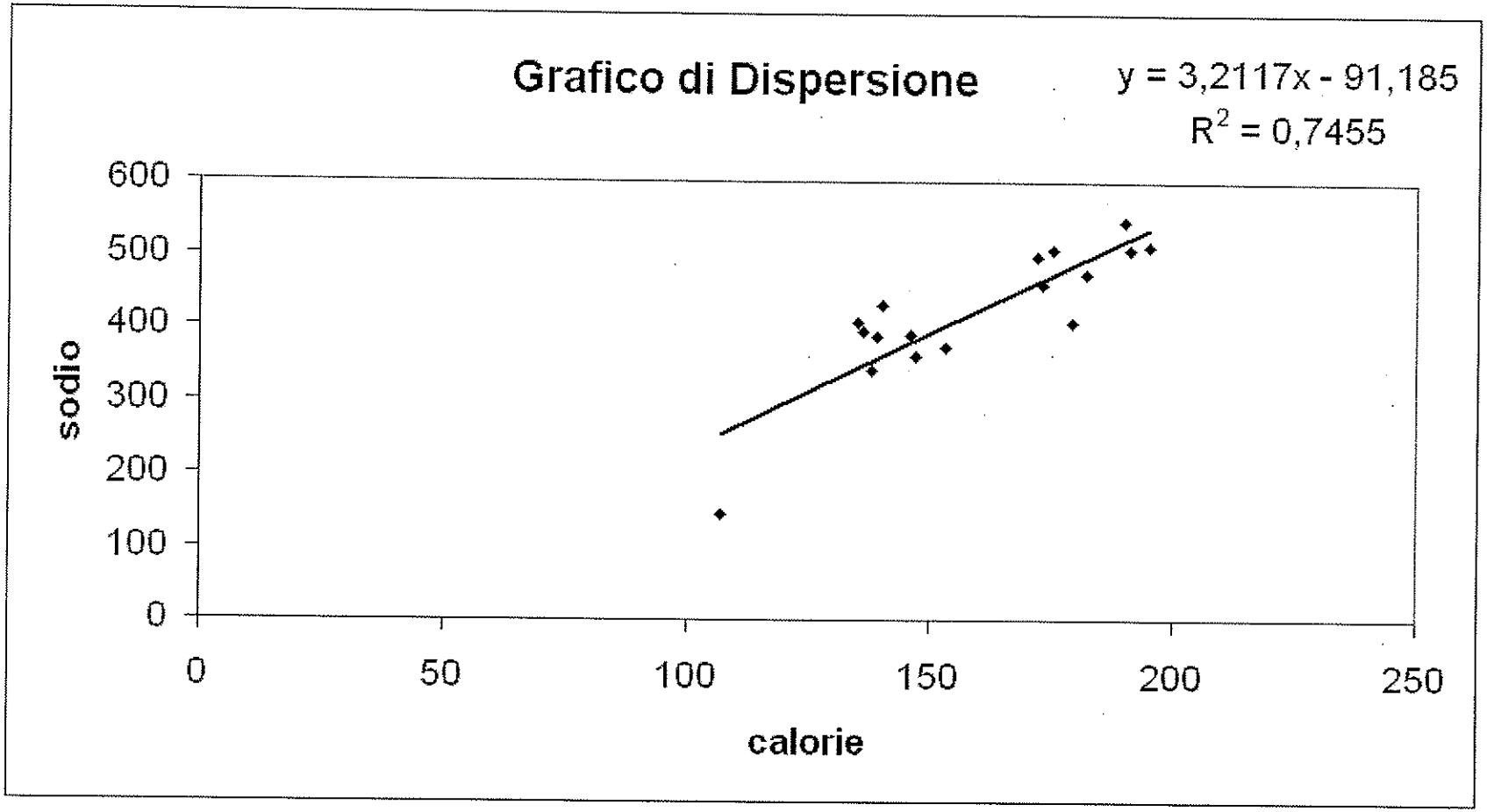
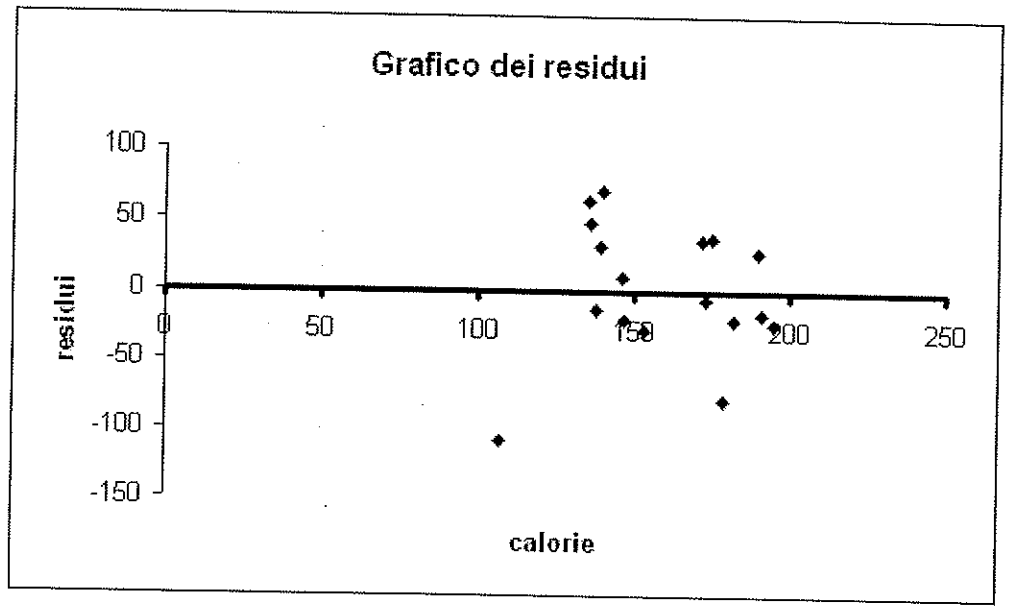
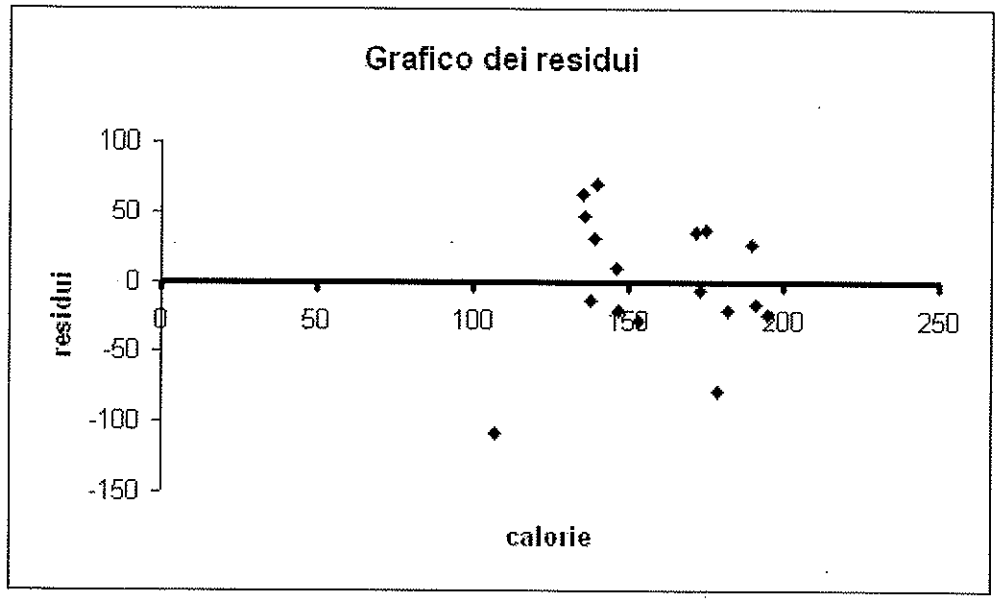


Grafico di Dispersione
y = 3,2117x - 91,185
R2 = 0,7455
sodio
600
500
400
300
200
100
0
0
50
100
150
200
250
calorie

Calorie	Sodio (mg)	m	b	Residui
173	458	3,211	-91,185	-6,318
191	506			-16,116
182	473			-20,217
190	545			26,095
172	496			34,893
147	360			-20,832
146	387			9,379
139	386			30,856
175	507			36,26
136	393			47,489
179	405			-78,584
153	372			-28,098
107	144			-108,392
195	511			-23,96
135	405			62,7
140	428			69,645
138	339			-12,933



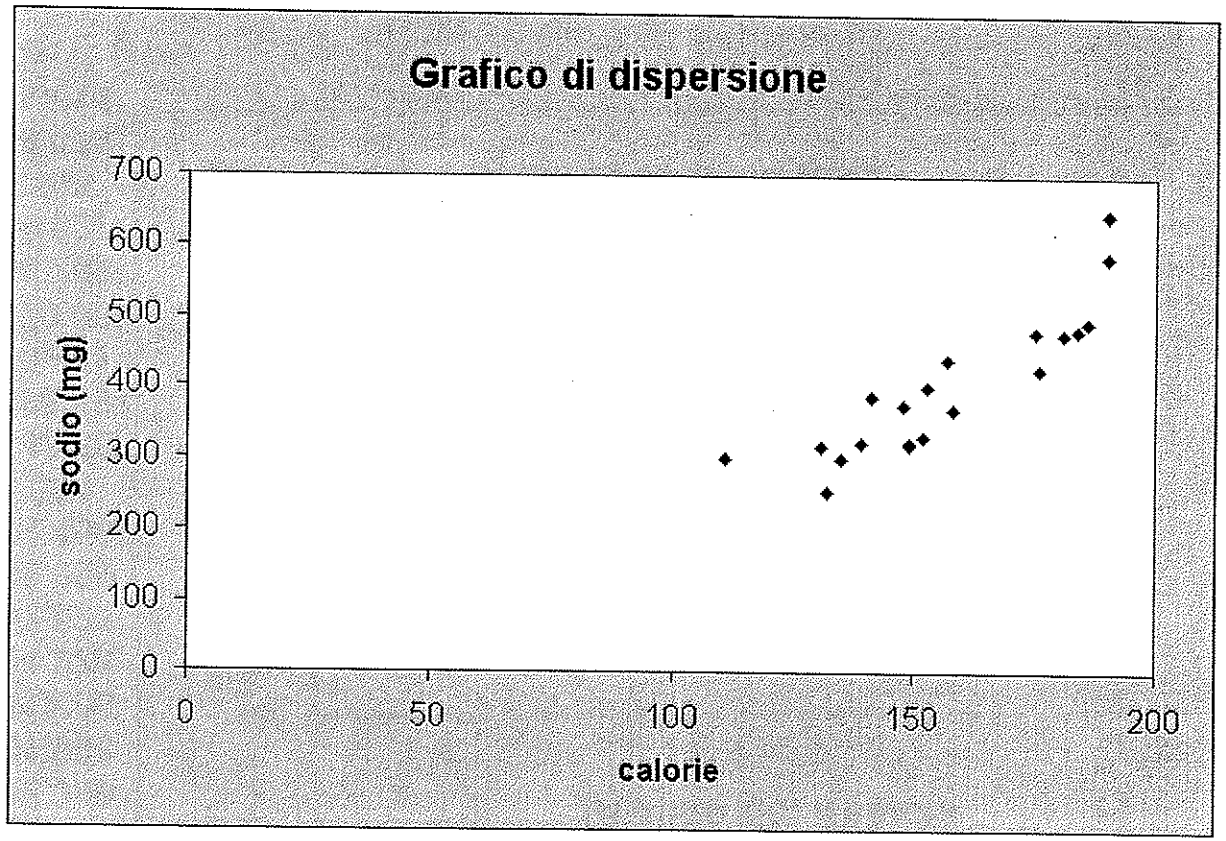
Calorie	Sodio (mg)	m	b	Residui
173	458	3,211	91,185	-6,318
191	506			-16,116
182	473			-20,217
190	545			26,095
172	496			34,893
147	360			-20,832
146	387			9,379
139	386			30,856
175	507			36,26
136	393			47,489
179	405			-78,584
153	372			-28,098
107	144			-108,392
195	511			-23,96
135	405			62,7
140	428			69,645
138	339			-12,933



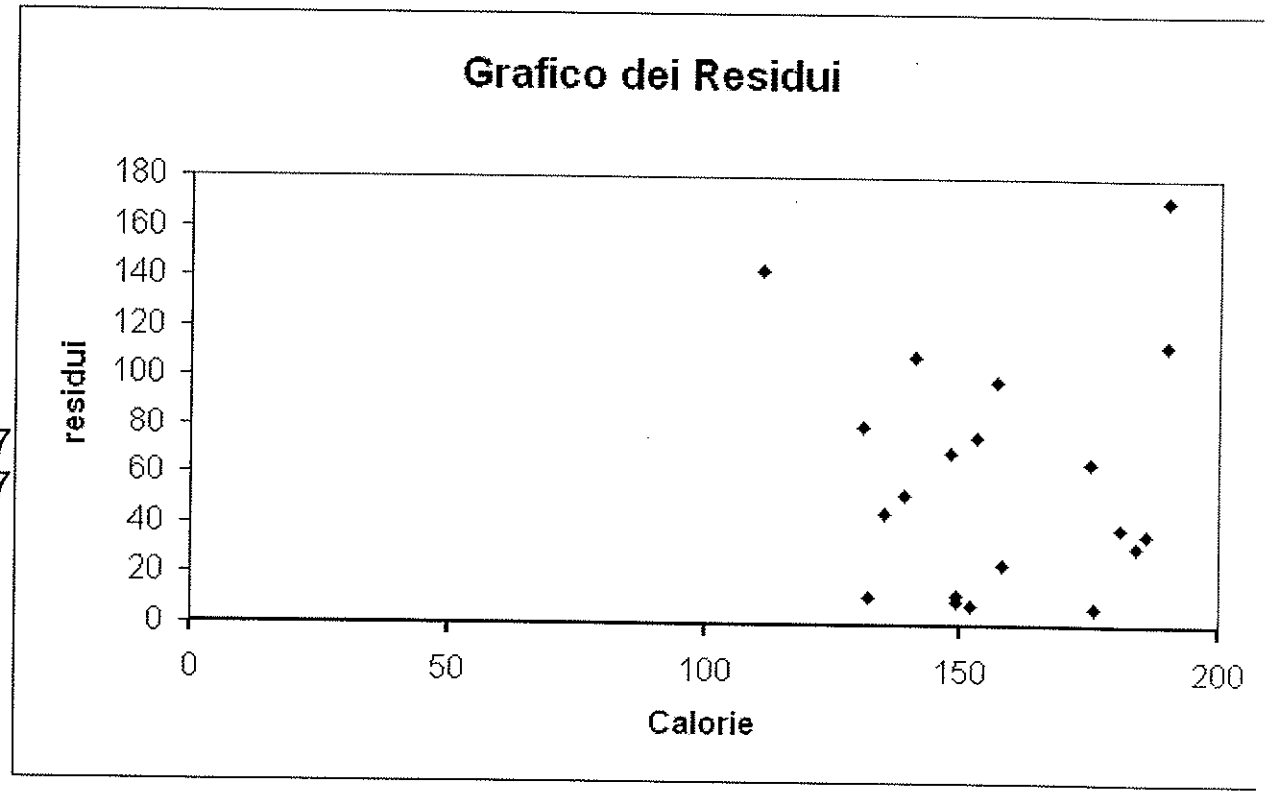
Calorie	Sodio(mg)
186	495
181	477
176	425
149	322
184	482
190	587
158	370
139	322
175	479
148	375
152	330
111	300
141	386
153	401
190	645
157	440
131	317
149	319
135	298
132	253

r

0,887087



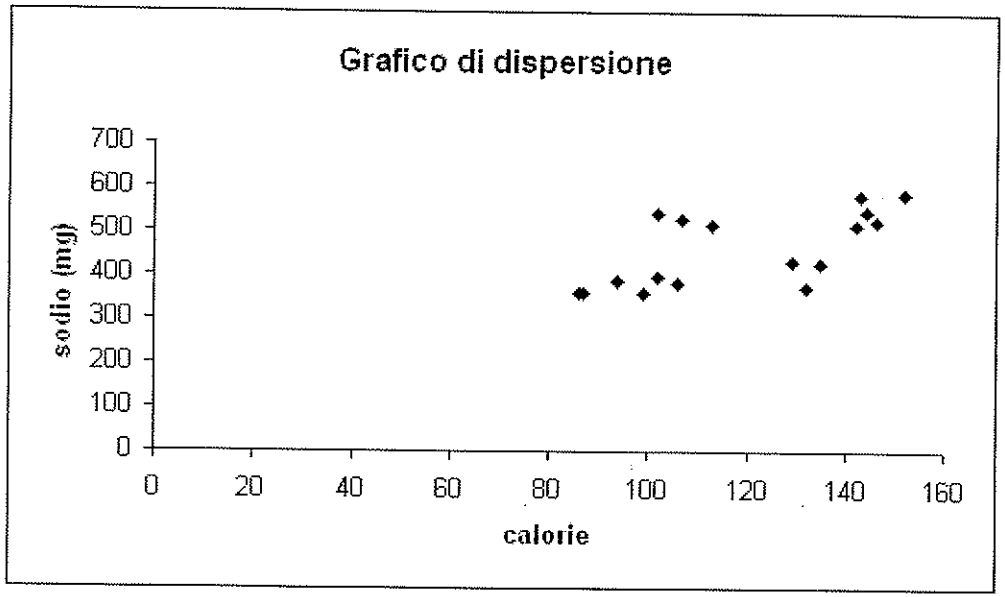
Calorie	Sodio(mg)	m	b	Residui
186	495	4,013	-288,33	36,912
181	477			38,977
176	425			7,042
149	322			12,393
184	482			31,938
190	587			112,86
158	370			24,276
139	322			52,523
175	479			65,055
148	375			69,406
152	330			8,354
111	300			142,887
141	386			108,497
153	401			75,341
190	645			170,86
157	440			98,289
131	317			79,627
149	319			9,393
135	298			44,575
132	253			11,614

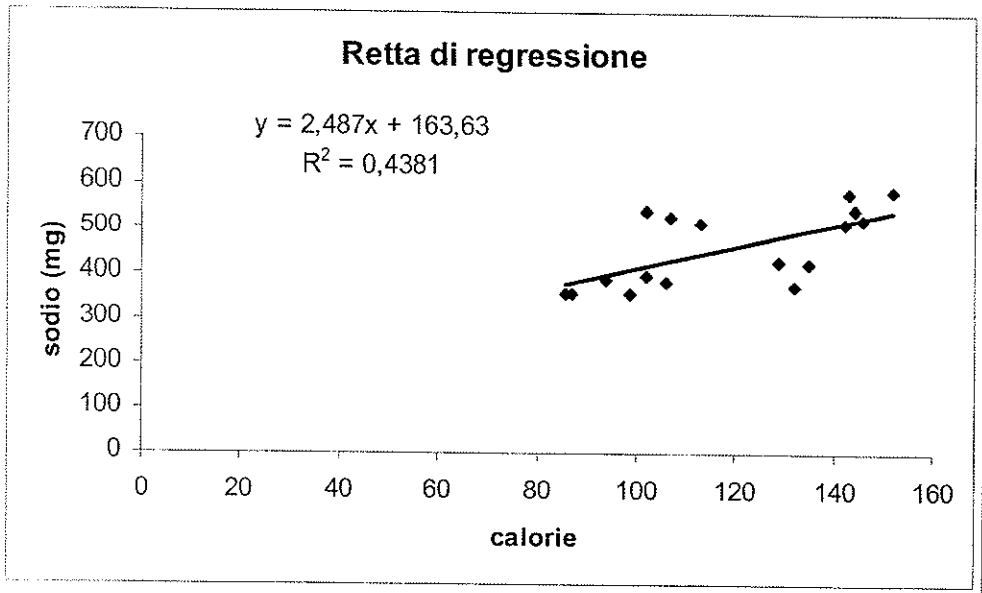


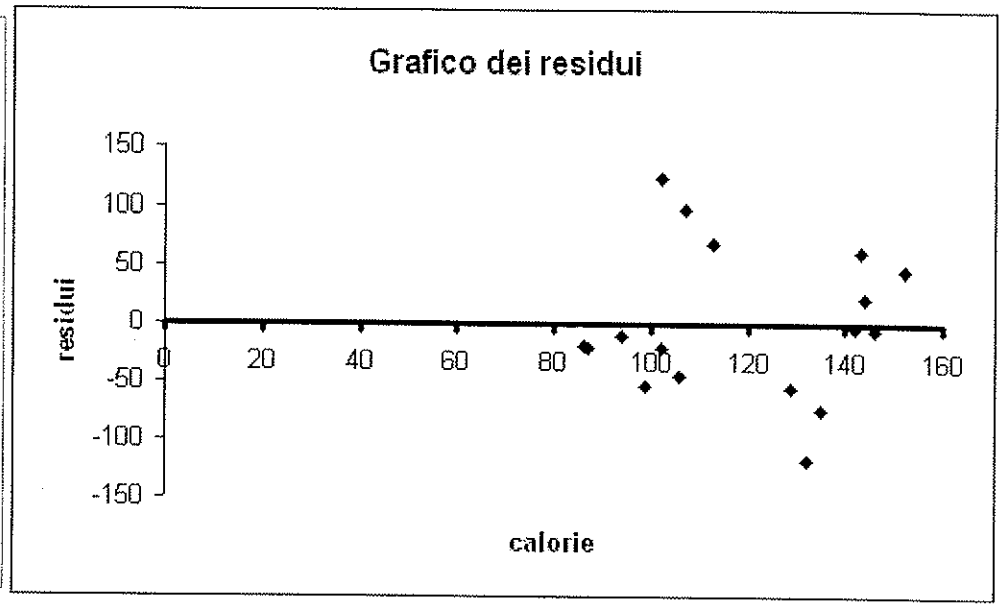
Calorie	Sodio (mg)
129	430
132	375
102	396
106	383
94	387
102	542
87	359
99	357
107	528
113	513
135	426
142	513
86	358
143	581
152	588
146	522
144	545

r

0,661862

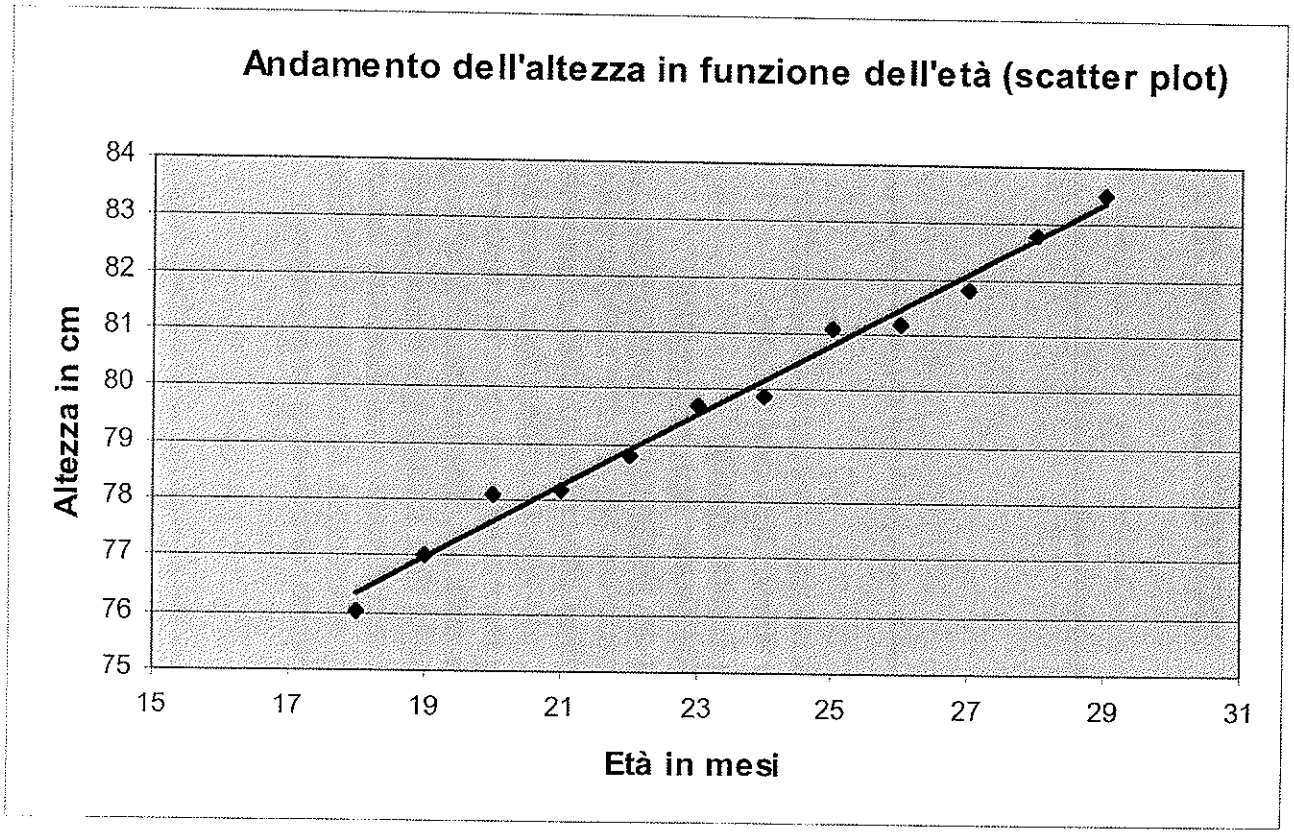






Calorie	Sodio (mg)	m	b	Residui
129	430	2,487	163,63	-54,453
132	375			-116,914
102	396			-21,304
106	383			-44,252
94	387			-10,408
102	542			124,696
87	359			-20,999
99	357			-52,843
107	528			98,261
113	513			68,339
135	426			-73,375
142	513			-3,784

86	358	-19,512
143	581	61,729
152	588	46,346
146	522	-4,732
144	545	23,242



ANALISI STATISTICHE BIVARIATE

Relazioni statistiche tra variabili quantitative: la correlazione e la regressione

1

Story Name: Age and height
Story Topics: Health
Datafile Name: Age and height
Methods: Scatterplot , Regression , Correlation
Abstract: The height of a child is not stable but increases over time. Since the pattern of growth varies from child to child, one way to understand the general growth pattern is by using the average of several children's heights, as presented in this data set. The scatterplot of height versus age is almost a straight line, showing a linear growth pattern.
The straightforward relationship between height and age provides a simple illustration of linear relationships, correlation, and simple regression.

2

Datafile Name: Age and height
Datafile Subjects: Health
Story Names: Age and height
Description: Mean heights of a group of children in Kalama, an Egyptian village that is the site of a study of nutrition in developing countries. The data were obtained by measuring the heights of all 161 children in the village each month over several years.
Number of cases: 12
Variable Names:

1. Age: Age in months
2. Height: Mean height in centimeters for children at this age

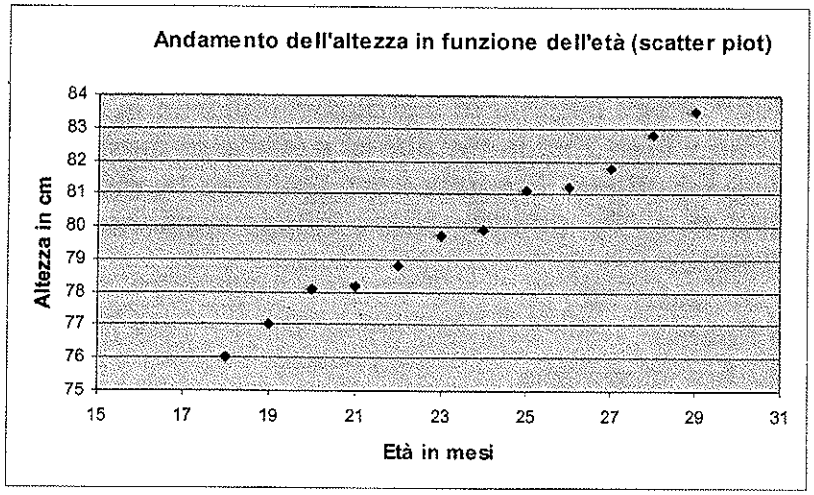
3

I dati

Age	Height
18	76,01
19	77,00
20	78,10
21	78,20
22	78,80
23	79,70
24	79,90
25	81,10
26	81,20
27	81,80
28	82,80
29	83,50

4

Lo scatter plot



5

Associazione tra variabili quantitative

- Consideriamo due variabili quantitative X e Y di cui interessa valutare l'associazione. Più precisamente la ***correlazione.***

Unità	Carattere X	Carattere Y
1	x_1	y_1
2	x_2	y_2
...	...	...
n	x_n	y_n

6

Visualizzazione: scatter plot

Per valutare *qualitativamente* l'esistenza di *associazione (correlazione)* tra due caratteri quantitativi, si procede ad una prima analisi grafica attraverso la costruzione di un **diagramma di dispersione o scatter plot.**

Si riportano in ascissa le misure osservate $x_1, x_2,\ldots,x_n$ del **carattere X**, in ordinata le corrispondenti misure osservate $y_1, y_2,\ldots,y_n$ di **Y.**

Le **singole osservazioni** (x_i,y_j) vengono così rappresentate con dei punti su un piano cartesiano.

7

Tipi di correlazione

- Se i **punti sono sparsi senza apparenti regolarità**

  ⇒ **non c'è correlazione tra i caratteri;**

- Se appare una certa regolarità, ad esempio se:
  - punti con ascissa piccola hanno ordinata piccola e punti con ascissa grande hanno ordinata grande ⇒ c'è correlazione diretta (positiva) tra i due caratteri
  - punti con ascissa piccola hanno ordinata grande e punti con ascissa grande hanno ordinata piccola ⇒ c'è correlazione inversa (negativa) tra i due caratteri

8

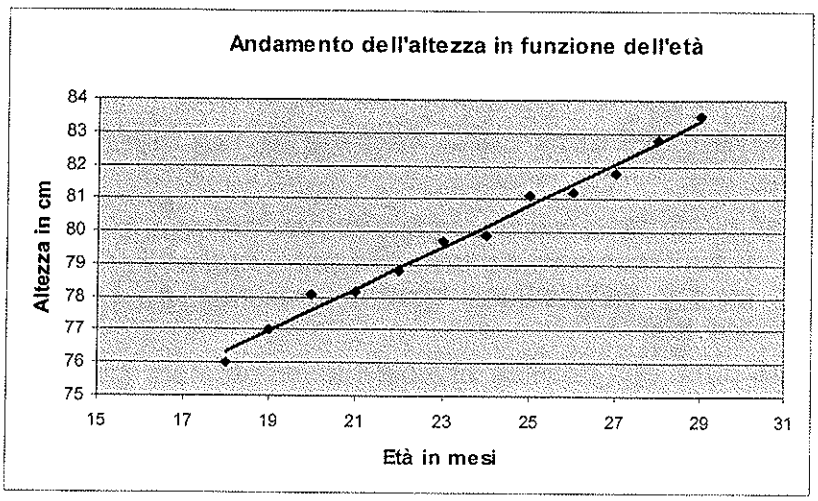
Correlazione lineare

- Se lo scatterplot evidenzia che i punti sono disposti attorno a una **retta crescente** o **decrescente**, si parla di **correlazione lineare**. In tal caso si può tracciare una **retta di regressione** (interpolante) a partire dai dati.
- Altre volte la **correlazione** è ancora evidente, ma la **linearità** non altrettanto.
- In ogni caso la costruzione di indici numerici è necessaria per valutare quantitativamente l'entità dell'associazione

Foglio di lavoro Microsoft Excel

9

La retta interpolante



10

Significato della retta

- La retta che interpola i punti sperimentali rappresenta il modello dei dati e permette di prevedere, data l'età di un bambino, quale dovrebbe essere in media la sua altezza.
- Se osserviamo altezze molto lontane dalla retta, abbiamo a che fare con dei casi anomali.

11

Esempio

- Supponiamo di voler studiare se esiste una relazione tra la misura dell'addome e il peso di 30 neonati.
- Come si sceglie la retta interpolante?
- Come si valuta?

12

PESO (in grammi)	SESSO	ADDOME (circonferenza in cm)
510	Femmina	19
1150	Femmina	24
1200	Femmina	24
1200	Femmina	25
1420	Femmina	25
1450	Femmina	26
1535	Femmina	28
1640	Femmina	28
1665	Femmina	28
1600	Femmina	29
1940	Femmina	29
1740	Femmina	30
1800	Femmina	30
1850	Femmina	30
2000	Maschio	29
2200	Maschio	29
2100	Maschio	30
2300	Maschio	30
2300	Maschio	30
2300	Maschio	30
2550	Maschio	30
2150	Maschio	31
2250	Maschio	31
2350	Maschio	32
2410	Maschio	32
2450	Maschio	32
2550	Maschio	32
2550	Maschio	32
2550	Maschio	32
2320	Maschio	33

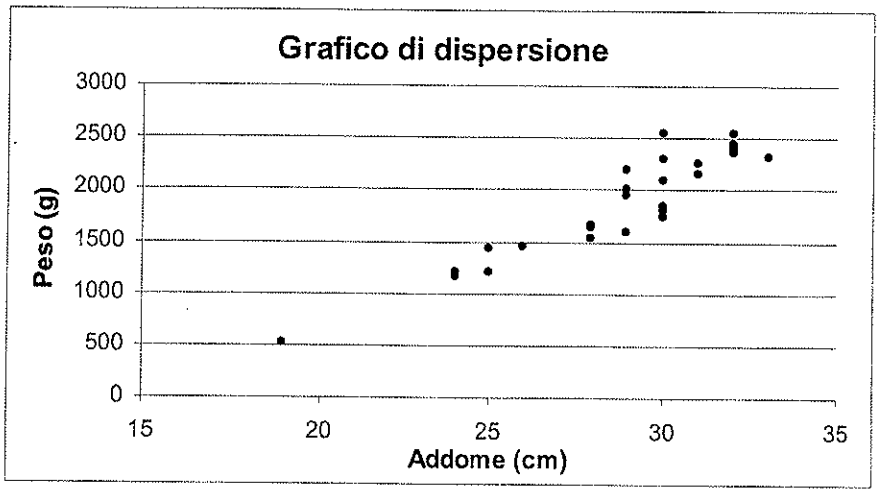
13

Dati grezzi

ADDOME (cm)	PESO (g)	ADDOME (cm)	PESO (g)	ADDOME (cm)	PESO (g)
29	2000	33	2320	30	2550
28	1665	31	2150	30	2300
30	2100	32	2550	30	1740
29	2200	24	1150	31	2250
30	2300	30	2300	32	2350
32	2550	29	1940	19	510
28	1535	28	1640	25	1420
32	2450	30	1800	30	1850
26	1450	29	1600	32	2410
25	1200	24	1200	30	2550

14

Scatter plot



15

Regressione lineare semplice

Variabile indipendente X
Variabile dipendente Y

assume che la relazione tra x e y possa essere riassunta graficamente sotto forma di una retta

$$Y = a + bx$$

pendenza

intercetta

16

Due possibili modalità di utilizzo:

- Predittività
- Correlazione

17

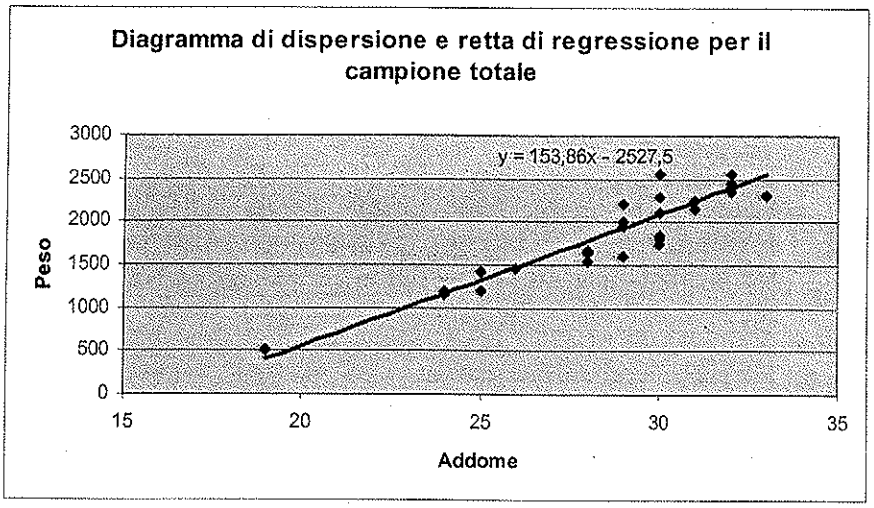
Predittività: Uso di x per predire i valori di y

Grado di correlazione tra x e y è alto

Y sarà un buon predittore di y

18

Interpoliamo i dati sui neonati (*)



19

In realtà la situazione è diversa per maschi e femmine

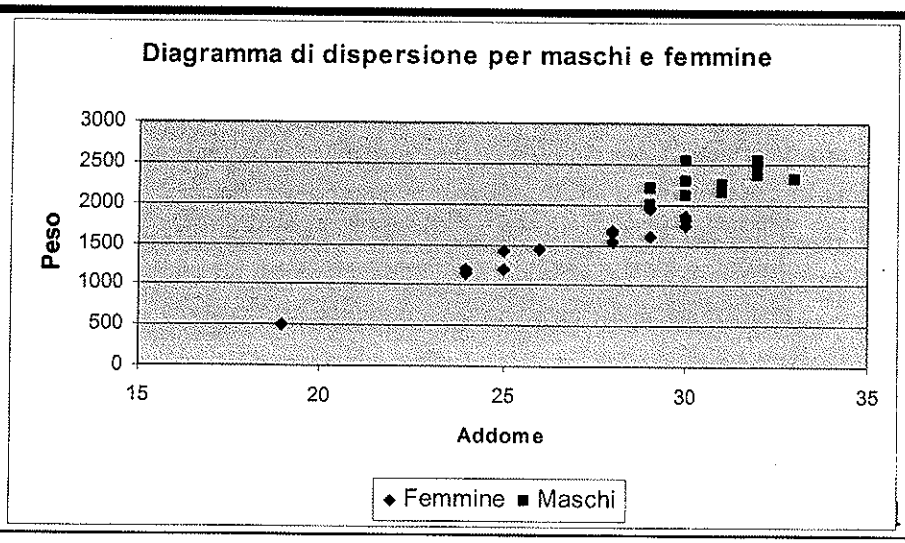
Femmine	
Addome	Peso
19	510
24	1150
24	1200
25	1200
25	1420
26	1450
28	1535
28	1640
28	1665
29	1600
29	1940
30	1740
30	1800
30	1850

Maschi	
Addome	Peso
29	2000
29	2200
30	2100
30	2300
30	2300
30	2300
30	2550
31	2150
31	2250
32	2350
32	2410
32	2450
32	2550
32	2550
32	2550
33	2320

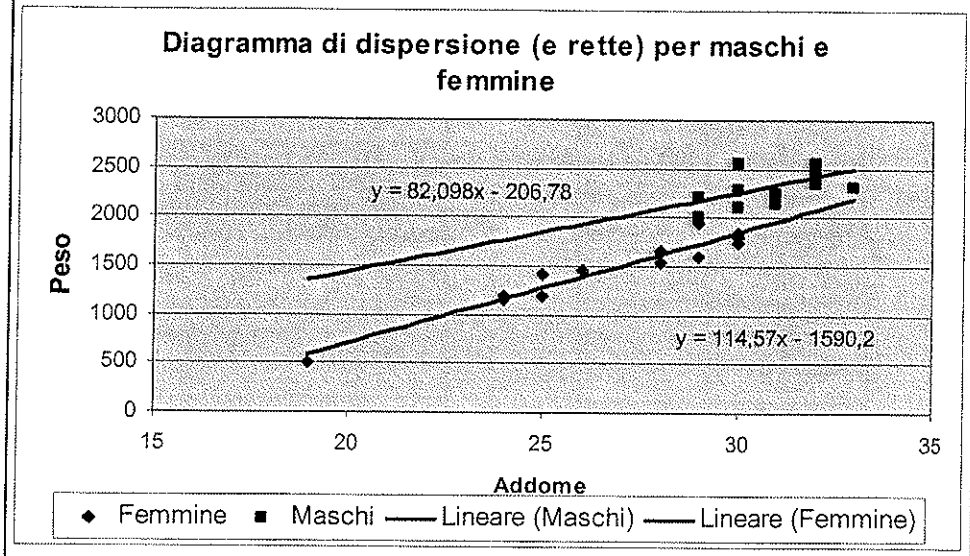
20

Separiamo i due gruppi

In termini tecnici la considerazione dei due gruppi separatamente si chiama stratificazione del campione e serve per effettuare le successive analisi su gruppi (strati) più omogenei tra loro.

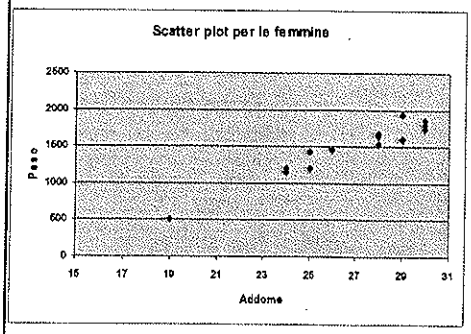


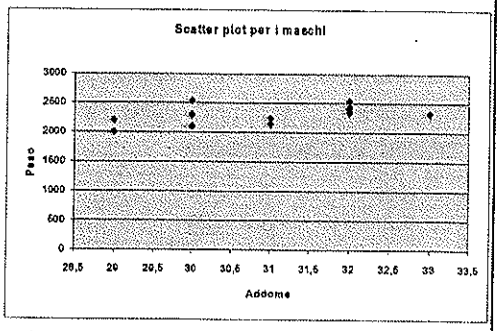
Le due rette



22

Guardiamo le nuvole di punti

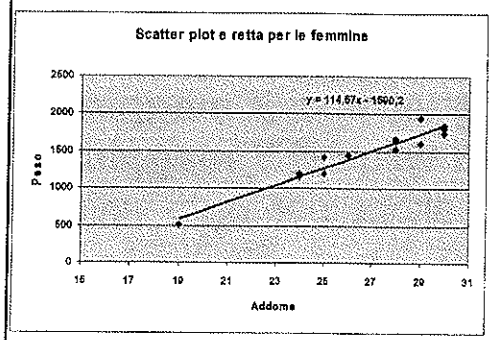


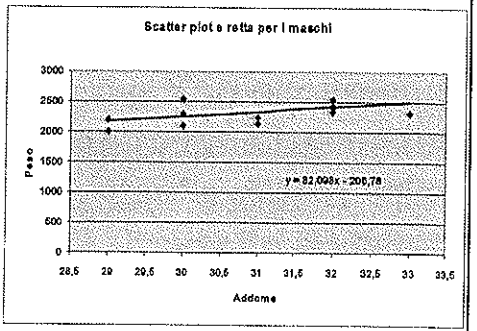


Osserviamo come le femmine mostrino una variabilità molto maggiore dei maschi per la variabile indipendente e questo permette un'analisi più accurata della relazione tra le due variabili. Conferma sulla variabilità si può avere dal calcolo dei coefficienti di variazione per tutte le variabili in gioco.

23

..e le rette interpolanti





24

Correlazione:

Grado di associazione tra la variabile x e la variabile y è espresso dal coefficiente di correlazione lineare ⟹ Grado di relazione lineare tra x e y

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

ρ_{xy} varia tra -1 e 1

Scarti standard delle due variabili

Covarianza

$$\sigma_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

25

Indici di correlazione

Se abbiamo **n** osservazioni congiunte di due caratteri, $\{(x_1,y_1),\ldots,(x_i,y_i),\ldots,(x_n,y_n)\}$, e le medie aritmetiche di X e Y e se a valori di X **maggiori** della media sono associati valori di Y **maggiori** della media e viceversa, allora per ogni unità statistica gli scarti $(x_i - \bar{x})$ e $(y_i - \bar{y})$ tendono a essere concordi, e quindi i loro **prodotti**

$$(x_i - \bar{x})(y_i - \bar{y}) > 0$$

sono con maggiore frequenza **positivi**.

26

Covarianza

Si dice covarianza dei due **caratteri X** e **Y** la funzione:

$$\sigma_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

n-1 sono i gradi di libertà della covarianza.

27

Covarianza

si può dire che se:

- $\sigma_{xy} > 0 \Rightarrow$ mediamente a valori grandi (piccoli) di X corrispondono valori grandi (piccoli) di Y, e **Y è direttamente correlato X, (X e Y sono direttamente correlati)** ;
- $\sigma_{xy} < 0 \Rightarrow$ mediamente a valori grandi (piccoli) di X corrispondono valori piccoli (grandi) di Y, e **Y è inversamente correlato a X, (X e Y sono direttamente correlati)** ;
- $\sigma_{xy} = 0 \Rightarrow$ **X e Y sono incorrelati o la loro relazione non è monotona**

28

Coefficiente di correlazione

Il valore di σ_{xy} non è utilizzabile per misurare **quanto** sia stretto il legame tra X e Y, perché dipende dalle unità di misura dei 2 caratteri.

Dividendo la σ_{xy} per il prodotto degli scarti standard delle due variabili si ottiene il ***coefficiente di correlazione di X e Y***:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

29

Calcolo degli indici per le femmine

x-media(x)	y-media(y)	(x-media(x))(y-media(y))
-7,79	-968,57	7545,16
-2,79	-328,57	916,71
-2,79	-278,57	777,21
-1,79	-278,57	498,64
-1,79	-58,57	104,84
-0,79	-28,57	22,57
1,21	56,43	68,28
1,21	161,43	195,33
1,21	186,43	225,58
2,21	121,43	268,36
2,21	461,43	1019,76
3,21	261,43	839,19
3,21	321,43	1031,79
3,21	371,43	1192,29
3,14	372,81	deviazione standard
	cov	1131,21
	corr	0,97

I coefficienti di variazione risultano rispettivamente:

CV(addome)=12%;

CV(peso)=25%.

30

Calcolo degli indici per i maschi

x-media(x)	y-media(y)	(x-media(x))(y-media(y))
-1,94	-333,13	646,27
-1,94	-133,13	258,27
-0,94	-233,13	219,14
-0,94	-33,13	31,14
-0,94	-33,13	31,14
-0,94	-33,13	31,14
-0,94	216,87	-203,86
0,06	-183,13	-10,99
0,06	-83,13	-4,99
1,06	16,87	17,88
1,06	76,87	81,48
1,06	116,87	123,88
1,06	216,87	229,88
1,06	216,87	229,88
1,06	216,87	229,88
2,06	-13,13	-27,05
1,24	170,28	deviazione standard
	cov	125,54
	corr	0,60

I coefficienti di variazione risultano rispettivamente:

CV(addome)=4%;

CV(peso)=7%.

Totale	
1436,00	cov
0,93	corr

Il valore del coefficiente di correlazione per la popolazione totale è intermedio, la covarianza è maggiore

31

Proprietà

- ρ_{xy} varia tra -1 e 1;
- ρ_{xy} è esattamente uguale a 1 se e solo se le due variabili sono correlate positivamente e i punti (x_i,y_i) risultano allineati;
- ρ_{xy} è esattamente uguale a -1 se e solo se le due variabili sono correlate negativamente e i punti (x_i,y_i) risultano allineati;
- Quanto più ρ_{xy} è prossimo a 1 in valore assoluto, tanto più stretta è l'associazione tra le due variabili.

32

Proprietà

- ρ_{xy} è un indice "normalizzato", cioè la cui grandezza ha un significato assoluto, è adimensionale.
- opera direttamente sui valori osservati dei caratteri quantitativi senza passare a misure raggruppate in classi, che producono perdita di informazione;
- se il legame tra le variabili è non lineare ρ_{xy} **basso in valore assoluto (prossimo a 0) non indica assenza di associazione.**

33

Residui (*)

- Si definiscono **valori stimati** i numeri:

$$y_i^* = a^* + b^* x_i$$

che rappresentano i valori del carattere **Y** stimati dalla retta di regressione in corrispondenza delle osservazioni x_i.

- si definiscono ***residui assoluti*** le differenze che corrispondono a x_i sulla retta e quelli effettivamente osservati:

$$r_i = y_i - y_i^*$$

- ***Per valutare la bontà dell'approssimazione ottenuta dal modello, rispetto all'andamento mostrato dai dati sperimentali, si può partire dallo studio dei residui.***

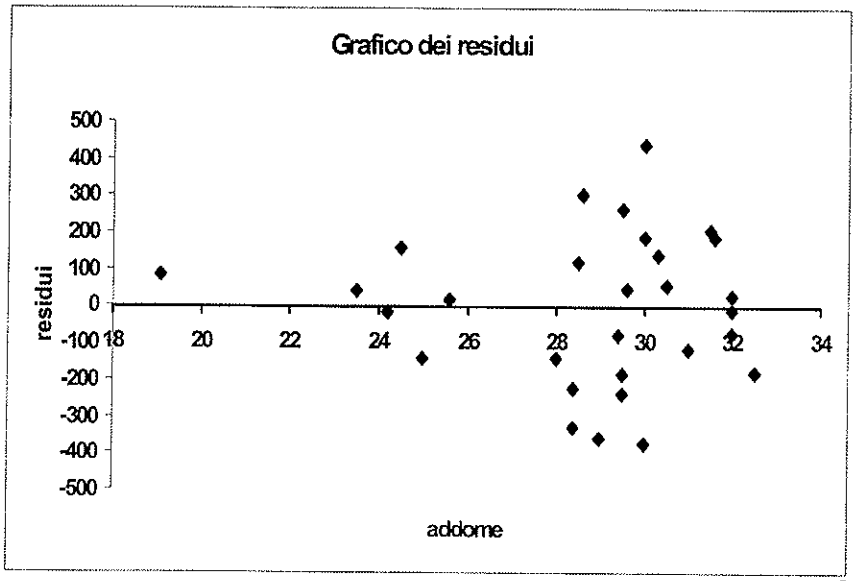
34

Regressione Lineare

Perché ci sia un ***buon adattamento del modello*** ai dati vogliamo che:

- i residui non individuino alcun andamento ulteriormente interpolabile con termini di ordine superiore.
- il segno dei residui sia "casuale", sia cioè, in qualche modo, ripartito equamente tra + e-, per escludere errori sistematici
  - ✓ rappresentando su di un grafico i punti (x_i, r_i) se la relazione tra **X** e **Y** è lineare dovremmo osservare dei punti che oscillano casualmente sopra e sotto lo 0.

35



36

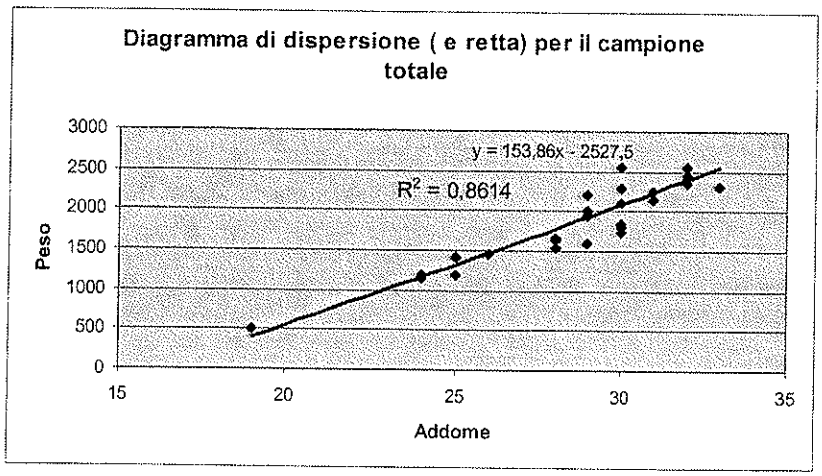
Il coefficiente di determinazione R^2

$$R^2 = (\rho_{XY})^2$$

- R^2 varia tra 0 e 1.
- R^2 prossimo a 1 $\Rightarrow$ *buon adattamento della retta di regressione ai dati osservati.*
- R^2 prossimo a 0 $\Rightarrow$ *cattivo adattamento della retta di regressione ai dati osservati.*

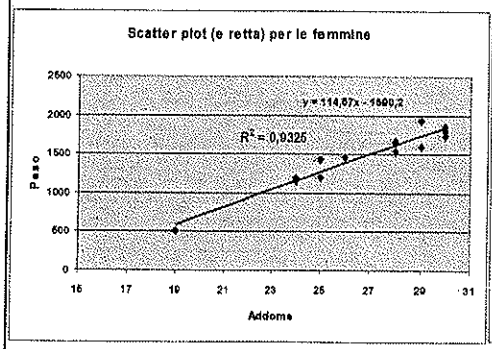
37

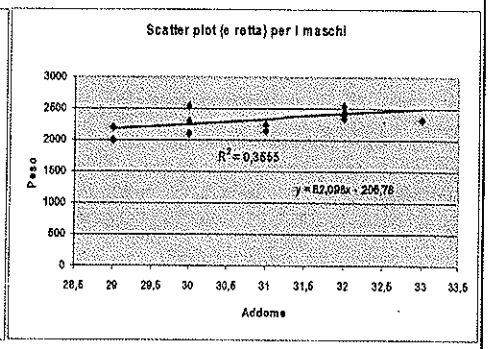
Valutiamo



38

Confrontiamo





39

Regressione lineare

Un possibile ed efficace utilizzo di R^2 è nel confronto di diversi modelli di regressione. Consideriamo il seguente esempio legato ad un problema agrario.

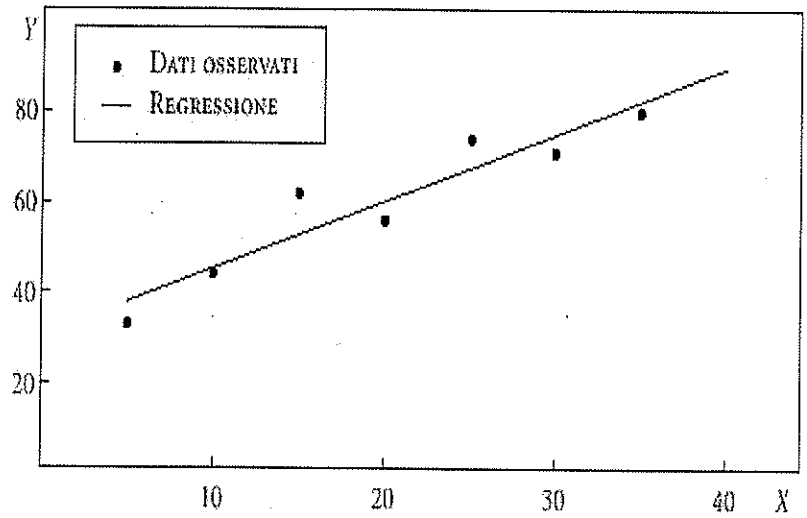
Si vuole misurare la crescita delle radici del mais in funzione del contenuto di saccarosio nel nutrimento fornito.

I seguenti dati si riferiscono alla crescita in millimetri della radice del mais (Y), coltivata in vitro per 10 giorni, in relazione al contenuto di saccarosio (X), misurato in g/l (grammi per litro), nel terreno di coltura:

$$x = \{5, 10, 15, 20, 25, 30, 35\}; \qquad y = \{33, 44, 62, 56, 74, 71, 80\}.$$

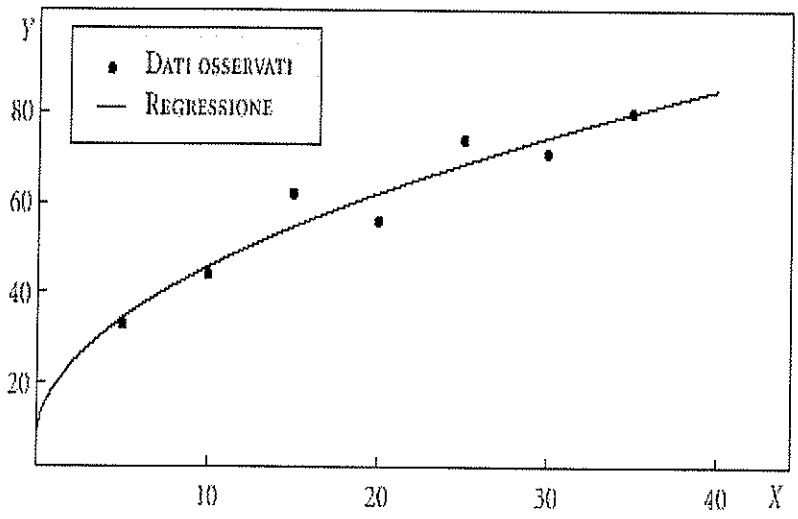
40

Regressione lineare R^2=0.889



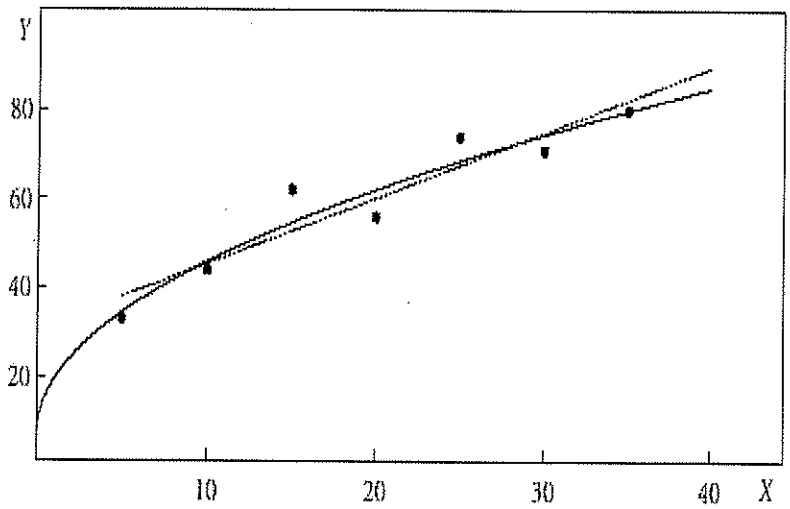
41

Regressione lineare R^2=0.92



42

Regressione lineare



43

I due modelli

- I due modelli differiscono in quanto il secondo (che ha come variabile indipendente $\sqrt{x}$ invece di x) introduce una curvatura che modella bene l'effetto di saturazione per dosi alte del nutriente, tipico in analisi del tipo dose-risposta.

44

Tabella Osservata

	positivo	negativo
grave	85	40
medio	125	95
lieve	150	145
Totale	360	280

TabellaAttesa

	positivo	negativo	Totale
grave	70,31	54,69	125
medio	123,75	96,25	220
lieve	165,94	129,06	295
Totale	360	280	640

Tabella delle differenze			Tabella differenze al quadrato		Tabella degli addendi dell'indice	
grave	14,69	-14,69	215,80	215,80	3,07	3,95
medio	1,25	-1,25	1,56	1,56	0,01	0,02
lieve	-15,94	15,94	254,08	254,08	1,53	1,97
Totale	0	0	**valore dell'indice**	10,54		

Tabella degli addendi dell'indice

	3,069209	3,945805
	0,012626	0,016234
	1,531178	1,968725
valore dell'	10,54378	

Tabella degli ODDS e ODDS RATIO

ODDS (grave/lieve)		
	0,57	0,28
ODDS RATIO		2,05
ODDS (grave/medio)		
	0,68	0,42
ODDS RATIO		1,62
ODDS (medio/lieve)		
	0,83	0,66
ODDS RATIO		1,27

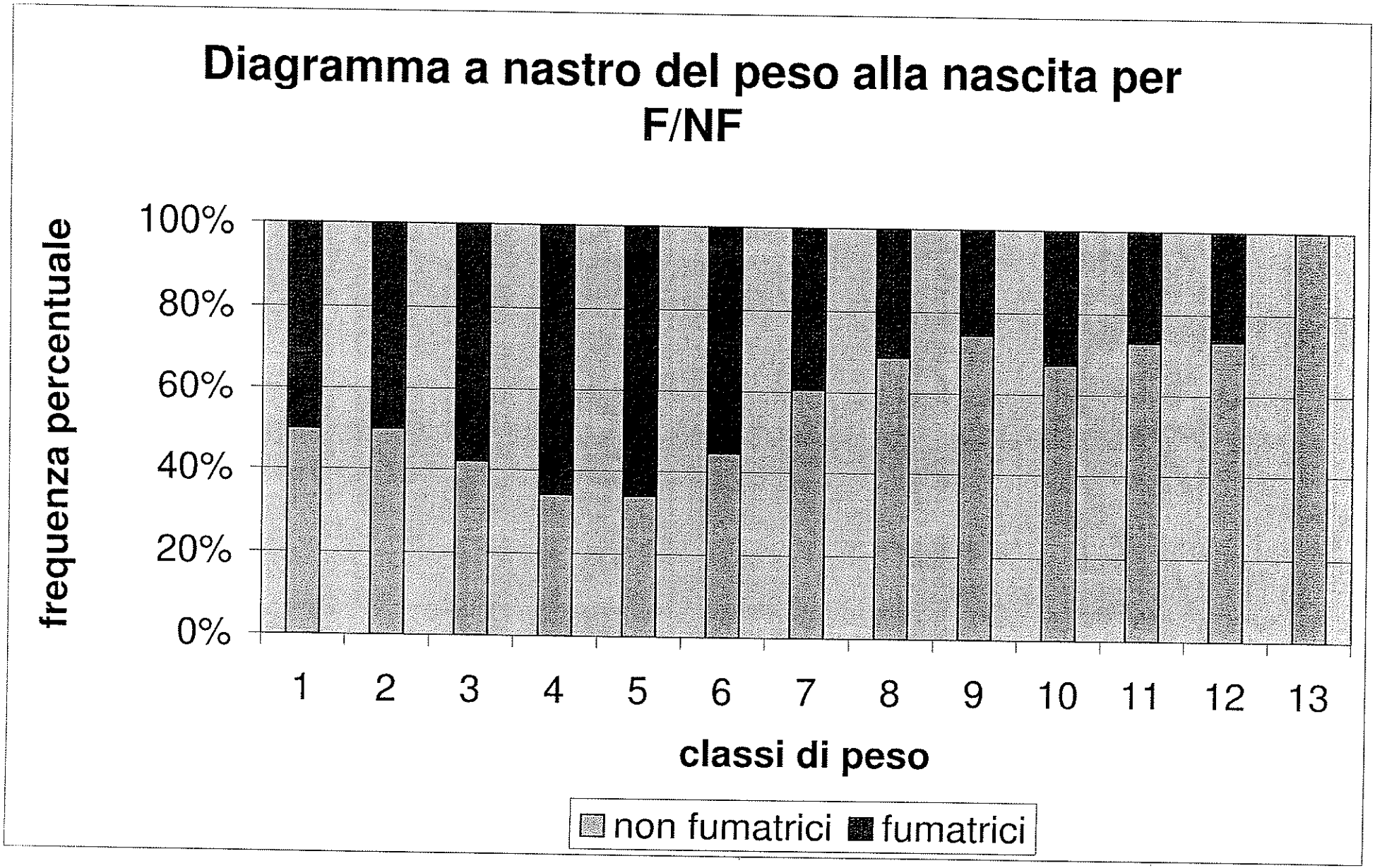


Diagramma a nastro del peso alla nascita per F/NF
frequenza percentuale
100%
80%
60%
40%
20%
0%
1
2
3
4
5
6
7
8
9
10
11
12
13
classi di peso
non fumatrici
fumatrici

ANALISI STATISTICHE BIVARIATE

Tabelle di contingenza

Analisi Statistica Bivariata

Generalmente, lo studio quantitativo di un fenomeno di interesse si svolge rilevando contemporaneamente più caratteri su ciascuna unità statistica, per spiegare il fenomeno attraverso le interconnessioni tra le variabili.

L'analisi statistica bivariata consiste nello *studio del comportamento dei due caratteri* ***congiuntamente*** *per ogni unità statistica.*

Analisi Statistica Bivariata

I dati vengono presentati mediante la distribuzione di frequenze doppie, detta **tabella di contingenza**, ottenuta suddividendo le unità in *classi* secondo entrambi i caratteri e *contando* poi le unità di ciascuna delle classi.

- ✓ *Caratteri qualitativi* ⇒ le classi corrispondono alle modalità;
- ✓ *Caratteri quantitativi* ⇒ le classi vengono identificate raggruppando prima i valori assumibili dai caratteri.

Tabella di contingenza

Per i dati relativi all'influenza dell'abitudine al fumo sul peso alla nascita è immediato costruire la tabella di contingenza, avendo raggruppato in classi i valori del peso alla nascita. Rispetto alla tabella già utilizzata per la costruzione dei grafici c'è una colonna ulteriore in cui compare la somma delle frequenze riga per riga. Mentre nell'ultima riga compare la somma colonna per colonna delle frequenza. L'ultima casella in basso a destra riporta il valore della numerosità campionaria. L'ultima riga e l'ultima colonna sono dette ***distribuzioni marginali*** e rappresentano le distribuzioni di ognuno dei due caratteri considerati singolarmente (distribuzioni univariate).

Classe birth weight	Frequenza assoluta non fumatrici	Frequenza assoluta fumatrici	Frequenza marginale
50-59	1	1	2
60-69	3	3	6
70-79	8	11	19
80-89	13	25	38
90-99	31	60	91
100-109	76	94	170
110-119	166	108	274
120-129	198	90	288
130-139	140	48	188
140-149	62	30	92
150-159	27	10	37
160-169	11	4	15
170-179	6	0	6
Frequenza marginale	742	484	1226

Come si costruisce una tabella di contingenza

X e Y sono le variabili osservate in ciascuna delle **n** unità del collettivo (abitudine al fumo e peso alla nascita) e $x_1, x_2, \ldots, x_s$ e $y_1, y_2, \ldots, y_t$ le modalità assunte.

n_{ij} è la frequenza assoluta delle unità in cui X assume modalità x_i e Y modalità y_j ;

$n_{i\bullet}$ la frequenza assoluta delle modalità x_i di X nel collettivo (frequenza marginale) e

$n_{\bullet j}$ quella della modalità y_j di Y.

Tabella di Contingenza

Modalità Carattere X	Modalità Carattere Y					Totale di riga
	y_1	…	y_j	…	y_t	
x_1	n_{11}	…	n_{1j}	…	n_{1t}	$n_{1\bullet}$
…	…	…	…	…	…	…
x_i	n_{i1}	…	n_{ij}	…	n_{it}	$n_{i\bullet}$
…	…	…	…	…	…	…
x_s	n_{s1}	…	n_{sj}	…	n_{st}	$n_{s\bullet}$
Totale di Colonna	$n_{\bullet 1}$	…	$n_{\bullet j}$	…	$n_{\bullet t}$	n

Distribuzioni bivariate

Anche una ***distribuzione congiunta di frequenze assolute*** può essere normalizzata per ottenere la *distribuzione congiunta di frequenze relative*. A partire dalla tabella è possibile procedere a diversi tipi di normalizzazione in relazione alle caratteristiche della distribuzione che si intende mettere in luce.

- Dividendo ogni cella della tabella per il numero totale di osservazioni, **n**, si ottiene la *distribuzione delle* ***frequenze relative dei due caratteri considerati congiuntamente***.
- A margine si ottengono le *distribuzioni di frequenze relative dei due caratteri considerati separatamente* (***distribuzioni univariate***).

Distribuzioni Condizionate

Può, però, essere di maggior interesse studiare come le distribuzioni riportate sulle varie righe dipendano dalla classificazione di colonna (o viceversa).

Cioè, una volta fissata una riga, che corrisponde a fissare una modalità del **carattere X**, si studia la distribuzione, su quella riga, del **carattere Y**.

In questo caso si studia la ***distribuzione condizionata di Y dato X*** e la ***variabile statistica condizionata*** si denota con **Y/X**.

Distribuzione relativa condizionata

La distribuzione normalizzata si ottiene dividendo i valori delle caselle della riga fissata per il totale di riga.

Ci sono tante distribuzioni condizionate di Y/X quante sono le righe della tabella.

Analogamente si calcolano le ***distribuzioni condizionate di X/Y***, scambiando il ruolo delle righe e delle colonne.

Lo studio delle distribuzioni condizionate, permette di evidenziare l'influenza di una delle variabili sulla variabilità dell'altra.

Distribuzione congiunta e condizionata di frequenze

Ricoveri in un servizio psichiatrico per sesso ed esito del ricovero

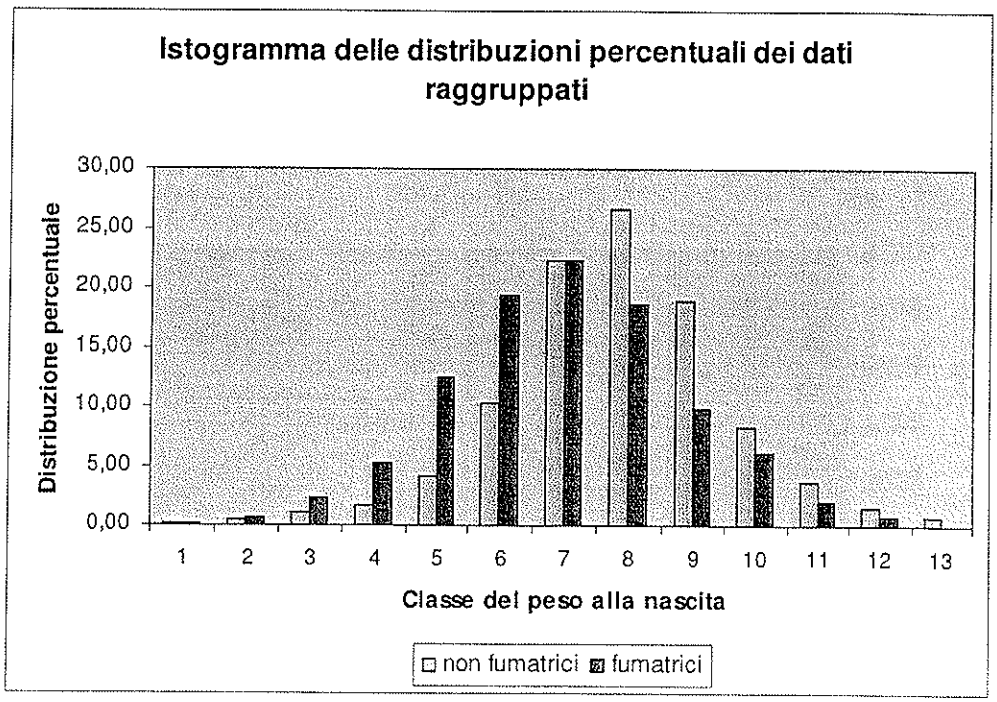
Esito	Maschi	Femmine	Totale
Allontanamento	35 (33,3%)	18 (22%)	53 (28,4%)
Dimissioni	44 (42%)	46 (56,1%)	90 (48,1%)
Trasferimento	5 (4,7%)	7 (8,5%)	12 (6,4%)
Decesso	-	1 (1,2%)	1 (0,5%)
Non Rilevati	21 (20%)	10 (12,2%)	31 (16,6%)
Totale	105 (100%)	82 (100%)	187 (100%)

Rappresentazioni grafiche

- Un primo tipo di rappresentazione consiste nella generalizzazione del ***diagramma a barre.***
- Si costruisce un diagramma a barre per ogni riga (o colonna) della tabella, differenziando i diversi diagrammi per colore.
- Si affiancano, poi, le barre di diverso colore che corrispondono alla stessa modalità dell'altro carattere.

Foglio di lavoro Microsoft Excel

Diagramma a barre affiancate

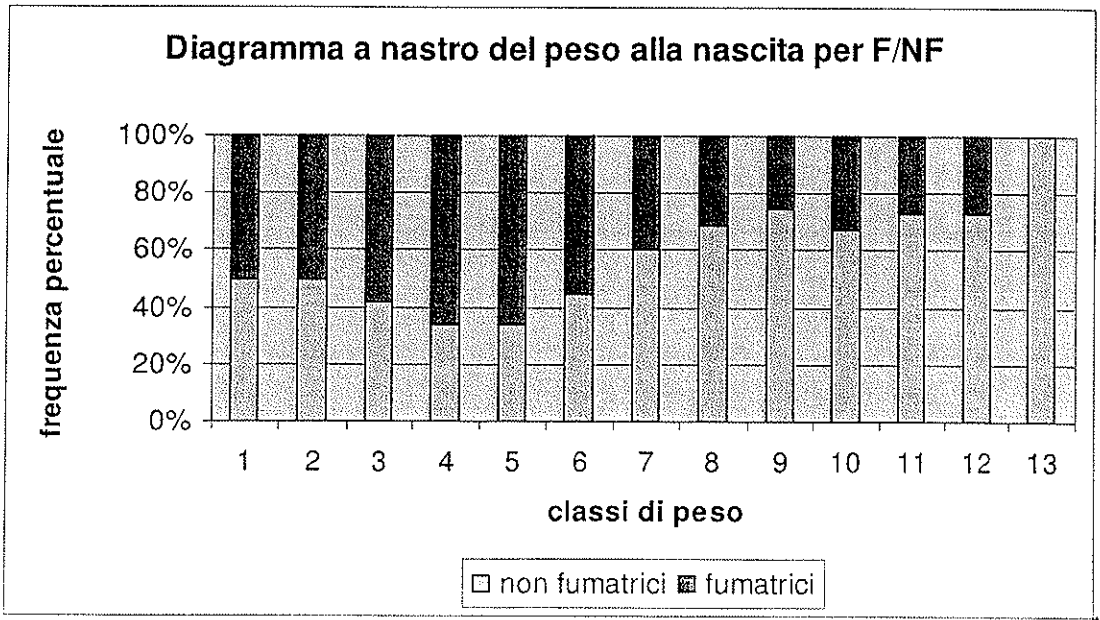


Associazione tra caratteri

- E' interessante studiare l'eventuale associazione tra due caratteri ovvero verificare se alcune modalità di uno dei due caratteri siano associate preferibilmente con alcune modalità dell'altro.
- Un primo approccio grafico a tale analisi può essere effettuato rappresentando le distribuzioni condizionate di riga, o di colonna, mediante i *diagrammi a nastro*.

Diagramma a nastro

Osserviamo come le percentuale di madri fumatrici sia maggiore tra i neonati con peso maggiore



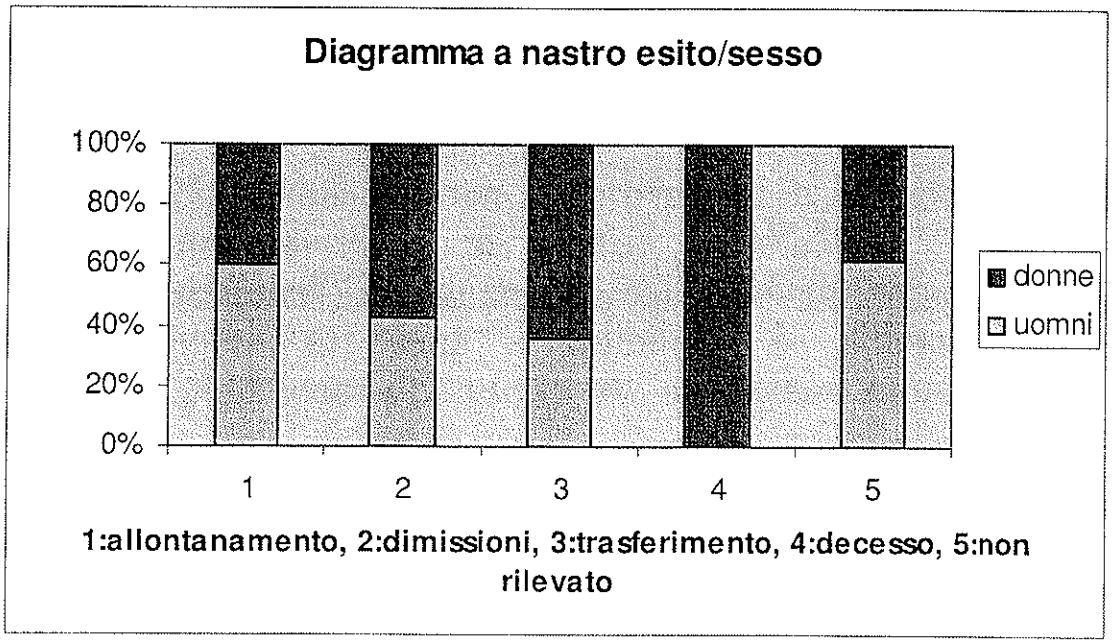
Diagrammi a nastro

I ***diagrammi a nastro*** rappresentano le *distribuzioni condizionate di riga*, o *di colonna:* sono idonei a **studiare qualitativamente l'associazione tra caratteri.**

Rispetto ai diagrammi a barre, si suddivide ogni rettangolo corrispondente ad una modalità, per esempio di riga, in parti proporzionali alle frequenze delle colonne relative a quella riga fissata. Il *rettangolo iniziale* avrà *dimensione proporzionale* alla *frequenza assoluta* se la distribuzione considerata è in forma di *frequenze assolute*, avrà, invece, *dimensione unitaria*, se si tratta di *distribuzione di frequenze relative.*

Foglio di lavoro Microsoft Excel

Ricoveri in un servizio psichiatrico per sesso ed esito del ricovero



Associazione esito/sesso

- La suddivisione dei rettangoli per sesso è piuttosto sbilanciata.
- Gli uomini mostrano con maggiore frequenza l'esito allontanamento rispetto alle donne che sembrano più frequenti nelle dimissioni e nei trasferimenti.
- E' possibile introdurre opportuni indici di connessione per misurare quantitativamente l'associazione tra le variabili.

Connessione Statistica

La connessione statistica studia di quanto le ***distribuzioni condizionate delle frequenze osservate*** si discostino da quelle che ci si aspetterebbe (***attese***) se i caratteri si comportassero in modo ***indipendente***.

Se le modalità di un carattere non avessero influenza sulle modalità dell'altro, tutte le distribuzioni relative condizionate, dovrebbero essere uguali (frequenze assolute proporzionali) e coincidere con la corrispondente distribuzione marginale. Varrebbe la proporzione:

$$n_{ij}^{*} : n_{i\bullet} = n_{\bullet j} : n$$

Colonne proporzionali

Se vale la proporzione, allora il prodotto degli estremi è uguale al prodotto dei medi e possiamo ricavare

$$n_{ij}^{*} = \frac{n_{\bullet j} n_{i\bullet}}{n}$$

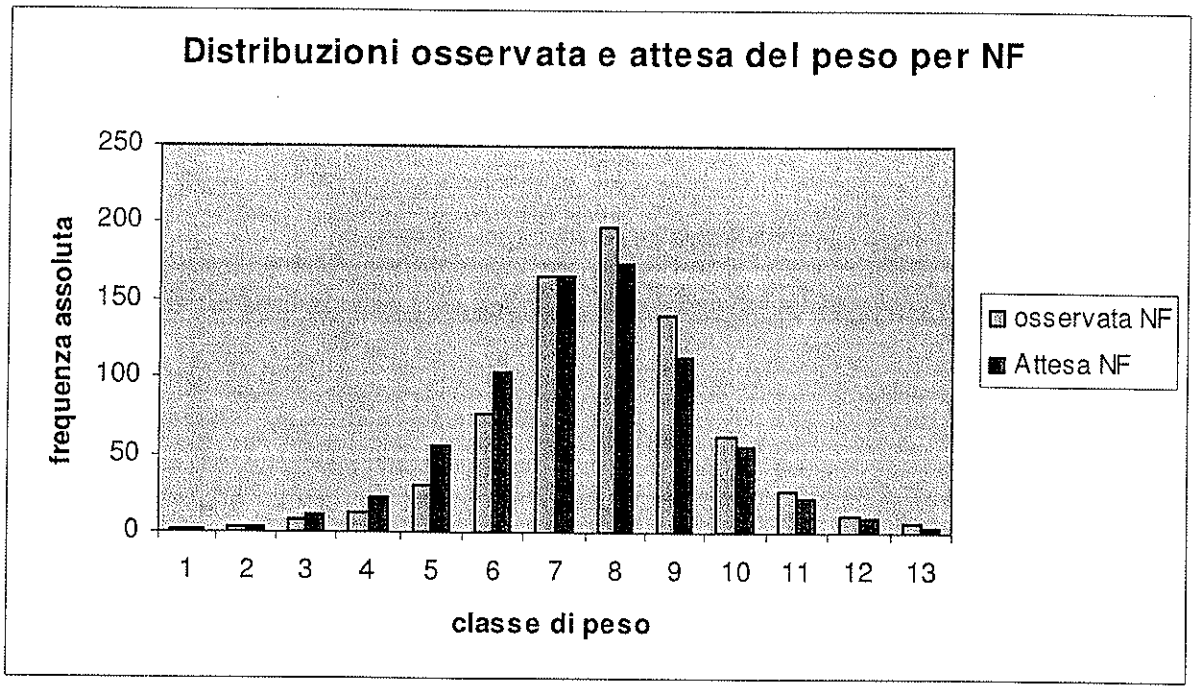
e confrontare con i valori osservati effettivamente (tabella osservata).

Classe birth weight	Frequenza assoluta non fumatrici	Frequenza assoluta fumatrici	Frequenza marginale
50-59	1	1	2
60-69	3	3	6
70-79	8	11	19
80-89	13	25	38
90-99	31	60	91
100-109	76	94	170
110-119	166	108	274
120-129	198	90	288
130-139	140	48	188
140-149	62	30	92
150-159	27	10	37
160-169	11	4	15
170-179	6	0	6
Frequenza marginale	742	484	1226

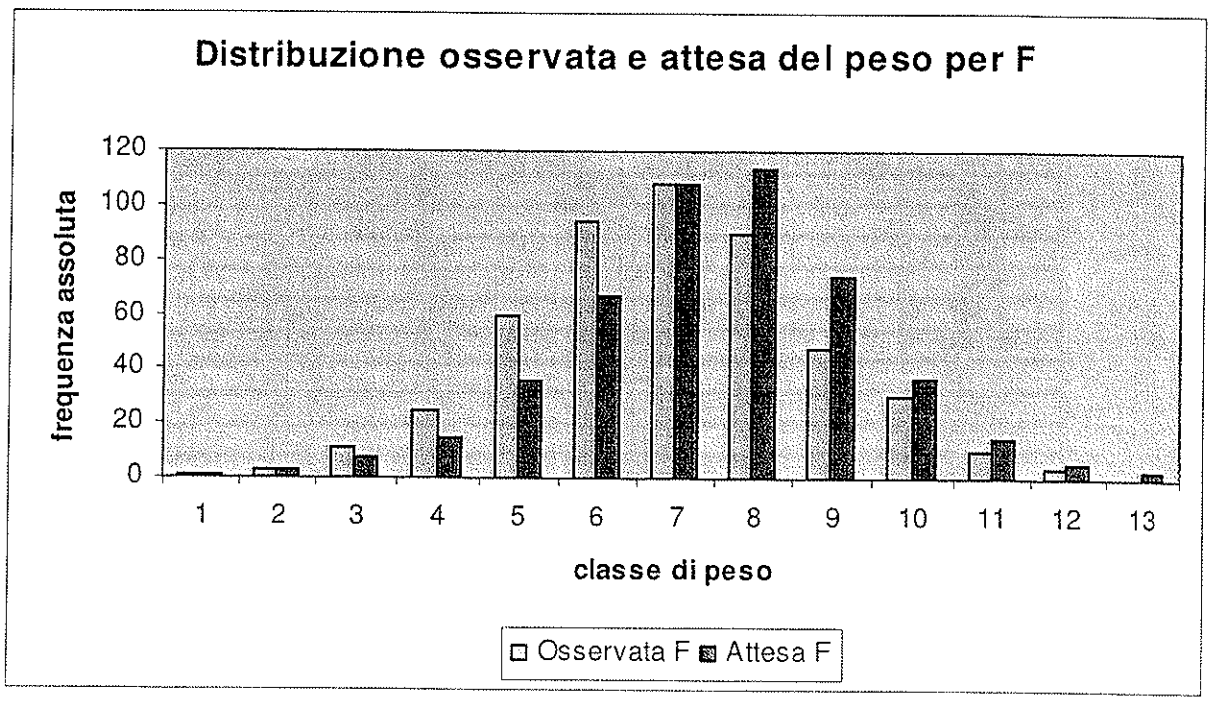
Tabella attesa

Birth weight	Osservata NF	Attesa NF	Osservata F	Attesa F	Marginale
50-59	1	1,21	1	0,79	2
60-69	3	3,63	3	2,37	6
70-79	8	11,50	11	7,50	19
80-89	13	23,00	25	15,00	38
90-99	31	55,08	60	35,92	91
100-109	76	102,89	94	67,11	170
110-119	166	165,83	108	108,17	274
120-129	198	174,30	90	113,70	288
130-139	140	113,78	48	74,22	188
140-149	62	55,68	30	36,32	92
150-159	27	22,39	10	14,61	37
160-169	11	9,08	4	5,92	15
170-179	6	3,63	0	2,37	6
marginale	742	484	742	484	1226

Confronto delle frequenze per NF



Confronto delle frequenze per NF



Ricerca di fattori di rischio

- Un ricercatore medico vuole studiare l'associazione tra gravità degli attacchi cardiaci e presenza di un tipo di anticorpo nel sangue. Per fare questo esamina 640 vittime di attacchi cardiaci.
- I risultati sono riportati nella tabella seguente:

Tabella osservata

Test per anticorpo / Gravità dell'attacco	+	-	Marginale
+++	85	40	125
++	125	95	220
+	150	145	295
Marginale	360	280	640

Connessione

- Nella prima tabella le distribuzioni condizionate non sembrano tra loro proporzionali.
- Possiamo, allora, calcolare la tabella attesa in caso di non influenza di un carattere sull'altro utilizzando la formula:

$$n^{*}_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

Tabella attesa

Test per anticorpo / Gravità dell'attacco	+	-	Marginale
+++	70,31	54,69	125
++	123,75	96,25	220
+	165,94	129,06	295
Marginale	360	280	640

Confrontare

Se confrontiamo le caselle estreme delle due tabelle possiamo osservare che c'e' un eccesso di soggetti osservati nella prima casella della prima riga e nell'ultima della terza, compensati da altrettanti casi in meno nell'altra casella della stessa riga. Ovvero, i casi con attacco grave e presenza di anticorpo sono più di quelli che ci si attenderebbe se questo non avesse influenza sulla gravità e, analogamente, i casi lievi senza anticorpo sono di più di quelli attesi. Questo è già un indizio di associazione tra presenza dell'anticorpo e gravità dell'attacco cardiaco.

Test per anticorpo / Gravità dell'attacco	+	-	Marginale
+++	85	40	125
++	125	95	220
+	150	145	295
Marginale	360	280	640

Test per anticorpo / Gravità dell'attacco	+	-	Marginale
+++	70,31	54,69	125
++	123,75	96,25	220
+	165,94	129,06	295
Marginale	360	280	640

Indice χ^2

- Dato che le tabelle, osservata e attesa, sono piuttosto diverse, vogliamo misurare tale differenza con un opportuno indice.
- L'indice è basato sulle differenze delle frequenze delle caselle corrispondenti delle due tabelle.
- Il χ^2 (**Chi quadrato**) è un indice statistico, introdotto da **Pearson**, e definito nel modo seguente:

$$\chi^2 = \frac{\sum_i (n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

Calcolo del χ^2

Il primo passo per il calcolo parte dalle tabelle osservata e attesa e costruisce la tabella delle differenze cella per cella elevate al quadrato.

Tabella Osservata			TabellaAttesa			
	positivo	negativo		positivo	negativo	Totale
grave	85	40	grave	70,31	54,69	125
medio	125	95	medio	123,75	96,25	220
lieve	150	145	lieve	165,94	129,06	295
Totale	360	280	Totale	360	280	640

Dalle differenze al quadrato, dividendo per i valori delle caselle corrispondenti della tabella attesa si calcola la tabella degli addendi da sommare per ottenere il valore dell'indice.

Tabella delle differenze			**Tabella differenze al quadrato**		**Tabella degli addendi dell'indice**	
grave	14,69	-14,69	215,80	215,80	3,07	3,95
medio	1,25	-1,25	1,56	1,56	0,01	0,02
lieve	-15,94	15,94	254,08	254,08	1,53	1,97
Totale	0	0	**valore dell'indice**	10,54		

Indice χ^2

- L'indice misura la distanza euclidea pesata tra le celle della tabella osservata e quelle corrispondenti della tabella attesa.
- Ogni differenza è pesata con l'inverso della frequenza attesa della cella per la necessità di considerare maggiormente influenti celle con frequenze alte. Infatti, mentre il termine al numeratore è di tipo quadratico, quello al denominatore è lineare e questo fatto rende maggiormente pesanti le differenze di celle con frequenza attesa maggiore.
- **Per le tabelle precedenti risulta χ^2 =10,54.**

Connessione statistica

- Dovremo studiare la variabilità teorica dell'indice per poter decidere quando considerarlo "grande" o "piccolo".
- La variabilità dipende dai gradi di libertà della tabella.
- Se abbiamo r righe e c colonne, con margini fissati, possiamo scegliere arbitrariamente r-1 righe e c-1 colonne. I gradi di libertà sono (r-1)(c-1).

Associazione di caratteri genetici

- Nelle tabelle che seguono sono riportati i dati osservati relativi a due gruppi di soggetti in cui si è rilevata la presenza di daltonismo rosso-verde per verificare se ci fosse associazione rispettivamente con il carattere sesso e con la presenza di sordità congenita.

Il daltonismo rosso-verde

Tabelle osservate			
Dalton r-v	maschio	femmina	Tot
positivo	420	68	488
negativo	4900	4600	9500
Tot	5320	4668	9988
Dalton r-v	sordo	non sordo	Tot
positivo	45	7500	7545
negativo	450	90000	90450
Tot	495	97500	97995

Il daltonismo rosso-verde

Tabelle attese			
Dalton r-v	maschio	femmina	Tot
positivo	259,93	228,07	488
negativo	5060,07	4439,93	9500
Tot	5320	4668	9988
Dalton r-v	sordo	non sordo	Tot
positivo	38,11	7506,89	7545
negativo	456,89	89993,11	90450
Tot	495	97500	97995

Risultati dei test

- Il valore del χ^2 è rispettivamente:
- χ^2=221,76
- χ^2=1,36
- In entrambi i casi il numero di gradi di libertà è 1
- **Il daltonismo è verosimilmente legato al sesso, ma non alla sordità.**