UNIVERSITA' DEGLI STUDI DI FIRENZE

Scuola di Economia e Management

Il Paradosso di Simpson nell'analisi statistica: Profilo storico e casi di studio

Autore: **Antonio Pupo**

Relatore: **Prof. Leonardo Grilli**

Anno Accademico 2012 - 2013

INDICE

INTRODUZIONE

Scopo di questo lavoro è la descrizione del cosiddetto “Paradosso di Simpson”, inquadrabile nei paradossi statistici, che permette una lettura dei dati a disposizione più approfondita rispetto a conclusioni, a prima vista incoerenti.

Partendo dalla definizione generica di paradosso, nel primo capitolo, si evidenzia la particolarità del paradosso statistico in questione, citando gli studiosi che hanno contribuito alla sua definizione, in modo particolare, George Udny Yule che ne evidenzia i tratti essenziali in un suo lavoro del 1903.

Non si tralasciano infine altri studiosi che con studi similari hanno contribuito alla definizione del lavoro di Simpson.

Si darà poi particolare risalto alla figura di Edward Hugh Simpson, descrivendo il lavoro che ha portato a definire compiutamente quello che, alcuni anni dopo, è stato definito, da altri, il paradosso di Simpson.

Verrà evidenziata l’opera di Blyth che, oltre a fornire, nel titolo di un suo lavoro, il nome attuale del paradosso, si è adoperato per la sua divulgazione cercando anche di semplificarne le modalità di presentazione.

Infine, negli anni più recenti, con l’avvento dei computer e lo sviluppo di software dedicati, sono stati costruiti degli applicativi che consentono, anche ai non esperti di statistica, di verificare la presenza del paradosso di Simpson nei casi oggetto dei loro studi.

A tal proposito verrà presentato un applicativo che, con i dati a disposizione, mostra in una forma grafica il problema e permette di studiarlo in tempo reale attraverso la sostituzione dei dati fondamentali, in quanto elabora in maniera immediata la nuova situazione.

Nel secondo capitolo si descriveranno alcuni dei casi più importanti per evidenziare la valenza del paradosso in questione.

Si riporterà il caso, ormai storico, dell'università di Berkeley, accusata, in un primo momento, di non favorire, nella fase di ammissione le candidate di sesso femminile e, successivamente, assolta attraverso l'utilizzo del paradosso che dimostrò una lettura non corretta dei dati delle ammissioni.

A seguire si evidenzierà la portata del paradosso attraverso un esempio similare costruito su dati reali di un'università italiana.

Si applicherà poi il paradosso a un caso di studio sui dati della disoccupazione con riferimento all'età ed al titolo di studio dei soggetti interessati.

Infine, preso atto che, nel corso di partite o incontri di alcuni sport, vengono raccolte moli di dati da analizzare, per stabilire in anticipo le possibilità di successo finale della squadra o dell'atleta saranno presentati due casi, Il primo caso, relativo al basket professionista americano, rappresenta un classico esempio del paradosso di Simpson. Il secondo caso, relativo agli incontri internazionali di tennis, rappresenta invece una situazione in cui il termine "paradosso di Simpson" viene usato con un'accezione diversa, e per questo è riportato in appendice.

CAPITOLO I

Profilo storico del Paradosso di Simpson

1.1 Definizione di paradosso

Per paradosso, dal greco *παρά* (*contro*) e *δόξα* (*opinione*), si intende un ragionamento che appare contrario al sistema di credenze comuni, principi o proposizioni scientifiche che si ritengono ben stabiliti, oppure un ragionamento che appare corretto, ma che porta ad una conclusione contraddittoria.

In filosofia e in economia il termine paradosso è usato spesso come sinonimo di antinomia. In matematica invece si distinguono i due termini: il paradosso consiste in una proposizione eventualmente dimostrata, logicamente coerente, ma lontana dall'intuizione; l'antinomia, invece, consiste in una vera e propria contraddizione logica.

Il paradosso è un potente stimolo alla riflessione che mette a nudo la debolezza della nostra capacità di discernimento ed i limiti di alcuni strumenti intellettuali a nostra disposizione che permettano di fare un ragionamento compiuto.

È stato così che paradossi basati su concetti semplici hanno spesso portato a grandi progressi intellettuali. Talvolta si è trattato di scoprire nuove regole matematiche o nuove leggi fisiche per rendere accettabili le conclusioni che all'inizio erano "apparentemente inaccettabili". Altre volte si sono individuati i sottili motivi per cui erano fallaci premesse o ragionamenti "apparentemente accettabili".

Secondo la definizione che ne dà Mark Sainsbury, il paradosso è “una conclusione apparentemente inaccettabile, che deriva da premesse apparentemente accettabili per mezzo di un ragionamento apparentemente accettabile”.

Sin dall'inizio della storia scritta, l'umanità si è sempre interessata ai paradossi, basti pensare a quelli enunciati da Zenone (la freccia scagliata dall’arco che non raggiungerà mai il bersaglio ed il pié veloce Achille che, essendo partito in ritardo, nella corsa non raggiungerà mai la tartaruga) spiegati un paio di millenni dopo, attraverso il calcolo infinitesimale.

1.2 I paradossi statistici

Nelle analisi statistiche, nell'ambito delle scienze sociali e mediche, ma non solo, è facile incorrere in errori di valutazione dei risultati ottenuti. Un esempio è quello sotto riportato.

I dati si riferiscono a come due ospedali (X e Y) operano su una certa malattia.

Il dato complessivo è che l'ospedale X ha il 55% di successi e quello Y, il 60%.

La prima conclusione è che per quella certa malattia converrebbe operarsi in Y.

Analizzando i dati scomposti, relativi ai due ospedali, risulta quanto segue:

- Per quella certa malattia, l’ospedale X opera sul 90% di casi gravi, ne risolve il 50% (45% del totale), mentre sui restanti 10%, lievi, ha il 100% di successo (10% del totale).
- Per l’ospedale Y il 40% sono casi lievi, ne risolve il 90%, (36% del totale), mentre sul 60% di casi gravi il successo è del 40% (24% del totale).

Da questi dati invece emerge che conviene sempre operarsi nell’ospedale X.

In pratica, la prima interpretazione dei dati era falsata da parametri non considerati.

1.3 Il precursore di Simpson

Nel corso di vari studi statistici, che prendevano in considerazione anche i dati provenienti dai sottogruppi che avevano contribuito alla formalizzazione del dato iniziale, emergeva un risultato in apparente contraddizione con il dato atteso.

1.3.1 Il contributo di George Udny Yule

Vari studiosi di statistica furono consci di tali problemi, fin dalla fine del secolo XIX. Ma è agli inizi del secolo ventesimo, con George Udny Yule, che il problema viene affrontato ed evidenziato nell'articolo *"Notes on the theory of association of attributes in Statistics"*, comparso in Biometrika (volume II, febbraio 1903). Ulteriori informazioni su tale studioso sono nella biografia riportata in appendice.

Per evidenziare il contributo di Yule, si riportano le conclusioni del capitolo 4, dell’articolo prima citato, intitolato: *“On the theory of complete independence of a series of Attributes”*, nel quale viene trattata la teoria dell’indipendenza completa dei dati di una serie di eventi.

Più in particolare Yule nota che “l’inadeguatezza del trattamento usuale d’indipendenza dei dati dipende dal fatto che si procede completamente a priori e generalmente ci si riferisce solamente a casi di probabilità artificiali. Il risultato è un'illusoria apparenza di semplicità". Quindi dimostra che "se un primo evento ha successo in a_1 e fallisce in b_1, un secondo evento ha successo in a_2 e fallisce in b_2 *e così via,* gli eventi combinati (successi e fallimenti) sono regolati dalla seguente espressione:

$$(a_1 + b_1)(a_2 + b_2) \dots (a_n + b_n)$$

I soli successi invece sono dati da:

$$a_1 \times a_2 \times \dots \times a_n$$

Quindi la probabilità di successo per tutti gli eventi considerati è:

$$\frac{a_1 a_2 \dots \dots \dots a_n}{(a_1 + b_1)(a_2 + b_2) \dots (a_n + b_n)}$$

Mentre la probabilità che il primo evento fallisca, mentre tutti i restanti abbiano successo è data da:

$$\frac{b_1 a_2 \dots a_n}{(a_1 + b_1)(a_2 + b_2) \dots (a_n + b_n)}$$

E così via per tutti gli altri casi. In definitiva, la probabilità degli eventi indipendenti combinati è il prodotto delle probabilità degli eventi separati".

Poi, sottolineando che nella realtà non si hanno, in generale, casi così semplici come quello (artificiale) appena trattato, osserva che la vera difficoltà sta nel capire se gli eventi, da analizzare, sono completamente indipendenti o lo sono solo in parte, nell'universo considerato.

Questa prima conclusione anticipa il successivo capitolo 5, dello studio in esame, intitolato significativamente: *"On the fallacies that may be caused by the mixing of distinct records independence of a series of Attributes"*, dove, con degli esempi, viene evidenziato l'errore che si introduce quando i dati considerati non sono indipendenti.

Di seguito si riporta l'esempio relativo ad un caso di ereditarietà, dove si considerano alcuni attributi che non sono ereditati né nella linea maschile, né in quella femminile, ma quando tali attributi si raggruppano emerge una considerevole ereditarietà apparente.

"Supponendo che un attributo sia presente nel 50% dei padri e dei figli, ma solo nel 10% delle madri e delle figlie, allora se non c'è ereditarietà in entrambe le linee dei discendenti, il quadro dei dati è approssimativamente il seguente.

Figura 1 - Tabella attributi padri - figli

Padri con attributo e figli con attributo	$[(0.5) \times (0.5)] \times 100 = 25$	25%
Padri con attributo e figli senza attributo	$[(0.5) \times (0.5)] \times 100 = 25$	25%
Padri senza attributo e figli con attributo	$[(0.5) \times (0.5)] \times 100 = 25$	25%
Padri senza attributo e figli senza attributo	$[(0.5) \times (0.5)] \times 100 = 25$	25%
Totale	100	100%

Figura 2 - Tabella attributi madri - figlie

Madri con attributo e figlie con attributo	$[(0.1) \times (0.1)] \times 100 = 1$	1%
Madri con attributo e figlie senza attributo	$[(0.1) \times (0.9)] \times 100 = 9$	9%
Madri senza attributo e figlie con attributo	$[(0.9) \times (0.1)] \times 100 = 9$	9%
Madri senza attributo e figlie senza attributo	$[(0.9) \times (0.9)] \times 100 = 81$	81%
Totale	100	100%

Se i dati di queste due tabelle si mescolano con le stesse proporzioni si ottiene:

Figura 3 - - Tabella attributi genitori - discendenti

Genitori con attributo e discendenti con attributo	$\frac{(25+1)}{2}$	13%
Genitori con attributo e discendenti senza attributo	$\frac{(25+9)}{2}$	17%
Genitori senza attributo e discendenti con attributo	$\frac{(25+9)}{2}$	17%
Genitori senza attributo e discendenti senza attributo	$\frac{(25+81)}{2}$	53%
Totale	100	100%

Quindi si rileva, ad esempio, che 13 discendenti su 30, di genitori con attributo ed in possesso loro stessi dell’attributo, corrispondono al 43,3% dei discendenti, ma erano solo il 30% dei discendenti in generale.

Questo esempio dimostra che si è ottenuta un’ereditarietà abbastanza ampia, ma falsa, dal semplice mescolamento dei dati delle due tabelle distinte”.

Quindi, dopo aver analizzato altre implicazioni derivanti dall'esempio, Yule conclude che un tale errore "potrebbe portare a risultati seriamente fuorvianti in parecchi casi dove avviene il mescolamento dei due sessi".

1.4 Il tutor di Simpson

All'Università di Cambridge, il professor Maurice Stevenson Bartlett assegna a Simpson la ricerca sulle interazioni e ne segue lo sviluppo.

1.4.1 Il contributo di Maurice Stevenson Bartlett

Nel 1946 il professor Maurice Bartlett fornisce a Simpson un suo articolo del 1935 (Contingency table interaction. *Journal of the Royal Statistical Society*, 2,248-252) e lo invita a lavorare sulle interazioni di secondo ordine delle variabili indicate nella tabella di contingenza. Il risultato di questo studio sarà l'articolo che dimostra l'esistenza del paradosso.

Notizie ulteriori sull'attività di Bartlett sono nella biografia riportata in appendice.

1.5 Altri studiosi citati da Simpson

Simpson nell'articolo utilizza e cita, oltre al documento di Bartlett, anche alcuni studi di R. A. Fisher, M. G. Kendall, H. W. Norton e G. W. Snedecor.

1.5.1 Il contributo di Ronald Aymler Fisher

Di Fisher viene citato uno studio del 1935 per confermare che la condizione di simmetria degli attributi, considerati nello studio, è soddisfatta da un rapporto di probabilità.

Fisher, la cui biografia è riportata in appendice, è considerato lo studioso che ha contribuito a far diventare la statistica una scienza moderna, in quanto ha sviluppato i concetti di riferimento della statistica matematica.

1.5.2 Il contributo di Maurice George Kendall

Di Kendall viene utilizzata l'opera *The advanced theory of statistics* allo scopo di verificare che in mancanza di interazioni di secondo ordine, gli attributi considerati sembrano essere indipendenti in tutta la popolazione. Anche di questo studioso è riportata in appendice la biografia.

1.6 Importanza del contributo di altri studiosi

Gli studiosi di statistica sopra citati hanno sicuramente influito in maniera diretta o indiretta alla riuscita del lavoro di Simpson.

In appendice sono state riportate le loro biografie per evidenziare che in molti casi hanno collaborato a ricerche comuni, in altri casi si sono scontrati per affermare le loro idee.

In ogni caso il risultato finale ha comportato, in generale, l'affermazione della statistica come supporto indispensabile ad altre discipline e, in particolare, un metodo che consentisse una lettura dei dati a disposizione più approfondita quando le prime conclusioni si presentavano incoerenti.

Tale metodo, che rientra nella definizione di paradosso, riportata nei paragrafi precedenti, fu sviluppato, con il supporto di alcuni esempi e la rappresentazione delle relazioni fra le variabili, dallo studioso di statistica britannico Edward Hugh Simpson nell'articolo "The interpretation of interaction in contingency tables" nel *Journal of the Royal Statistical Society* (1951).

Tale teoria fu poi chiamata "Paradosso di Simpson" da Colin Blyth nel 1972 (nell'articolo "On Simpson's Paradox and the sure-thing principle" pubblicato nel *Journal of the American Statistical Association, 67*(338), 364-366).

Bisogna però precisare che alcuni studiosi vi fanno riferimento chiamandolo "Effetto Yule-Simpson" oppure dandogli un nome impersonale, come "Paradosso dell'inversione", volendo rimarcare il fatto che Simpson non era stato il primo a scoprirlo.

1.7 Biografia di Edward Hugh Simpson

Edward Hugh Simpson (nato nel 1922) è uno statistico inglese molto conosciuto per aver descritto il paradosso che porta il suo nome. Dopo gli studi di matematica e statistica lavora come cripto analista (1942 – 1945).

Nel 1946, da studente laureato all'Università di Cambridge, incomincia ad occuparsi del problema dell'interazione dei dati, utilizzando l'articolo, prima indicato, del professor Maurice Bartlett ed a studiarne le implicazioni sotto la guida dello stesso Bartlett. Tale studio verrà poi pubblicato nel 1951, su richiesta di Bartlett che voleva citarlo in altri suoi studi.

Successivamente ha lavorato per il Ministero dell'Educazione britannico ed ha avuto vari incarichi nel settore educazione del Commonwealth.

1.8 Il paradosso di Simpson

Il paradosso di Simpson è, in statistica, la situazione in cui una relazione tra due fenomeni è apparentemente modificata o persino invertita a causa di altri fenomeni non presi in considerazione nell'analisi dei dati in possesso

È alla base di frequenti errori nelle analisi statistiche nell'ambito delle scienze sociali e mediche, ma anche di altre discipline.

Il paradosso di Simpson fu introdotto per la prima volta da Yule nel 1903 (vedi paragrafo 1.3.1) il quale evidenzia "the fallacies that may be caused by the mixing of distinct records'' (pages 132-133).

Nell'articolo intitolato: THE INTERPRETATION OF INTERACTION IN CONTINGENCY TABLES, pubblicato nel "Journal of the Royal Statistical Society" (Series B - Methodological - Vol. 13, No. 2 (1951), pp. 238-241), E. H. Simpson, senza citare lo studio di Yule, descrive il fenomeno attraverso l'interpretazione delle interazioni nelle tabelle di contingenza.

Nella sintesi dell'articolo, dichiara di prendere per buona la definizione, fornita da Bartlett, di interazione di secondo ordine in una tabella (2 x 2 x 2), ma "si propone di dimostrare, con un esempio, che la mancanza di questa interazione di secondo ordine non necessariamente giustifica la procedura meccanica di formazione delle tre tabelle 2 x 2 e la verifica di ciascuna di queste secondo i metodi standard".

Nel testo dell'articolo esamina prima il caso di completa indipendenza dei tre attributi ed utilizza "una tabella 2 x 2 x 2 costruita classificando gli ingressi in accordo con il loro possesso degli attributi A o Ā, B o $\overline{B}$, C o $\overline{C}$), dove come d'uso Ā indica "non-A" e così via e si indicano con n_{11}, n_{12}, n_{13}, n_{14} , n_{21}, n_{22}, n_{23} , n_{24} le probabilità che un ingresso cadrà in una delle otto classi così formate. Si avrà la tabella sotto riportata.

Figura 4 - Tabella attributi - probabilità

	CB	C$\overline{B}$	$\overline{C}$B	$\overline{C}$ $\overline{B}$
A	n_{11}	n_{12}	n_{13}	n_{14}
Ā	n_{21}	n_{22}	n_{23}	n_{23}

Ovviamente $n_{11} + n_{12} + n_{13} + n_{14} + n_{21} + n_{22} + n_{23} + n_{24}$ = 1.

L'estensione a questo caso dell'ipotesi d'indipendenza, che è comunemente applicato alle tabelle 2 x 2, vale a dire che la probabilità della classe AB è il prodotto delle probabilità delle classi A e B, è il punto di partenza dell'ipotesi di completa indipendenza".

Successivamente, attraverso l'esempio sotto riportato, verifica cosa succede in presenza o in assenza dell'interazione di secondo ordine.

Figura 5 - Tabella dati degli effetti di una terapia

	Maschi	Maschi	Femmine	Femmine
	Non Trattati	*Trattati*	*Non Trattate*	*Trattate*
Vivi	4/52	8/52	2/52	12/52
Morti	3/52	5/52	3/52	15/52

Da questi dati deduce che "c'è un'associazione positiva fra terapia e sopravvivenza sia tra i maschi sia tra le femmine, ma se i dati delle tabelle vengono combinati si ritrova nuovamente che non c'è associazione fra terapia e sopravvivenza nella popolazione complessiva", (vedi tabella sotto riportata).

Figura 6 - Tabella che evidenzia la non associazione fra terapia e sopravvivenza

	Non Trattati	*Trattati*
Vivi	6/52	20/52
Morti	6/52	20/52

A questo punto c'è da chiedersi qual è la corretta interpretazione di questi dati. La risposta che fornisce Simpson è la seguente: "la terapia difficilmente può essere eliminata come non valida quando dà dei benefici sia ai maschi sia alle femmine".

L'esposizione di Simpson ma soprattutto gli esempi, che evidenziavano la teoria del paradosso, furono subito considerati rivoluzionari in quanto fornivano il vero significato delle correlazioni e le facevano meglio comprendere.

1.9 Il divulgatore del paradosso

Dopo qualche anno, C. R. Blyth battezza questa teoria con il nome di Paradosso di Simpson (piuttosto che Paradosso di Yule) nel suo lavoro del 1972 "On Simpson's Paradox and the sure-thing principle", pubblicato nel *Journal of the American Statistical Association, 67*(338), 364-366 e da quel momento è conosciuta con tale nome.

Quindi la espone in un modo ancora più semplice, avvalendosi dell'esempio del dottore che utilizza una nuova terapia per alcuni suoi pazienti.

"Un dottore ha intenzione di provare una nuova terapia su un certo numero di pazienti, la maggior parte del luogo (gruppo C) e una piccola parte di Chicago (gruppo C'). Uno studioso di statistica gli consiglia di utilizzare una tabella di numeri casuali costruita nel modo sotto riportato.

- Paziente gruppo C - Nuova terapia - Probabilità 0,91.
- Paziente gruppo C - Vecchia terapia - Probabilità 0,09
- Paziente gruppo C' - Nuova terapia - Probabilità 0,01
- Paziente gruppo C - Vecchia terapia - Probabilità 0,99

Quando il dottore sottopone allo studioso di statistica la tabella complessiva dei risultati (vedi sotto), questi gli fa notare che la nuova terapia era evidentemente risultata molto dannosa e lo critica per aver insistito a provare su così tanti pazienti.

Figura 7 – Tabella complessiva dei risultati delle due terapie

Terapia	*Standard*	*Nuova*
Morti	5950	9005
Vivi	5050 (46%)	1095 (11%)

Il dottore replica che aveva continuato, perché convinto che la nuova terapia era ovviamente migliore, in quanto aveva quasi raddoppiato il tasso di sopravvivenza in entrambe le città, come evidenziato dalla tabella sotto riportata".

Figura 8 – Tabella dei risultati dei gruppi separati

	Pazienti C	**Pazienti C**	**Pazienti C'**	**Pazienti C'**
Terapia	*Standard*	*Nuova*	*Standard*	*Nuova*
Morti	950	9000	5000	5
Vivi	50 (5%)	1000 (10%)	5000 (50%)	95 (95%)

Dai risultati sui singoli gruppi emergeva che i pazienti, soggetti alla nuova terapia, avevano un tasso di sopravvivenza più alto rispetto a quelli sottoposti alla terapia standard.

Quando le variabili dei gruppi di pazienti invece erano combinate, era la terapia standard a fornire il tasso di sopravvivenza più alto.

Con tale esempio Blyth evidenzia il Paradosso di Simpson e ne fornisce la spiegazione in termini di probabilità.

"Dati gli eventi A, B, e C (e i complementi B'e C') è possibile avere:

$$P(A|B) < P(A|B')$$

e simultaneamente avere anche

$$P(A|BC) \geq P(A|B'C) \text{ e } P(A|BC') \geq P(A|B'C').$$

Se, dall'esempio sopra riportato, si pone A = Vivi, B = Nuova terapia e C = Pazienti locali, allora P( ) è la probabilita' di un paziente scelto a caso fra quelli inseriti nella tabella di fig. 8 e coincide con le proporzioni indicate nella tabella, ora prese sul totale della popolazione disponibile.

In tale ipotesi si ottiene:

$$P(A|B) = 0{,}11 \;<\; P(A|B') = 0{,}46$$

$$P(A|BC) = 0{,}10 \;>\; P(A|B'C) = 0{,}05$$

$$P(A|BC') = 0{,}95 \;>\; P(A|B'C') = 0{,}50$$

Il fatto inizialmente sorprendente è che una media di 0,10 e di 0,95 è molto piu' piccola di una media di 0,05 e 0,50 ma, se gli eventi B e C fossero indipendenti, per esempio nel caso in cui la proporzione di chi riceve la nuova terapia è la stessa per i pazienti del gruppo C e del gruppo C', allora il paradosso non si manifesterebbe".

1.10 L’illustrazione del paradosso

Nonostante lo sforzo di Blyth di semplificare il paradosso di Simpson, esso è di difficile comprensione da parte degli studenti che iniziano lo studio della statistica e in generale per chi non si occupa di statistica. Diventa quindi particolarmente sfidante spiegare il paradosso di Simpson fin dai corsi introduttivi di statistica.

A tale scopo tornano utili i tool grafici e le applicazioni software di recente sviluppo, una delle quali viene riportata nel paragrafo seguente.

È presa dallo studio intitolato: *An Applet for the Investigation of Simpson’s Paradox* di Kady Schneiter, Jürgen Symanzik (Utah State University), pubblicato nel *Journal of Statistics Education* Volume 21, Number 1 (2013).

1.10.1 Descrizione dell’applicazione grafica

Tale applicazione, denominata SP (in: *www.math.usu.edu/~schneit/CTIS/SP/*), attraverso la rappresentazione grafica di Baker-Kramer (BK plot), illustra, per un determinato set di dati, il paradosso di Simpson agli studenti dei corsi introduttivi di statistica.

L’applicazione è costituita da due componenti principali: una tabella ed un grafico, che sono costruiti dall’applicazione al caricamento dei dati.

Il caso ipotetico in esame confronta le percentuali di cani e gatti tenuti in casa dai loro proprietari.

Le colonne della tabella riportano il numero dei soggetti delle categorie da confrontare (in tal caso cani e gatti), il numero delle condizioni d'interesse e la percentuale di ogni categoria riferita alla condizione di interesse.

Tali numeri sono disposti su due righe e si riferiscono ai due livelli (piccolo e grande) della variabile di dimensione.

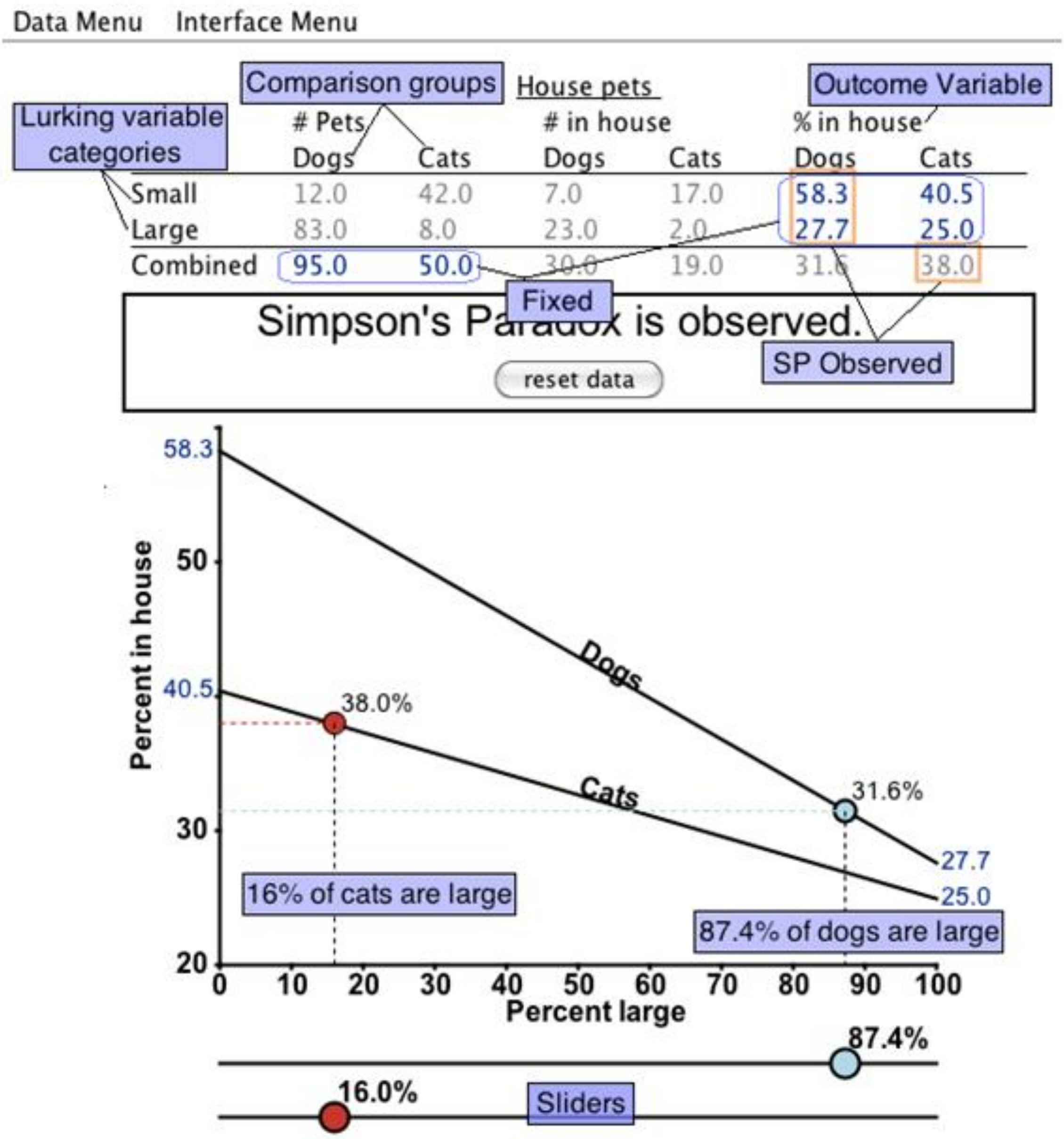
Quando l'utente manipola l'applicazione compare il numero totale delle osservazioni riferito alle due categorie con le relative percentuali di interesse.

In modo automatico si aggiornano tutte le altre celle, quelle stabilite dall'utente sono in blu, quelle dinamiche sono invece di colore grigio.

Nel grafico, sull'asse orizzontale sono rappresentate le percentuali delle osservazioni relative ad uno specifico livello della variabile, mentre l'asse verticale riporta le percentuali dei soggetti interessati. La relazione fra queste variabili è data dal grafico di ogni categoria interessata (cani e gatti).

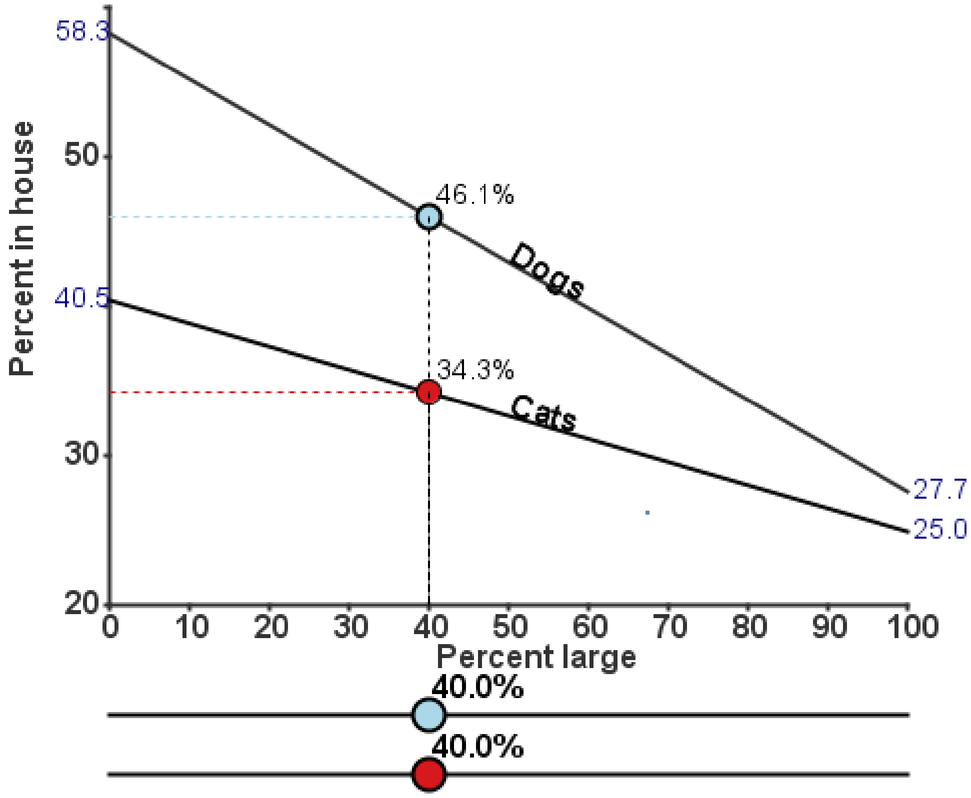
C'è un bottone su ogni linea che può essere spostato per evidenziare la percentuale richiesta. La variazione della percentuale si riflette sulla riga dei dati combinati della tabella.

Figura 9 - Applicazione SP - In alto la tabella dei dati, in basso il grafico relativo.



L’applicazione consente attraverso lo spostamento dei bottoni di posizionarsi sulla percentuale di interesse. Se per esempio entrambi i bottoni vengono posizionati sulla percentuale del 40%, il grafico mostra che sia i cani sia i gatti sono grandi e sono rispettivamente il 46,1% ed il 34,3% (vedi sotto).

Figura 10 - Grafico ottenuto dalle percentuali inserite (40%)



La tabella (non riportata) naturalmente si aggiorna in automatico, evidenziando i nuovi dati in funzione delle percentuali inserite.

Il paradosso di Simpson è osservato tutte le volte che le percentuali dei soggetti del gruppo, interessato al confronto, sono più alte per entrambi i livelli della variabile nascosta, laddove la percentuale "complessiva" dei soggetti interessati è più alta rispetto agli altri gruppi interessati al confronto.

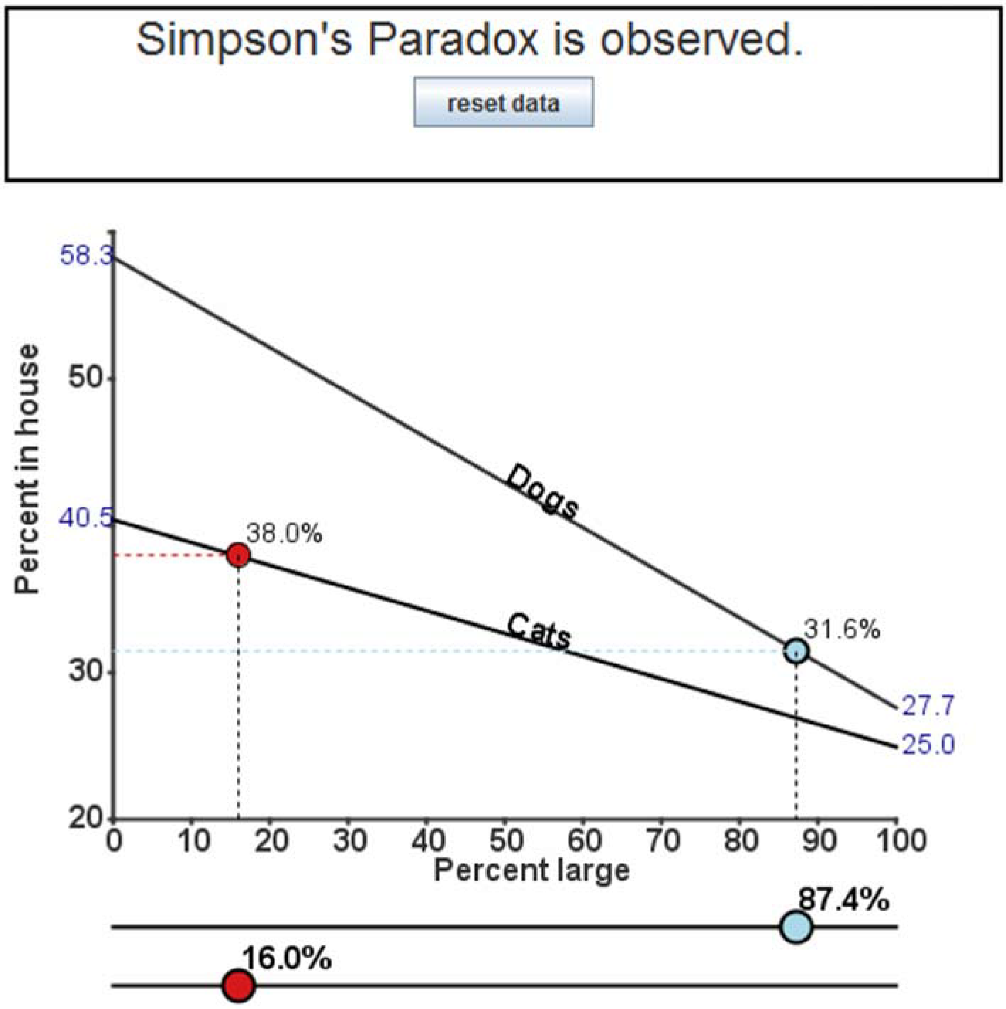
Ad esempio, ciò potrebbe accadere quando si rileva una percentuale più alta di cani domestici piccoli rispetto a quella dei gatti domestici piccoli ed una percentuale più alta di cani domestici grandi rispetto a quella dei gatti domestici grandi, ma una percentuale "complessiva" più alta di gatti domestici.

Nella tabella tali percentuali sono evidenziate con un riquadro rosso, nelle ultime due colonne, mentre sul grafico, il paradosso è evidenziato tutte le volte che al bottone della linea più in basso corrisponde una percentuale più alta sulla curva d’interesse rispetto al bottone della linea più in alto.

Figura 11 - Evidenza del paradosso su grafico e tabella

House pets

	# Pets		# in house		% in house	
	Dogs	Cats	Dogs	Cats	Dogs	Cats
Small	12.0	42.0	7.0	17.0	58.3	40.5
Large	83.0	8.0	23.0	2.0	27.7	25.0
Combined	95.0	50.0	30.0	19.0	31.6	38.0



Con riferimento agli animali domestici, la percentuale dei cani è più grande della percentuale dei gatti, per ogni specifica percentuale riferita agli animali grandi. Allora il paradosso di Simpson si verifica quando le percentuali di gatti e cani domestici sono scelte in modo tale da avere una percentuale più alta di gatti domestici rispetto a tutti i cani domestici.
L'applicazione non fornisce istruzioni dettagliate su cosa succede sulla tabella e sul grafico quando si presenta il paradosso di Simpson: ciò è stato fatto intenzionalmente per permettere all'istruttore di spiegarlo oppure agli studenti di scoprirlo.

CAPITOLO II

Casi di studio del Paradosso di Simpson

2.1 Caso di studio 1 – Ammissione all’Università

2.1.1 Universita’ di Berkeley

Un’importante verifica del paradosso di Simpson si è avuta nel 1973 all’Università di Berkeley quando vennero ammessi ai corsi di laurea il 44% dei ragazzi ed il 35% delle ragazze che avevano fatto domanda.

I dati si riferivano alla tabella reale sotto riportata.

Figura 12 - Ammissioni a Berkeley (1973)

	Ammessi	Rifiutati
Ragazzi	3738 (44.27%)	4704 (55.73%)
Ragazze	1494 (34.57%)	2827 (65.43%)

In definitiva le domande presentate erano state 12763 (8442 ragazzi e 4321 ragazze) con le percentuali di ammissioni prima indicate.

Le associazioni femministe non solo protestarono vivacemente, ma fecero causa all'Università, accusandola di discriminazione sessista.

L'Università si difese (con successo) dimostrando che il risultato cumulativo dipendeva dal paradosso di Simpson, invece se si guardava alle situazioni dei singoli dipartimenti, emergeva che le ammissioni erano sostanzialmente equilibrate, spesso anzi con una prevalenza a favore delle ragazze.

Il sostegno scientifico alla linea difensiva dell'amministrazione dell'Università di Berkeley è apparso nell'articolo scientifico di Bickel, Hammel e O'Connell, pubblicato sulla prestigiosa rivista Science nel 1975, dal titolo: *Sex bias in graduate admissions: Data from Berkeley.*

Gli autori, partendo dai dati reali della tabella di fig. 12, analizzarono i dati relativi ai 101 dipartimenti dell'università, poi si concentrarono sui dipartimenti (85) dove avevano presentato domanda sia i ragazzi sia le ragazze ed alla fine dello studio dimostrarono la tesi, sopra riportata, della non discriminazione.

Per far capire il concetto presentarono un esempio ipotetico, considerando solo due dipartimenti fittizi, con i dati della tabella sotto riportata.

Figura 13 - Dati ammissione di 2 dipartimenti fittizi

		Ammessi	Rifiutati
Ragazzi	Dipartimento MM	200 (50%)	200
Ragazze	Dipartimento MM	100 (50%)	100
Ragazzi	Dipartimento SW	50 (33%)	100
Ragazze	Dipartimento SW	150 (33%)	300
Ragazzi	Totali	250 (45%)	300
Ragazze	Totali	250 (38%)	400

Da essa rilevarono che al dipartimento MM avevano presentato domanda 400 ragazzi e 200 ragazze e gli ammessi erano stati in eguale proporzione, 200 e 100 rispettivamente, pari al 50%, mentre al dipartimento SW avevano presentato domanda 150 ragazzi e 450 ragazze e gli ammessi erano stati in eguale proporzione, 50 e 150 rispettivamente, pari al 33%.

Ma, quando si accorpavano i due dipartimenti, risultava che la percentuale dei ragazzi ammessi scendeva al 45% (250 su 550) e quella delle ragazze saliva al 38% (250 su 650), come dalle ultime due righe della tabella sopra riportata.

Inoltre, considerando il totale dei ragazzi, risultava che il 73% (400 su 550) aveva fatto domanda al dipartimento MM ed il 27% (150 su 550) al dipartimento SW, mentre, considerando il totale delle ragazze, risultava che il 69% (450 su 650) aveva fatto domanda al dipartimento SW ed il 31% (200 su 650) al dipartimento MM (vedi tabella sotto riportata).

Figura 14 - Dati iscrizione a 2 dipartimenti fittizi

		Richiesta Iscrizione	Ammessi su Totale Iscrizioni	Totale Iscrizioni
Ragazzi	Dipartimento MM	400 (73%)	200 (36%)	550
Ragazze	Dipartimento MM	150 (27%)	100 (18%)	550
Ragazzi	Dipartimento SW	200 (31%)	50 (8%)	650
Ragazze	Dipartimento SW	450 (69%)	150 (23%)	650

Se si considerava il numero degli ammessi, rispetto al totale delle iscrizioni ad ogni dipartimento, risultavano più penalizzati i gruppi richiedenti più numerosi (i ragazzi al dipartimento MM e le ragazze al dipartimento SW).

In definitiva il dato complessivo, che aveva originato il caso, si era dimostrato non attendibile a seguito delle verifiche a livello di dipartimento.

2.1.2 Universita' della Calabria

L'esempio seguente è stato sviluppato da Aljosa Volcic nello studio: *"La bella addormentata e altre illusioni probabilistiche"*, è chiaramente ispirato al famoso caso accademico (e giuridico) sopra riportato, ma è calato sul sistema reale di ammissioni ai corsi dell’Universita’ della Calabria.

Lo statuto dell'Università della Calabria (che è un'università residenziale) prevede che ogni corso di laurea stabilisca, anno per anno, il numero massimo degli iscritti al primo anno, tenendo conto del numero dei docenti, delle aule, dei laboratori, e così via. La selezione dei candidati avviene solitamente in base ad una graduatoria per la quale conta solo il voto di maturità, ma supponiamo che la Facoltà di Scienze decida di far fare ai candidati un esame attitudinale e di formare la graduatoria esclusivamente in base ad esso.

Per ogni corso di laurea viene dunque creata una commissione che procede all'esame degli aspiranti.

Presso la Facoltà di Scienze la graduatoria è effettivamente selettiva solamente per i due corsi di laurea più richiesti, quelli di Informatica e di Biologia, perché per tutti gli altri corsi di laurea le domande sono solitamente meno numerose dei posti disponibili.

Supponiamo che per Informatica i posti disponibili siano 100, mentre per Biologia siano 140. Per questi due corsi di laurea vengono presentate in tutto 400 domande, così suddivise, per corso di laurea e genere:

Figura 15 - Tabella domande di ammissione

	Informatica	Biologia
Maschi	150	50
Femmine	50	150
Totale	200	200

Questi dati sono realistici, perché le ragazze preferiscono Biologia, mentre i ragazzi preferiscono Informatica, inoltre Biologia offre più posti di informatica.

Le due commissioni predispongono dei test che non privilegiano né i ragazzi, né le ragazze.

A Informatica viene accettata la metà delle domande e, a seguito del test, viene ammessa la metà dei ragazzi e la metà delle ragazze, che hanno fatto domanda.

A Biologia viene accolto il 70% delle domande, anche qui il test seleziona con equità e passano la selezione il 70% dei ragazzi ed il 70% delle ragazze.

I risultati delle ammissioni sono riportati nella tabella seguente.

Figura 16 - Tabella degli ammessi

	Informatica	Biologia
Maschi	75	35
Femmine	25	105
Totale	100	140

.

Il giornale locale, riportando i dati dell'ammissione alla Facoltà di Scienze, constata che, dei 240 ammessi ai due corsi di laurea, 110 sono ragazzi e 130 sono ragazze ed osserva che tra i candidati c'erano 200 ragazzi ed altrettante ragazze, quindi il titolo che riassume l'articolo non può essere che il seguente: "L'Università della Calabria privilegia le ragazze?".

Naturalmente noi, avendo seguito passo passo come si è arrivati a quel dato globale, sappiamo che il titolo del giornale avanza un'ipotesi che non ha alcun fondamento. In questo, appunto, sta il paradosso.

Si noti che anche se i risultati dei test fossero stati leggermente più favorevoli ai ragazzi rispetto alle ragazze, questo fatto sarebbe rimasto "mascherato" nell'aggregazione dei dati.

Se ad esempio ci fossero stati 78 ragazzi ammessi ad Informatica e 41 a Biologia, le ragazze complessivamente ammesse sarebbero state comunque più numerose.

La spiegazione dell'inversione di tendenza nella fase di aggregazione dei dati sta nel fatto che i ragazzi hanno fatto prevalentemente domanda ad Informatica, che ha messo meno posti a disposizione e dove era quindi più difficile entrare.

Al contrario le ragazze hanno preferito scegliere il corso di laurea in Biologia, dove c'erano più posti a disposizione ed era quindi più facile entrare.

2.2 Caso di studio 2 – Tasso di disoccupazione

L''articolo, tratto dal The Wall Street Journal del 2 dicembre 2009, intitolato: *When Combined Data Reveal the Flaw of Averages* di Cari Tuna si occupa del paradosso di Simpson confrontando il tasso di disoccupazione della recessione del 2009 con quella del 1983 negli Stati Uniti.

“Si chiede subito se l’andamento economico del 2009 è peggiore della recessione dei primi anni ’80 e, prendendo in considerazione il tasso di disoccupazione, la prima risposta è no o al massimo non ancora, in quanto la percentuale dei disoccupati nel 2009 è attestata al 10,2%, in confronto al picco del 10,8% registrato alla fine del 1982.

Ma vista in altro modo la recessione del 2009 sembra peggiore, in quanto la percentuale di disoccupati tra i laureati ed i diplomati è più alta rispetto a quella registrata durante la recessione degli anni ’80.

Allora perché il tasso di disoccupazione complessivo è più basso nel 2009, ma è più alto considerando ciascun gruppo?

Questa anomalia è un esempio del paradosso di Simpson, un comune fenomeno di errore statistico che ha origine nella differente dimensione dei sottogruppi analizzati. Detto più semplicemente il paradosso di Simpson mostra che i dati aggregati possono sembrare in controtendenza rispetto ai risultati dei singoli gruppi”.

Il fenomeno appena evidenziato si spiega con l’esempio seguente (tratto da Wikipedia).

Si ipotizzi una situazione nella quale, a parità di età, tra i diplomati (maturità o laurea) la percentuale di disoccupati sia circa la metà di quella riferita a chi non ha conseguito il diploma. Si consideri pure il fatto che, per motivi storici, tra gli anziani, i diplomati (maturità o laurea) siano in numero molto minore e che, per motivi legati al mercato del lavoro, il tasso di disoccupazione tra i giovani sia più elevato che tra gli anziani.

Su tali ipotesi sono state costruite le tabelle sotto riportate.

Figura 17 - Tabella dati per eta' e titolo di studio

Lavoratori	senza diploma	con diploma	Totale
Giovani	20	80	100
Anziani	120	30	150
Totale	140	110	250

Figura 18 - Tabella tassi di disoccupazione

Tasso disoccupazione	senza diploma	con diploma
Giovani	30%	15%
Anziani	5%	3,33%

Da questa tabella risulta che, in entrambi i casi, la disoccupazione è circa doppia tra i non diplomati, rispetto ai diplomati. Si calcola quindi il numero di disoccupati utilizzando la tabella dati precedente.

Figura 19 - Tabella disoccupati per eta' e titolo di studio

Disoccupati	senza diploma	con diploma	Totale
Giovani	6	12	18
Anziani	6	1	7
Totale	12	13	25

Questi valori assoluti ci permettono ora di calcolare il tasso di disoccupazione per i non diplomati e per i diplomati, senza tenere conto dell'età (vedi sotto).

Figura 20 - Tabella disoccupati per titolo di studio

Lavoratori	Rapporto disoccupazione	Tasso disoccupazione
senza diploma	12/140	8,6%
con diploma	13/110	11,8%

Improvvisamente si scopre che tra i diplomati il tasso di disoccupazione invece che essere la meta e di un quarto maggiore che tra i non diplomati, esattamente il contrario di quello che si era ipotizzato.

Questo è un ulteriore esempio del paradosso di Simpson ed è dovuto al fatto che il tasso di disoccupazione è nettamente maggiore nel gruppo che ha una maggiore percentuale di diplomati: trascurare l'esistenza di due relazioni fondamentali (quella tra disoccupazione e età, nonché quella tra età e titolo di studio) fa giungere a conclusioni errate.

Mentre in questo caso preparato a tavolino la contraddizione è evidente, nelle analisi statistiche reali può capitare di non accorgersi delle relazioni implicite esistenti tra le variabili.

Se ci si limita ad analizzare dati aggregati, senza incrociarli con le variabili essenziali; la contraddizione non viene percepita e si traggono conclusioni completamente opposte alla vera distribuzione, con conseguenze potenzialmente molto gravi.

In situazioni meno estreme di quelle dell'esempio, le stesse cause del paradosso di Simpson possono portare a sovrastimare o sottostimare differenze tra gruppi, senza però capovolgere il "segno" della relazione.

I dati prodotti dal paradosso di Simpson chiaramente non sono sbagliati in sé, ma semplicemente devono essere letti in modo diverso.

Nel nostro caso, un lettore o analista superficiale sarebbe portato ad affermare che, tra persone con diploma, ci sono piu disoccupati, che tra persone senza diploma e, se usasse il concetto di causa-effetto in maniera superficiale, arriverebbe alla conclusione che, avere un diploma, è la causa di una maggiore disoccupazione.

Invece, volendo usare concetti di causa effetto, tenendo nella giusta considerazione tutti i dati a disposizione, si concluderebbe che:

- I giovani sono sei volte più soggetti alla disoccupazione rispetto agli anziani.
- In generale, avere un diploma, riduce il "rischio disoccupazione" alla metà.

2.3 Caso di studio 3 – Campionato di basket

Il paradosso di Simpson è stato evidenziato in molti studi riferiti a contesti scientifici e sociali, compreso il mondo dello sport.

In ambito sportivo è stato messo in luce nel baseball e nel cricket e, solo in casi limitati, nel basket.

Lo studio in esame: *Simpson's Paradox and Other Reversals in Basketball: Examples from 2011 NBA Playoffs* di Y. Zee Ma and Andrew M. Ma, pubblicato nel *International Journal of Sports Science and Engineering, Vol. 05 (2011) No. 03, pp. 145-154,* cerca di colmare tale lacuna, analizzando le partite della fase dei playoff del campionato di basket statunitense (NBA) del 2011.

Il paradosso è evidenziato quando vengono estrapolate le statistiche, relative ai punteggi ottenuti con tiri da 2 punti e con tiri da 3 punti, dal contesto generale delle tipologie di punteggio, che comprendono anche i tiri i tiri liberi.

Le ragioni, che hanno indotto a studiare i playoff NBA del 2011, sono legate, innanzi tutto, al fatto che, tale specifica manifestazione, ha avuto un andamento molto interessante a causa dei continui cambi della squadra di testa o per l'esito, non rispettato, delle previsioni degli esperti. Inoltre, in molte partite di questo torneo, la squadra che era in testa, per i primi tre quarti dell'incontro, poi l'ha perso nell'ultimo quarto o addirittura negli ultimi minuti per la rivalsa della squadra avversaria.

Dalla situazione appena descritta sono emersi alcuni esempi del paradosso di Simpson che andremo ad illustrare.

Una delle principali statistiche, che fornisce NBA, è quella relativa alla percentuale di tiro delle due squadre per ogni partita, ma essendo tre le tipologie di tiro che si possono trasformare in punteggio (tiri da 2 punti, da 3 punti e liberi), le tabelle ufficiali NBA non associano, ai punteggi, i dati sulle tipologie di tiro. Allora si è presa in considerazione la partita 5 del torneo, fra i San Antonio Spurs e i Memphis Grizzlies e sono state ricavate per ciascuna squadra le percentuali relative ai tiri, da 2 punti e da 3 punti, effettuati in quella partita.

Il risultato è quello della tabella seguente, che riporta altresì i risultati e le percentuali complessivi delle due tipologie di tiro.

Figura 21 - Statistiche sui tiri della partita 5 fra Spurs e Grizzlies

Squadra	Tiri da 2 punti		Tiri da 3 punti		Totale	
Grizzlies	49.4%	38/77	30.0%	3/10	47.1%	41/87
Spurs	50.8%	32/63	31.8%	7/22	45.9%	39/85

Da essa emerge che i Grizzlies hanno una percentuale complessiva di tiro più alta pur avendo la percentuale sui tiri da 3 punti più bassa. Allora, si sarebbe portati a pensare che i Grizzlies abbiano avuto una percentuale più alta sui tiri da 2 punti, per compensare il deficit sui tiri da 3 punti ed ottenere il risultato complessivo più alto. Invece, con sorpresa, dalla tabella si vede che i Grizzlies hanno anche la percentuale sui tiri da 2 punti più bassa rispetto agli Spurs.

Come si spiega, allora, che la squadra, che ha le percentuali più basse su entrambe le tipologie di tiro considerate, ha il risultato complessivo migliore?

Questo è un esempio ulteriore del paradosso di Simpson che si evidenzia quando due o più variabili di una categoria di dati sono combinate. Alcuni ricercatori evidenziano il paradosso attraverso l'individuazione delle categorie singole da una tabella complessiva, in tal caso si ottiene il paradosso di Simpson inverso, ma, indipendentemente dalla procedura usata (combinazione piuttosto che scomposizione dei dati), la natura del paradosso rimane invariata.

Prima di procedere con l'analisi della partita 5, fra i San Antonio Spurs e i Memphis Grizzlies, occorre precisare che i Grizzlies erano dati per sfavoriti da commentatori ed esperti, in quanto era la prima volta che arrivavano a tale fase del torneo (nelle altre tre partecipazioni non erano riusciti a vincere nemmeno una partita della fase preliminare), mentre i loro avversari avevano già vinto 4 volte il campionato ed avevano in squadra parecchi campioni.

In ogni caso i Grizzlies erano giunti alla partita 5 (sulle 7 previste) con il risultato di 3 a 1 sugli Spurs e, ad una manciata di secondi dalla fine di questa partita, conducevano 95-92. A questo punto un tiro da 3 di Ginobili degli Spurs, poi validato da 2 dalla prova video, portò il risultato a 95-94, seguirono due tiri liberi (andati a segno) per i Grizzlies e, quando mancavano 2 secondi alla fine del tempo regolamentare, il risultato era di 97-94. Alla ripresa del gioco, uno spettacolare tiro da 3 di Gary Neal, andato a segno prima del suono della sirena, portò il risultato in parità (97-97).

Questo tiro da 3, negli ultimi 2 secondi dei tempi regolamentari, modificò irreversibilmente l'andamento della partita. Nei tempi supplementari, infatti, gli Spurs superarono facilmente i Grizzlies, vincendo alla fine la partita 110-103.

Interessante notare che tale risultato effettivo è in controtendenza con il risultato statistico emerso dalla tabella precedente e può essere spiegato con il fatto che gli schemi di gioco dei Grizzlies prevedevano solo occasionalmente i tiri da 3, e quindi andavano in difficoltà con quelle squadre che invece li utilizzavano molto, come gli Spurs.

A tal proposito si mettono a confronto solo le statistiche sui tiri di Manu Ginobili degli Spurs con quelle di Tony Allen dei Grizzlies (vedi tabella sotto riportata).

Figura 22 - Statistiche dei tiri di 2 giocatori (partita 6)

Squadra	Tiri da 2 punti		Tiri da 3 punti		Totale	
Manu Ginobili	66.7%	4/6	25%	2/8	42.9%	6/14
Tony Allen	50.0%	4/8	0%	0/1	44.4%	4/9

I due giocatori hanno ottime percentuali sui tiri da 2, non altrettanto sui tiri da 3.

Ginobili pur avendo siglato, come prima visto, dei tiri da 3 decisivi, ne ha sbagliato 6 su 8 tentativi effettuati, mentre Allen ha sbagliato tutti e 3 le volte che ha provato. Per le due categorie considerate, Ginobili ha percentuali superiori rispetto ad Allen, ma considerando le due categorie combinate risultano inferiori. In poche parole emerge ancora il paradosso di Simpson.

I dati della tabella si riferiscono alla partita 6 nella quale Ginobili riesce a segnare due canestri da 3 punti, dei quali, l'ultimo dei due, allo scadere del terzo quarto, portando il risultato a 70-66 per gli Spurs. Quando tutti immaginavano una situazione simile alla partita 5, nell'ultimo quarto, invece, i Grizzlies riuscirono a rovesciare la tendenza vincendo la partita ed anche la serie.

In conclusione, attraverso il paradosso di Simpson nel basket, si è notato che le statistiche da analizzare, oltre che la percentuale di tiro (suddivise per tipologia) dovrebbero riguardare anche il numero dei tentativi di tiro, allo scopo di fornire agli allenatori le indicazioni su come impostare le partite, ad esempio dando l'incombenza a più giocatori. Inoltre bisognerebbe tenere conto (attraverso opportuni pesi) della diversa difficoltà che hanno i giocatori nell'effettuare le tre tipologie di tiro, allo scopo di mitigare i risultati contrastanti che emergono dai dati complessivi messi a confronto con quelli delle categorie singole.

CONCLUSIONI

Quanto riportato nei capitoli precedenti, porta a concludere che lo studio di Simpson ha posto le basi per la dimostrazione del paradosso.

E' comunque da evidenziare il contributo fondamentale di Yule che, circa cinquant'anni prima, aveva intuito che "la vera difficoltà sta nel capire se gli eventi, da analizzare, sono completamente indipendenti o lo sono solo in parte, nell'universo considerato". Non a caso alcuni studiosi parlano di paradosso di Yule ed altri di paradosso di Yule-Simpson.

Bisogna però riconoscere che la fortuna dello studio di Simpson è derivata dal lavoro di Blyth, che, oltre ad assegnargli il nome, ne ha favorito la diffusione.

I primi utilizzatori del lavoro di Blyth sono stati gli studiosi del caso dell'Università di Berkeley, che per la sua risonanza ha contribuito a far conoscere ancora di più il paradosso di Simpson.

L'avvento dei sistemi informatici ha alimentato ancor più il suo utilizzo, soprattutto nell'ambito dello sport, dove, per l'abbondanza dei dati a disposizione, l'analisi si può effettuare solo con mezzi automatici.

Infine, in anni più recenti, soprattutto nel mondo anglosassone, il paradosso di Simpson è utilizzato in modo più esteso per evidenziare preventivamente alcune situazioni ritenute anomale allo scopo di meglio comprenderne l'origine e la natura.

APPENDICE A

A.1 “Quasi-Simpson’s Paradox”: un neologismo per l’inversione negli incontri di tennis

Il termine "paradosso di Simpson" è divenuto sinonimo di "inversione" e quindi viene talvolta usato in modo non appropriato, come nella tesi di B. Wright intitolata "*Best Of N Contests: Implications Of Simpson's Paradox In Tennis”* pubblicata nel 2012 in *Electronic Theses, Treatises and Dissertations-The Graduate School* dell’Università Statale della Florida.

L'autore, a pagina 31, ammette che "Instances identified in this thesis may be considered as quasi-Simpson’s Paradox since a third component is required for direct application of the statistical abnormality". Poi, quando tale lavoro di tesi è sfociato in un articolo su rivista, gli stessi autori hanno incluso nel titolo il termine "quasi-Simpson’s Paradox", vedi: Wright, Benjamin; Rodenberg, Ryan M.; Sackmann, Jeff (2013) “*Incentives in Best of N Contests: Quasi-Simpson's Paradox in Tennis*” pubblicato in *International Journal of Performance Analysis in Sport*, Volume 13, Number 3, December 2013 , pp. 790-802(13).

Il contenuto della tesi sopra citata può essere riassunto come segue, partendo dalla considerazione che la statistica viene frequentemente applicata al mondo dello sport per la mole di dati che le varie discipline sono in grado di fornire. Non fa eccezione il mondo del tennis professionista maschile, al quale il lavoro di B. Wright è rivolto, con lo scopo iniziale di individuare la presenza del paradosso di Simpson nei tornei analizzati.

Una partita di tennis è considerata un esempio di paradosso quando un giocatore fa più punti del suo avversario ma alla fine perde la partita. I risultati delle partite analizzate, che manifestano tale paradosso, danno l'opportunità di azzardare delle strategie per trarre profitto dal sistema di punteggio utilizzato nel tennis.

Nella prima fase del torneo di Wimbledon del 2010 sono di fronte l'americano John Isner ed il francese Nicolas Mahut. L'incontro inizia alle sei di sera e, dopo circa tre ore, viene sospeso per oscurità quando i giocatori sono in parità di set (2-2). La partita riprende alle due del pomeriggio successivo e dopo circa sette ore di gioco viene nuovamente sospesa per l'oscurità. Al momento della seconda sospensione i giocatori erano ancora alla pari con 59 giochi a testa, un punteggio senza precedenti nel tennis.

La partita si concluse il terzo giorno, dopo un'altra ora di gioco e vinse John Isner con il punteggio di 6-4, 3-6, 6-7 (7-9), 7-6 (7-3), 70-68.

La precedente partita più lunga (durata 6 ore e 40 minuti) era stata superata di circa tre ore, già alla seconda sospensione. La partita fra Isner e Mahut, oltre a essere quella più lunga, ha battuto vari record fra i quali: il set più lungo, il numero più alto di giochi in un set, il numero più alto di giochi in una partita, il numero più alto di "aces" fatti da un singolo giocatore in una partita (Isner, 113), il numero più alto di "aces" in una partita (216).

L’ultimo record battuto è stato quello dei punti totali fatti da ciascun giocatore, vinto da Mahut con 502, rispetto ai 478 collezionati da Isner.
Il fatto che Isner abbia vinto l’incontro con un numero più basso di punti ha sollevato molte perplessità fra i tifosi di questo sport, ma ha anche avviato gli studi statistici atti a comprendere il fenomeno per il quale, in una partita, aveva vinto chi aveva fatto meno punti dell’avversario.
Una tale situazione nel mondo del tennis professionistico, spinge a ricercare ed identificare ulteriori “anormalità” statistiche, sfuggite alle ricerche precedenti.
Quest’analisi è mirata ad identificare il paradosso di Simpson a livello dei punti fatti durante le gare maschili effettuate dal 1991 al 2011 nei circuiti mondiali. Ciò potrà portare ad un miglioramento delle strategie nel mondo del tennis. Ad esempio, un giocatore può adottare uno sforzo minore durante le risposte ai servizi dell’avversario, se pensa di avere più possibilità di fare punti quando a servire è lui, oppure se è in ritardo di punteggio può evitare di fare sforzi straordinari, usando tale fase per riposare ed avere, poi, le energie necessarie a capovolgere l’esito della partita. Esempi del paradosso di Simpson possono produrre strategie, usate nel tennis professionale, atte, per esempio, a fare punti solo in certe fasi della partita, in ciò aiutati dal sistema del maggior punteggio da realizzare su un numero dispari di gare (Best of N Scoring), usato nel tennis, ma anche in altri sport.
I punti accumulati dai due contendenti della partita prima considerata, sono stati ricavati dai punteggi classici usati per assegnare il gioco (game) nel modo sotto indicato.

Figura 23 - Tabella conversione punteggi (game)

Per ogni gioco (game)	Punteggio classico	Punteggio all’unità	Punti tot.
1°vantaggio	15	1°punto	1
2°vantaggio	30	2°punto	2
3°vantaggio	40	3°punto	3
4°vantaggio	“game”	4°punto	4

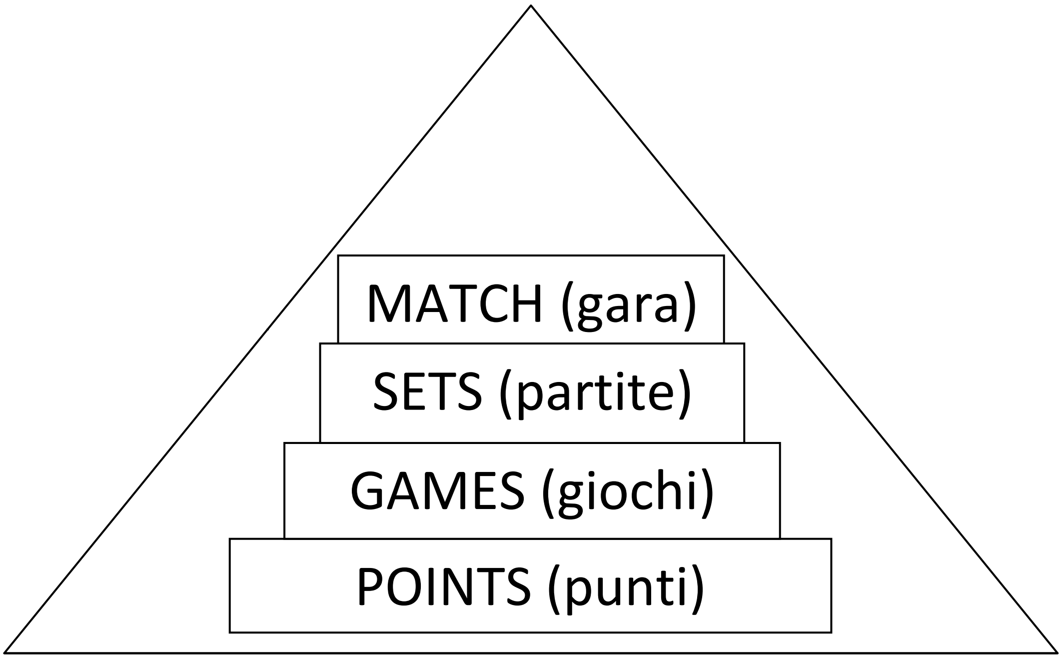
I punteggi classici (15, 30, 40 e “game”), come appena indicato, possono essere trasformati in punteggi all’unità (1°punto, 2°punto, 3°punto e 4°punto, rispettivamente), allo scopo di poter fare i confronti più facilmente. Si vince il gioco al 4° punto realizzato. Se i due giocatori sono a 3 punti, il 4° punto realizzato si chiama “vantaggio” e ne occorre un secondo per vincere il gioco.

Lo stesso sistema si utilizza per la partita (set): Vince chi per primo colleziona 6 giochi; se entrambi i giocatori sono a 5 giochi, vince chi ha un margine di altri 2 giochi. Infine la gara si vince al meglio delle 3 partite, al meglio delle 5 partite nei tornei professionali maschili (Best of N Scoring).

Definito cosa s’intende per punteggio all’unità, si stabilisce che il paradosso di Simpson si manifesta quando, in una partita, il vincitore totalizza meno punti dell’avversario. È quindi necessario conoscere i punti totali della partita ed i punti totali del vincitore per fare la valutazione.

Per comprendere i legami fra points, games, sets e match si allega la figura con la gerarchia del sistema di punteggio nel tennis (From ATP Official Website, 2012).

.



Per ogni incontro si calcolano i punti totali di ogni partita, i punti totali del vincitore e si calcola la relativa percentuale. Questi dati si mettono su un foglio elettronico, organizzato nel modo seguente:

- Colonna a: Punti totali dell’incontro
- Colonna b: Punti totali del vincitore
- Colonna c: percentuale dei punti del vincitore (Valore b / Valore a)
- Colonna d: Nome del vincitore
- Colonna e: Nome dell’avversario
- Colonna f: Risultato dell’incontro
- Colonna g: Data dell’incontro
- Colonna h: Nome del torneo
- Colonna i: Tipo di superficie dell’incontro

Le colonne a, b e c contengono i dati numerici dell'analisi, le altre colonne contengono le informazioni atte ad individuare i protagonisti, il risultato e la data dell'incontro considerato (vedi tabella semplificata con le prime 5 colonne ed alcune righe, sotto riportata).

Figura 24 - Tabella punteggi (all'unità) gare con paradosso

Punti totali	Punti vincitore	% vincitore	Nome vincitore	Nome avversario
980	478	48,775	*John Isner*	*Nicolas Mahut*
184	85	46,196	*Jim Courier*	*Brett Steven*
341	170	49,853	*Marat Safin*	*Andre Agassi*

Con riferimento all'incontro analizzato, la percentuale denota che si è di fronte ad un esempio di paradosso, in quanto il vincitore dell'incontro ha totalizzato meno punti del suo avversario. Si riportano a seguire due ulteriori esempi reali fra tutti quelli evidenziati dallo studio, i dati del quale sono stati ricavati dai siti ufficiali dei vari tornei che utilizzano la tecnologia IBM's PointStream. Tale tecnologia fornisce agli utenti, per ogni incontro, i punteggi dei game e dei set, inoltre tiene traccia dei punti totalizzati da ogni giocatore alla fine della partita, vale a dire i dati necessari allo studio in esame.

Sono state analizzate circa 55.776 gare individuali, giocate nei 21 anni considerati (1991-2011) e, con riferimento al sistema di punteggio all'unità, prima definito, sono stati individuati 2.793 gare dove si è manifestato il paradosso, corrispondenti al 5,01% delle gare complessive considerate.

Tale risultato ha due implicazioni pratiche. La prima è che i dirigenti di tale sport dovrebbero farsi carico dell'anormalità statistica secondo la quale il vincitore non necessariamente deve fare più punti dell'avversario e cercare di risolverla.
La seconda è che tale situazione potrebbe essere sfruttata dagli scommettitori per fare profitti, spingendo un giocatore a perdere i punti intenzionalmente.
In definitiva il paradosso manifestatosi sul 5% delle gare non deve essere tenuto in considerazione solo dagli studiosi di statistica, ma anche da chi governa il mondo del tennis per le possibili implicazioni sopra accennate.

APPENDICE B

B.1 Biografia di George Udny Yule

George Udny Yule (Beech Hill presso Haddington , Scozia, 18 febbraio 1871 – Cambridge , Inghilterra, 26 giugno 1951) è stato uno studioso di statistica britannico, pioniere della statistica moderna, che ha dato, tra l'altro, fondamentali contributi alla teoria della correlazione, della regressione lineare e delle serie storiche.

Di famiglia di tradizioni letterarie e amministrative, frequenta la scuola di Winchester e a 16 anni si iscrive ai corsi di ingegneria presso l'University College di Londra.

Inizia la sua attività di ricerca con studi sulle onde elettromagnetiche e, in occasione di un suo soggiorno a Bonn nel 1892, segue le lezioni di Heinrich Hertz.

Il risultato di tale attività sono sei articoli sulla teoria elettromagnetica.

Nel 1893 viene indirizzato da Karl Pearson alla statistica e scrive il primo articolo sull'argomento nel 1895 "On the Correlation of Total Pauperism with Proportion of Outrelief" dove introduce i coefficienti di correlazione per studiare le numerose tabelle di contingenza ante litteram. Da considerare che la parola tabella di contingenza viene introdotta da Karl Pearson solo nel 1904.

Fortemente influenzato da K. Pearson, al quale rimane molto legato, contribuisce dal 1897 al 1899 alla teoria della regressione e della correlazione.

Nel 1900 pubblica in Philosophical Transactions of the Royal Society of London l'articolo "On the association of attributes in statistics" dove introduce l'indice di associazione per tabelle di contingenza 2x2 detto Q di Yule.

Dal 1902 al 1909 è professore di statistica all'University College e dalle sue lezioni nasce il celebre "Introduction to the Theory of Statistics". Con "Notes on the theory of association of attributes in Statistics", comparso in Biometrika nel 1903, descrive quello che diventerà noto come paradosso di Simpson.

La sua massima produzione scientifica è negli anni 1920 durante i quali introduce il correlogramma e fonda la teoria delle serie storiche autoregressive. Nello stesso periodo (dal 1924 al 1926) è presidente delle Royal Statistical Society.

La morte di Karl Pearson nel 1936 lo affligge seriamente. Trova un proprio ultimo campo di ricerca svolgendo studi sulla frequenza della parole (soprattutto sostantivi) in vari testi, e pubblica nel 1944 "The Statistics of Literary Vocabulary".

Con i suoi numerosi contributi aprì nuove strade alla statistica moderna, spianando tra l'altro la strada a Ronald Fisher nell'ambito della distribuzione delle correlazioni parziali.

B.2 Biografia di Maurice Stevenson Bartlett

Maurice Stevenson Bartlett (Chiswick-Londra, 18 giugno 1910 – Exmouth-Devon, 8 gennaio 2002) è stato uno statistico britannico che ha dato importanti contributi all'analisi di dati con caratteristiche spazio-temporali.

I suoi genitori erano poveri, ma lui potè studiare con una borsa di studio per la Scuola Superiore Latymere, nel 1929, con una seconda borsa di studio per il Queens' College di Cambridge.

Nel 1933, Bartlett fu reclutato da Egon Pearson (1895-1980) che aveva fondato il nuovo dipartimento di Statistica presso l'University College di Londra. Qui Bartlett trovò un ambiente molto stimolante in quanto ebbe modo di lavorare con Jerzy Neyman(1894-1981), Ronald A. Fisher (1890-1962) e JBS Haldane(1892-1964).

Nel 1934 Bartlett si trasferì al centro di ricerche per l'agricoltura (ICI) dove ha fatto ricerche sulla genetica e la caratterizzazione dell' intelligenza.

Nel 1938 è passato all'Università di Cambridge. Quando è scoppiata la seconda guerra mondiale ha lavorato al Ministero dei Rifornimenti, dove ha fatto numerose ricerche insieme a Frank Anscombe (1918-2001), David Kendall (1918-2007) e Pat Moran (1917-1988).

Dopo la seconda guerra mondiale Bartlett ha fatto numerose visite negli Stati Uniti. Nel 1947 è diventato professore presso l'Università di Manchester, dove ha effettuato studi di epidemiologia. Nel 1960 è ritornato come professore di

Statistica all'University College di Londra. Gli ultimi otto anni della sua carriera accademica Bartlett li ha spesi come professore di biomatematica presso l'Università di Oxford.

Nel 1975 si è ritirato dagli incarichi accademici, ma è rimasto attivo nel campo della statistica.

B.3 Biografia di Ronald Aymler Fisher

Ronald Aylmer Fisher (Londra, 17 febbraio 1890 – Adelaide, 29 luglio 1962) è stato uno statistico, matematico e biologo britannico. Viene considerato colui che ha fatto della statistica una scienza moderna, in quanto ha fondato i concetti di riferimento della statistica matematica moderna.

Dal 1919 al 1933 è stato docente presso la stazione sperimentale di Rothamsted, poi, dal 1933 al 1943 a capo del dipartimento di eugenetica all'University College di Londra e infine, dal 1943 al 1957 titolare della cattedra di genetica a Cambridge.

Nel 1918 dimostrò matematicamente che i caratteri genetici (argomento di forte interesse per il neodarwinismo) seguono le regole indicate da Mendel e che si distribuiscono secondo una curva di Gauss.

È stato tra i primi a comprendere l'importanza del campionamento casuale per poter generalizzare i risultati, in opposizione ai campionamenti fatti secondo criteri vari di opportunità.

Nel 1925 perfezionò il metodo ideato da William Sealy Gosset (alias *Student*) per confrontare due medie, ideando il test "t di Student" attualmente usato e introducendo il concetto di gradi di libertà.

Importante sua innovazione è stata la cosiddetta *analisi della varianza*, ma è un suo allievo (George W. Snedecor) a utilizzare una distribuzione diversa da quella gaussiana, introducendo la variabile casuale F di Snedecor, dove la *F* è in onore al maestro (Fisher).

Con *The Design of Experiments* (1935) introdusse la regola che gli esperimenti devono essere programmati (*designed*, progettati) *prima* di essere effettuati, affinché i test statistici potessero avere una loro validità. In questo ambito coniò i concetti di ipotesi nulla (H_0) e ipotesi sperimentale (H_1).

Ha affermato (e si tratta di una grande novità in ambito del metodo scientifico) che nessuna ricerca sperimentale può dimostrare l'ipotesi sperimentale, ma solo "accettare" o "respingere" l'ipotesi nulla, anche se effettuare tanti esperimenti in cui si rigetta l'ipotesi nulla aumenta la credibilità che l'ipotesi sperimentale sia vera.

Nel 1930 propose la *Teoria genetica della selezione naturale* (The genetical theory of natural selection) nella quale studiava in maniera innovativa diversi concetti nel campo dell'evoluzione.

Nonostante la sua abbondante produzione scientifica, è stato presidente della Royal Statistical Society, primo presidente della International Biometric

Society, presidente della Société de Biométrie e presidente dell'Istituto Internazionale di Statistica (IIS).

Il fatto che, in seguito alle divergenze con il proprio maestro Karl Pearson, divenne direttore della stazione agraria di Rothamsted, alimenta tuttora la leggenda che la statistica moderna e la sua metodologia sia nata in ambito agrario, mentre in realtà Fisher come i suoi predecessori si forma nell'ambito della biometria, psicologia sperimentale o eugenetica. Infatti dopo aver diretto per 14 anni la stazione sperimentale gli venne assegnata la cattedra di eugenetica fondata da Galton e appartenuta a Pearson, e successivamente la Cattedra di Genetica all'Università di Cambridge.

Nel 1936 introdusse con *The use of multiple measurements in taxonomic problems* l'analisi discriminante (nella fattispecie quella lineare).

B.4 Biografia di Maurice George Kendall

Maurice George Kendall (Kettering, 6 settembre 1907 – Redhill, 29 marzo 1983) è stato uno statistico inglese.

Trasferitosi da bambino a Derby, si interessa, durante le scuole superiori, inizialmente alle lingue per poi passare alla matematica. Grazie alle sue capacità viene ammesso al St John's College a Cambridge, dove conclude in matematica. Viene in seguito assunto al ministero per l'agricoltura, dove viene a contatto con la statistica e produce lavori di tale qualità da diventare nel 1934 membro della Royal Statistical Society (della quale sarà presidente dal 1960 al 1962). Nel 1945 la RSS gli attribuisce la "Guy Medal in Silver", nel 1968 la "Guy Medal in Gold".

Con la sua ampia ed eccellente produzione scientifica contribuì tra l'altro alle teorie sulle serie storiche e la correlazione dei ranghi.

Nel 1980 le Nazioni Unite lo premiano con la medaglia per la pace per i suoi studi sulla fertilità.

Opere

- *The advanced theory of statistics*: primo volume nel 1943, il secondo volume nel 1946
- *Rank correlation methods, 1948*
- *An Introduction to the Theory of Statistics* (coautore George Udny Yule), 1968
- *Studies in the history of statistics and probability* (coautore Egon Pearson), 1969

BIBLIOGRAFIA / SITOGRAFIA

- Abbagnano N. *(1993) Dizionario di filosofia* Definizione di Paradosso TEA.

- Appleton D. R., French J. M., Vanderpump M. P. J. (November 1996). Ignoring a Covariate: An Example of Simpson's Paradox. *The American Statistician, Vol. 50, No. 4, pp. 340-341*
  http://www.jstor.org/stable/2684931

- Bartlett M. S. Biografia *Wikipedia*
  http://en.wikipedia.org/wiki/M._S._Bartlett
  Consultato il 1/1/2014

- Bickel P. J., Hammel E. A., O'Connell J. W. (February 1975).Sex Bias in Graduate Admissions: Data from Berkeley. *Science, Vol. 187*
  http://www.unc.edu/~nielsen/soci708/cdocs/Berkeley_admissions_bias.pdf

- Blyth C. R. (June 1972). On Simpson's Paradox and the Sure-Thing Principle. *Journal of the American Statistical Association, Volume 67, Number 338*
  http://www.stat.cmu.edu/~fienberg/Statistics36-756/Blyth-JASA-1972.pdf

- Fisher R. A. Biografia *Wikipedia*
  http://it.wikipedia.org/wiki/Ronald_Fisher
  Consultato il 1/1/2014

- Gelman A. (December 2009). Simpson's Paradox not always such a paradox. *Statistical Modeling, Causal Inference and Social Science*
  http://andrewgelman.com/2009/12/03/simpsons_parado/
  Consultato il 26/12/2013

- Kendall M. G..Biografia *Wikipedia*
  http://it.wikipedia.org/wiki/Maurice_George_Kendall
  Consultato il 1/1/2014

- Ma Y. Z., Ma A. M. (August 2011). Simpson's Paradox and Other Reversals in Basketball: Examples from 2011 NBA Playoffs. *International Journal of Sports Science and Engineering Vol. 05 (2011) No. 03, pp. 145-154*
  http://www.worldacademicunion.com/journal/SSCI/SSCIvol05no03paper03.pdf

- Malinas G, Bigelow J. (August 2009). Simpson’s Paradox. *Stanford Encyclopedia of Philosophy, First Published Feb 2,2004, Revised Aug 6,2009* http://plato.stanford.edu/entries/paradox-simpson/
- Paradosso di Simpson-Esempio tasso di disoccupazione. *Wikipedia* http://it.wikipedia.org/wiki/Paradosso_di_Simpson Consultato il 26/12/2013
- Paradosso definizione. *Wikipedia* http://it.wikipedia.org/wiki/Paradosso Consultato il 26/12/2013
- Paradosso di Zenone. *Wikilibri* http://www.wikilibri.it/pdf/Paradossi.pdf Consultato il 26/12/2013
- Sainsbury M. *Wikilibri* Definizione di Paradosso http://www.wikilibri.it/pdf/Paradossi.pdf Consultato il 26/12/2013
- Schneiter K., Symanzik J.(2013). An Applet for the Investigation of Simpson’s Paradox. *Utah State University*. *Journal of Statistics Education Volume 21, Number 1* www.amstat.org/publications/jse/v21n1/schneiter.pdf
- Simpson’s paradox. *Wikipedia* http://en.wikipedia.org/wiki/Simpson's_paradox Consultato il 26/12/2013
- Simpson E. H.(May 1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society, Series B 13: 238–241* http://www.epidemiology.ch/history/PDF%20bg/Simpson%20EH%201951%20the%20interpretation%20of%20interaction.pdf
- Tuna C. (December 2009). When Combined Data Reveal the Flaw of Averages. *The Wall Street Journal* http://online.wsj.com/news/articles/SB125970744553071829?mg=reno64-wsj&url=http%3A%2F%2Fonline.wsj.com%2Farticle%2FSB125970744553071829.html
- Volcic A. (Novembre 2009). La Bella Addormentata e altre Illusioni Probabilistiche. *Presentato alla Conferenza della Mathesis di Firenze* http://web.math.unifi.it/users/mathesis/conferenze/files-presentazioni/0910/Volcic.pdf

- Wagner C. H. (February 1982). Simpson's Paradox in Real Life. *The American Statistician* 36 (1): 46–48
  http://www.jstor.org/discover/10.2307/2684093?uid=3738296&uid=2134&uid=2&uid=70&uid=4&sid=21103363223863

- Wright B. (September 2012). Best of N Contests: Implications of Simpson's Paradox in Tennis. *The Florida State University*. *The Graduate School*. *Electronic Theses, Treatises and Dissertations, Paper 5279* http://diginole.lib.fsu.edu/etd/5279/

- Wright B., Rodenberg, R. M., Sackmann, J. (2013). Incentives in Best of N Contests: Quasi-Simpson's Paradox in Tennis. *International Journal of Performance Analysis in Sport, Volume 13, Number 3, December 2013 , pp. 790-802(13).*

- Yule G. U. (February 1903). Notes on the theory of association of attributes in Statistics. *Biometrika Vol. II*
  http://www.epidemiology.ch/history/PDF%20bg/Yule%20U%201903%20notes%20on%20the%20theory%20of%20assoc%20of%20attrib%20in%20stats.pdf

- Yule G. U. Biografia *Wikipedia*
  http://it.wikipedia.org/wiki/George_Udny_Yule
  Consultato il 1/1/2014