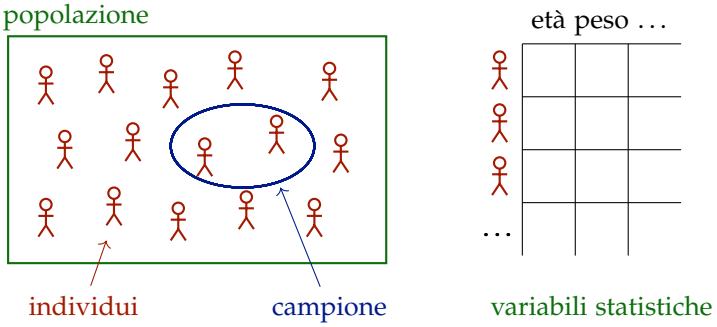
§16.2 Popolazione, individui, campione

Si usa la statistica in una situazione di questo genere. C'è una popolazione, cioè un insieme di individui, ciascuno dei quali ha delle caratteristiche dette variabili statistiche. Si osserva un campione, che è un sottoinsieme della popolazione, preso a caso, e da questa osservazione si vuole dedurre qualcosa a proposito dell'intera popolazione. Le parole preso a caso sono il cuore del discorso: la statistica è la disciplina che permette di controllare, e addirittura sfruttare, il caso.



Gli individui non sono necessariamente esseri umani, e neppure necessariamente oggetti concreti. Se studiamo le conseguenze degli incidenti stradali, la popolazione potrebbe essere l'insieme degli incidenti stradali avvenuti in un certo anno in un certo paese; se studiamo le conseguenze di una malattia, la popolazione potrebbe essere l'insieme di tutti i casi di quella malattia: sia quelli presenti, che quelli passati e futuri. La popolazione è l'insieme di tutti gli oggetti omologhi che stiamo considerando.

A ciascun individuo della popolazione sono associate delle caratteristiche: se sono esseri umani, per esempio, l'età, il peso, il sesso, lo stato malato o non malato di influenza, il colore degli occhi, il nome, il cognome. Queste caratteristiche sono le variabili statistiche. Ogni individuo ha un valore per ciascuna variabile.

Infine, un campione è un sottoinsieme casuale degli individui. Questo è il concetto più difficile da definire, perché qui occorrerebbe dire cosa sia il caso. Trascurando questa questione filosofica, resta il problema concreto di come prendere un campione casuale: problema che può essere ostico, su cui torneremo.

Lo scopo della statistica è questo: vogliamo capire qualcosa a proposito di una popolazione, ma non possiamo esaminare l'intera popolazione, perché è troppo vasta o addirittura è un ente puramente concettuale (tutti i casi concepibili di una malattia non si possono concretamente esaminare). Abbiamo accesso soltanto a un campione, e vogliamo estendere quello che vediamo sul campione all'intera popolazione, in maniera ragionevole.

Prendere un campione, però, non vuol dire sceglierlo: se gli individui sono scelti con un criterio, da quelli non si può dedurre nulla sull'intera popolazione. Se scegliessimo come campione degli studenti di un corso tutti quelli seduti al primo banco, ne concluderemmo che sono tutti diligentissimi. Il modo sensato di prendere un campione rappresentativo della popolazione è a caso. Se vogliamo sapere la proporzione di uomini e donne fra gli studenti, possiamo prenderne dieci a caso: è ragionevole che

la proporzione nel campione sia simile a quella dell'intera popolazione; se invece scegliessimo soltanto gente con la barba, chiaramente non lo sarebbe.

La statistica è, in questo senso, teoria della probabilità applicata: parte da un campione casuale e, sfruttando il fatto che è casuale, estende le conclusioni delle osservazioni fatte sul campione all'intera popolazione, misurando quanto questa estensione sia ragionevole. Se, su un campione, un vaccino sembra efficace, la statistica permette di decidere, tenendo conto di quanto sembra efficace e di quanto è numeroso il campione, se la conclusione è estendibile all'intera popolazione, oppure se potrebbe essere un effetto del caso. Tutta la ricerca medica è fatta così.

Nota (piano del corso)
La statistica si divide in due rami:

- la statistica descrittiva, cioè i metodi per descrivere una situazione: tabelle, diagrammi, misure riassuntive;
- la statistica inferenziale, che si occupa di trarre delle conclusioni: dal campione, dedurre se una cosa è plausibile oppure no, basandosi su un criterio scientifico.

Cominceremo dalla statistica descrittiva, per poi passare alla statistica inferenziale.

§16.3 Tipi di variabili statistiche

categoriche	
nominali	es. provincia di nascita, stato di occupazione
ordinali	es. titolo di studio
binarie	due valori, es. sano/malato
numeriche	
discrete	numeri interi, es. numero di fratelli
continue	valori in $\mathbb{R}$, es. peso, statura

Le variabili numeriche sono quelle espresse da un numero: età, peso. Le variabili categoriche sono quelle che non sono espresse da un numero: sesso, colore degli occhi, stato di occupazione.

Le variabili numeriche si distinguono in discrete, che assumono valori interi (non si possono avere due fratelli e mezzo), e continue, che assumono valori reali, su una scala infinitamente divisibile (in teoria il peso si può misurare al milligrammo). Alle volte si può scegliere: l'età in anni compiuti è discreta, l'età misurata esattamente dalla nascita è continua.

Le variabili categoriche si possono suddividere ulteriormente. Le binarie possono assumere due valori (maschio/femmina, sano/malato). Le nominali possono assumere più di due valori, ma senza un ordine (stato di occupazione: disoccupato, in cerca di lavoro, occupato, in pensione; provincia di nascita). Le ordinali sono quelle in cui le categorie hanno un ordine naturale — per esempio il titolo di studio: licenza elementare, licenza media, &c.

Vedremo metodi diversi per trattare le variabili categoriche e quelle numeriche: per esempio con le variabili numeriche si possono calcolare medie e mediane, cosa che con le variabili categoriche non si può fare (qual è lo stato di occupazione medio?).

Importante: variabile numerica significa variabile *quantitativa*, non semplicemente espressa per mezzo di numeri. Il comune di nascita, per esempio, si potrebbe esprimere indicando il CAP, ma questo non lo renderebbe una variabile numerica, perché non ha senso, per esempio, sommare fra loro dei CAP. Alle volte, la distinzione fra una variabile meramente ordinale e una numerica discreta può essere opinabile: per esempio, i valori di soddisfazione "per niente soddisfatto", "poco soddisfatto", "così così", "soddisfatto", "entusiasta" non sono numerici, ma potrebbero anche essere tradotti nei numeri 1–5, e non sarebbe del tutto fuori questione farne la media. Personalmente, non tratterei allo stesso modo i valori "guarigione", "miglioramento", "condizione stazionaria", "peggioramento", "decesso".

§16.4 Frequenze e tabelle

Esempio (indice di massa corporea)
L'indice di massa corporea (IMC, in inglese BMI) è

$$\text{IMC} = \frac{\text{peso}}{\text{statura}^2}$$

misurato in kg/m^2; è pensato come misura dello stato di nutrizione. I dati che seguono sono un estratto, creato a scopo didattico, dal *Framingham Heart Study*, uno dei primi grandi studi sulle malattie cardiovascolari.

IMC	frequenza	freq. relativa
$< 20\,\text{kg}/\text{m}^2$	210	4.9%
20–$25\,\text{kg}/\text{m}^2$	1744	40.3%
25–$30\,\text{kg}/\text{m}^2$	1809	41.8%
30–$35\,\text{kg}/\text{m}^2$	437	10.1%
$\geqslant 35\,\text{kg}/\text{m}^2$	129	3.0%
totale	4329	100.1%

L'IMC è una variabile numerica, ma in questa tabella è stato trasformato in una variabile categorica (ordinale), suddividendo i valori in cinque classi. È un procedimento standard, che si usa per esempio per applicare a variabili numeriche metodi pensati per le variabili categoriche. La convenzione generalmente usata è che le classi includano l'estremo sinistro ed escludano il destro: 20–25 vuol dire da 20 incluso a meno di 25.

Il numero di individui in ciascuna categoria si chiama frequenza (o frequenza assoluta) di quella categoria nel campione. La somma di tutte le frequenze, qui 4329, è la numerosità del campione (o della popolazione, se la tabella riguarda l'intera popolazione). La frazione di individui che ricade in ciascuna categoria, di solito espressa in percentuale, si chiama frequenza relativa:

$$\text{frequenza relativa} = \frac{\text{frequenza}}{\text{numerosità}} \qquad \text{per esempio} \qquad \frac{210}{4329} = 4.9\%$$

Nota importante
Quando si calcolano le frequenze relative conviene fare la somma per controllare che faccia 100%. Può venire 99.8% o 100.1%, come nella tabella qui sopra, per gli arrotondamenti, e va benissimo; se viene molto diverso, c'è un errore.

§16.5 Tabelle di contingenza

Le variabili statistiche non si considerano quasi mai da sole: quello che si cerca, di solito, sono le correlazioni fra variabili diverse. Aggiungiamo alla tabella dell'IMC il numero di individui che, nei 24 anni di osservazione successivi, sono stati ricoverati per un infarto.

	esito		
IMC	infarto	non infarto	totale
$< 20\,\mathrm{kg/m^2}$	8	202	210
$20–25\,\mathrm{kg/m^2}$	118	1626	1744
$25–30\,\mathrm{kg/m^2}$	195	1614	1809
$30–35\,\mathrm{kg/m^2}$	46	391	437
$\geqslant 35\,\mathrm{kg/m^2}$	18	111	129
totale	385	3944	4329

Questa è una tabella di contingenza (o tabella a doppia entrata) e presenta due variabili statistiche: l'IMC e l'esito, una variabile categorica binaria (infarto/non infarto). Ci permette di studiare (vedremo in seguito con quali metodi) la correlazione fra le due variabili: per esempio, se un IMC alto è un fattore di rischio per l'infarto. A occhio sembra di sì: la percentuale di ricoverati per infarto va dal 3.8% (8/210) nella prima classe al 14.0% (18/129) nell'ultima.

Il corpo della tabella contiene la distribuzione congiunta delle due variabili: per ogni combinazione di valori, quanti individui la presentano. I totali di riga e di colonna, ai margini, sono le distribuzioni marginali: le distribuzioni di ciascuna variabile presa da sola. Conoscendo soltanto le distribuzioni marginali, in generale, non si può ricostruire la distribuzione congiunta.

§16.6 Studi clinici randomizzati

Uno dei primi studi condotti con un metodo statistico adeguato, e che oggi si studia nei libri di testo, è quello che nel 1948 ha stabilito l'efficacia della streptomicina nel trattamento della tubercolosi polmonare (Medical Research Council, *British Medical Journal*, 30 ottobre 1948).

valutazione radiologica a sei mesi	streptomicina		controllo	
netto miglioramento	28	51%	4	8%
miglioramento moderato o lieve	10	18%	13	25%
nessun cambiamento sostanziale	2	4%	3	6%
peggioramento moderato o lieve	5	9%	12	23%
netto peggioramento	6	11%	6	11%
decesso	4	7%	14	27%
totale	55	100%	52	100%

Questo studio è famoso per essere stato il primo randomizzato: il campione è stato diviso a caso in due gruppi; il gruppo di trattamento è stato curato con la streptomicina, il gruppo di controllo no. Modernamente si dà spesso un placebo, ma non sempre è possibile; in questo studio il gruppo di controllo è stato trattato come si trattava la tubercolosi all'epoca — vale a dire in nessun modo, perché la tubercolosi non aveva una cura: i pazienti venivano semplicemente tenuti a riposo. Gli autori discutono il problema etico di lasciare alcuni pazienti senza una vera cura: in questo caso, la questione è risolta dal fatto che stavano usando tutta la streptomicina disponibile, e il caso è un metodo di allocazione, se non altro, equo.

I pazienti sono stati trattati con streptomicina per sei mesi, e le loro radiografie sono state valutate da radiologi a cui non era stato detto in che gruppo stessero. La tabella — di contingenza — mostra gli esiti nei due gruppi. In situazioni come questa si parla di variabile indipendente (qui il gruppo) e di variabile dipendente (qui l'esito): la prima è quella che ci aspettiamo possa influenzare la seconda. Qui, l'efficacia del farmaco è evidente a occhio. Quando l'effetto è più debole, però, occorre avere un criterio preciso. I test di significatività (capitolo 21) — ce ne sono vari, che si applicano in base al tipo di esperimento — forniranno precisamente questo criterio.

Perché assegnare i pazienti ai gruppi a caso, e non, per esempio, formando coppie di pazienti il più possibile simili per età, sesso, gravità della malattia? Perché non sappiamo quali variabili influenzino l'esito. Se si divide un gruppo di persone in due a caso, qualunque variabile possa giocare un ruolo (foss'anche i gusti musicali) tende a distribuirsi in modo simile nei due gruppi: il caso uniforma i gruppi rispetto non soltanto alle variabili che conosciamo, ma anche a quelle che non conosciamo. In secondo luogo, sull'effetto del caso abbiamo tutta una serie di metodi matematici, che permettono di distinguerlo dai fenomeni veri e propri — paradossalmente, l'effetto di una scelta deliberata, ancorché ragionevole, sarebbe più difficile da controllare.

§16.7 Variabili confondenti

Come si fa a stabilire se un vaccino funziona? Questa è la storia di uno dei più celebri esperimenti di questo genere: il vaccino contro la poliomielite — una malattia che colpisce i bambini e può causare paralisi permanenti — di Jonas Salk. A posteriori, sappiamo che ha funzionato: grazie alla vaccinazione, questa malattia, un tempo endemica, è stata eliminata dai paesi occidentali, e ne rimangono pochi casi al mondo.

Seguiamo il ragionamento con cui si arriva a un buon esperimento.

1. Si potrebbero vaccinare tutti i bambini, diciamo di una regione, e poi guardare se i casi di poliomielite calano. Non funziona: le malattie infettive vanno a ondate, e i casi possono calare o aumentare da un anno all'altro per conto loro.

2. Potremmo vaccinare soltanto metà dei bambini, per poi confrontare i casi fra vaccinati e non vaccinati: le ondate epidemiche colpiscono tutti allo stesso modo. Non è fattibile: non si può obbligare nessuno a sottomettersi all'esperimento, tanto più che un vaccino si somministra a persone sane: c'è il problema del consenso.

3. Potremmo vaccinare chi acconsente e usare gli altri come controllo. Non funziona: può darsi che l'esposizione alla malattia sia correlata con la volontà di vaccinarsi. Se chi acconsente è più esposto, il vaccino sembrerà meno efficace di quanto è; se è meno esposto, più efficace.

4. Prima si chiede il consenso, poi, fra chi ha acconsentito, si decide a caso chi vaccinare e chi mettere nel gruppo di controllo. Così i due gruppi sono uniformi. Resta un problema: chi sa di essere vaccinato potrebbe comportarsi diversamente (prendere più rischi) da chi non lo è.

5. Allora la scelta si fa a caso, e al gruppo di controllo si fa un'iniezione di soluzione fisiologica, indistinguibile dal vaccino: né il paziente né il medico sanno in che gruppo è il paziente. Si parla di esperimento in doppio cieco.

Ancora non abbiamo un esperimento perfetto, principalmente perché la statistica riguarda soltanto chi acconsente, e si deve supporre che le conclusioni ricavate da questo gruppo si applichino a tutta la popolazione. Questa è un'ipotesi ragionevole, e, d'altro canto, senza ledere la libertà dei pazienti non c'è altra soluzione. Più in generale, se il trattamento non si esaurisce in una singola iniezione, qualcuno potrebbe interromperlo a metà. La scelta più sensata è classificarlo comunque secondo il gruppo a cui era stato assegnato all'inizio (analisi intention to treat), ma sfavorisce il trattamento, e non c'è una soluzione perfetta. Infine, non in tutti gli studi è possibile il doppio cieco: chi viene operato sa di essere stato operato.

Esempio (il vaccino di Salk)
Nel 1954 fu sperimentato su larga scala il vaccino contro la poliomielite sviluppato da Jonas Salk. La tabella seguente riporta i risultati di due studi condotti in parallelo (T. Francis, *American Journal of Public Health*, 1955); il tasso è il numero di casi di poliomielite ogni 100 000 bambini.

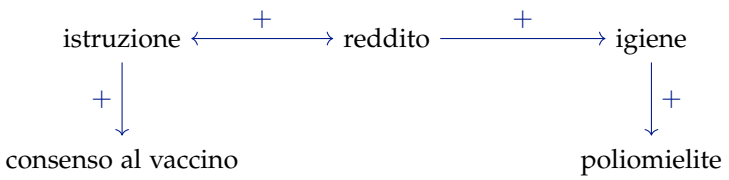
studio randomizzato in doppio cieco			studio NFIP		
	numero	tasso		numero	tasso
trattamento	200 000	28	classe seconda (vaccino)	225 000	25
controllo	200 000	71	classi prima e terza (controllo)	725 000	54
non consenso	350 000	46	classe seconda (non consenso)	125 000	44

Si sperimentarono due diversi protocolli. Nello studio di destra (della National Foundation for Infantile Paralysis), si è proceduto per classi scolastiche: si sono vaccinati i bambini di seconda, e i bambini di prima e di terza hanno fatto da controllo. Questo è un difetto, perché le malattie infettive sono correlate con il tempo e con lo spazio: se il contagio si diffonde in certe classi seconde, sembrerà che il vaccino non funzioni; se si diffonde in certe classi prime o terze, sembrerà che funzioni — e il vaccino non c'entra niente. Fattori come questi si chiamano variabili confondenti: variabili correlate sia con il trattamento sia con l'esito che si studia. C'è un secondo difetto: il gruppo di trattamento è costituito solo da bambini i cui genitori hanno dato il consenso, mentre questa selezione non è stata operata sul gruppo di controllo.

Lo studio di sinistra è quello fatto bene, secondo lo schema del punto 5: controllato (c'è un gruppo di controllo), randomizzato (l'assegnazione ai gruppi è casuale) e in doppio cieco (le fiale di vaccino e di soluzione fisiologica erano contrassegnate da un codice, e solo chi aveva progettato l'esperimento sapeva che cosa contenesse ciascuna). La tabella riporta tre gruppi: trattamento, controllo e non consenso — il consenso è stato acquisito prima della divisione in trattamento e controllo (né può essere altrimenti, stante che ambo i gruppi hanno ricevuto un'iniezione).

Il vaccino funziona: 28 casi ogni 100 000 bambini nel gruppo di trattamento, contro 71 nel gruppo di controllo. Ma si guardi il gruppo del non consenso: non è stato vaccinato, eppure ha solo 46 casi ogni 100 000, molti più dei vaccinati, ma molti meno del gruppo di controllo! Chi decide di vaccinarsi ha una probabilità di ammalarsi diversa da chi decide di non farlo. Questo mostra che era necessario assegnare i gruppi dopo avere chiesto il consenso: se si fosse confrontato chi decide di vaccinarsi con chi non si vaccina, il confronto sarebbe stato falsato.

Perché chi non voleva vaccinarsi aveva una probabilità di ammalarsi inferiore rispetto al gruppo di controllo? La ragione è una catena di correlazioni:



All'epoca, la poliomielite era endemica: quasi tutti la prendevano, ma da piccolissimi, e, presa nei primissimi anni di età, dà luogo per lo più a sintomi lievi e nessuna conseguenza. I casi paralitici si sviluppano, invece, in chi la contrae da più grande. Questo ha una conseguenza apparentemente paradossale: i casi più gravi hanno una correlazione positiva con il livello di igiene — infatti, i bambini che vivevano in condizioni igieniche migliori erano meno esposti al contagio, quindi contraevano la malattia da più grandi, ossia nel periodo pericoloso. L'igiene aveva, poi, una correlazione positiva col reddito delle famiglie, e il reddito è correlato col grado di istruzione. Infine, l'istruzione ha — o almeno aveva — una correlazione positiva con la propensione a far vaccinare i propri figli. Morale: le variabili confondenti possono fare giri strani; uno studio randomizzato in doppio cieco serve proprio a difendersi da quelle che non si conoscono.

§16.8 Studi osservazionali

Gli studi visti finora sono studi di intervento: si interviene sugli individui. A volte non è possibile. La tabella seguente viene da uno studio sulla possibile correlazione fra uso della pillola anticoncezionale e pressione arteriosa (pressione sistolica, in millimetri di mercurio; *Walnut Creek Contraceptive Drug Study*).

	età 17–24		età 25–34		età 35–44		età 45–58	
pressione	no	sì	no	sì	no	sì	no	sì
sotto 90	–	1	1	–	1	1	1	–
90–95	1	–	1	–	2	1	1	1
95–100	3	1	5	4	5	4	4	2
100–105	10	6	11	5	9	5	6	4
105–110	11	9	11	10	11	7	7	7
110–115	15	12	17	15	15	12	11	10
115–120	20	16	18	17	16	14	12	9
120–125	13	14	11	13	9	11	9	8
125–130	10	14	9	12	10	11	11	11
130–135	8	12	7	10	8	10	10	9
135–140	4	6	4	5	5	7	8	8
140–145	3	4	2	4	4	6	7	9
145–150	2	2	2	2	2	5	7	9
150–155	–	1	1	1	1	3	2	4
155–160	–	–	–	1	1	1	1	3
160 e oltre	–	–	–	–	1	2	2	5
totale (%)	100	98	100	99	100	100	99	99
numero	1206	1024	3040	1747	3494	1028	2172	437

I valori sono percentuali di ciascuna colonna; "sì" e "no" indicano l'uso della pillola; – vuol dire trascurabile. Le classi includono l'estremo sinistro e non il destro; per gli arrotondamenti le colonne possono non sommare a 100.

Uno studio randomizzato qui non si può fare: né il placebo né il doppio cieco sono proponibili. Si tratta di uno studio osservazionale: si sono semplicemente raccolti i dati di donne che usano, o non usano, la pillola, per scelta propria. In questa situazione non ci si può fidare del caso per neutralizzare le variabili confondenti. La più evidente è l'età: la proporzione di utilizzatrici fra le donne giovani è molto diversa da quella fra le donne più anziane, e la pressione è correlata con l'età. Se si confrontasse la pressione media di tutte le utilizzatrici con quella di tutte le non utilizzatrici, le seconde

risulterebbero in media più anziane, e il confronto sarebbe falsato. Per questo la tabella è suddivisa anche per età: presenta tre variabili statistiche (pressione, età, uso della pillola). La speranza è che, facendo il confronto colonna per colonna, l'effetto della distribuzione di età si annulli. In uno studio osservazionale non c'è altro da fare che pensare a tutte le possibili variabili confondenti e tenerne conto in qualche modo.

§16.9 Il paradosso di Simpson

Le statistiche vengono spesso usate per argomentare circa lo stato delle cose e le decisioni da prendere. C'è però sempre un passaggio logico nascosto: quello che porta dal dato numerico alla conclusione sostanziale. Per sapersi difendere, intellettualmente, dall'errore comune, o dalla malizia del venditore di aspirapolveri di turno, occorre rendersi conto della necessità di capire cosa i dati grezzi dicano davvero, e poi avere i dati, e anche saperlo fare.

Esempio
In una grande università, l'accesso ai corsi di laurea è regolato da una prova d'ingresso. Il tasso di successo è

$$\text{tasso di successo} = \frac{\text{numero ammessi}}{\text{numero iscritti alla prova}}$$

Due burocrati, per sostenere agende politiche diverse, fanno le seguenti affermazioni:

Burocrate A: "Il tasso di successo delle donne è minore di quello degli uomini!"

Burocrate B: "In tutti i corsi, il tasso di successo delle donne è maggiore!"

Possono avere ragione tutti e due? Ecco i dati — la vicenda è avvenuta davvero; i numeri sono semplificati; i due burocrati sono personaggi di fantasia.

	donne			uomini		
corso	iscritte	ammesse	tasso di s.	iscritti	ammessi	tasso di s.
A	100	80	80%	800	600	75%
B	200	150	75%	500	250	50%
C	400	80	20%	400	60	15%
D	500	50	10%	200	10	5%
totale	1200	360	30%	1900	920	48%

In ciascun corso le donne hanno un tasso di successo maggiore degli uomini; eppure, nel complesso, gli uomini hanno un tasso maggiore: 48% contro 30%. Hanno ragione tutti e due. Com'è possibile? Il corso A è chiaramente il più accessibile (entrano quasi tutti) e il corso D il più esclusivo (non entra quasi nessuno); la maggioranza degli uomini si è iscritta al corso A, la maggioranza delle donne al corso D. Le donne hanno un tasso di successo complessivo minore perché, in maggioranza, hanno tentato la prova di ingresso più difficile.

Questo fenomeno si chiama paradosso di Simpson. Ci dice che nessuno dei due burocrati ha capito che cosa stesse succedendo: la questione vera è che le iscrizioni non sono distribuite uniformemente fra i corsi, e che la difficoltà delle prove non è uniforme. Entrambi dicevano il vero, ed entrambi non avevano capito niente. La statistica è una cosa complicata, e conviene studiarla nei dettagli.