38 I Test Statistici

IL PRIMO COMPITO DELLA STATISTICA INFERENZIALE:

DISTINGUERE FRA
– VARIAZIONI SIGNIFICATIVE
– E NON SIGNIFICATIVE:
PER ESEMPIO PER DISTINGUERE
PLAUSIBILI EFFETTI CASUALI IN UN TRIAL CLINICO DA
SPERABILI EFFETTI CAUSALI DEI FARMACI.

38.1 Test simile a quello della signora del tè

Un'esperienza statistica che si avvicina a riprodurre il famoso esperimento della signora del tè è questo il seguente.

Sa in soggetto riconoscere le parole cinesi da quelle giapponesi, quando traslitterate come nell'uso comune, omettendo gli accenti?

Lo scrivente ha posto a ChatGPT una richiesta di questo tipo per il cinese, e poi similmente per il giapponese:

Mi scrivi 4 nomi di animali in cinese, ma non banali come cane e gatto, bensì di più raro uso?

Le parole sono state private degli accenti e numerate da 1 a 8.
Poi ha fatto questa richiesta:

Mi scrivi una permutazione casuale dei numeri da 1 a 8?

Con la permutazione ha potuto mescolare le 8 parole straniere così:

1: rakuda

2: bianfu

3: kongkue

4: komori

5: kujaku

6: xiniu

7: kujira

8: haitun

Se il soggetto trova le 4 parole cinesi (e quindi le 4 parole giapponesi) il test respinge quest'ipotesi nulla:

H_0: il soggetto non sa distinguere le parole giapponesi da quelle cinesi traslitterate come nell'uso comune, omettendo gli accenti.

Le parole cinesi erano:

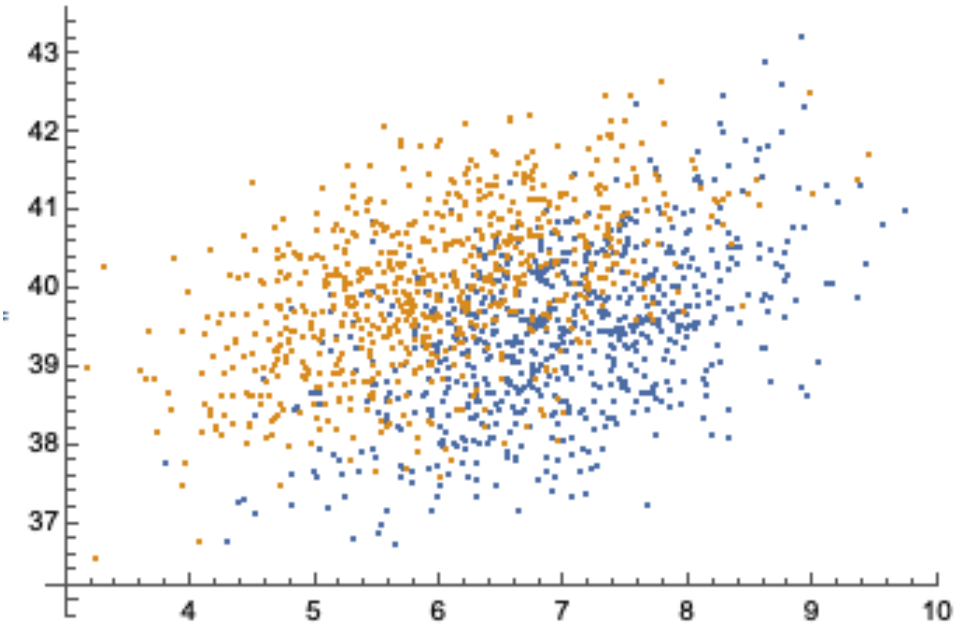
3: Kǒngquè - Pavone
6: Xīniú - Rinoceronte
2: Biānfú - Pipistrello
8: Hǎitún - Delfino

Le parole giapponesi erano:

5: Kujaku - Pavone
7: Kujira - Balena
1: Rakuda - Cammello
4: Kōmori - Pipistrello

La permutazione dei numeri da 1 a 8 era:
7, 3, 1, 8, 5, 2, 6, 4.

38.2 Introduzione ai Test Statistici



Arruoliamo molte persone con BMI di 40.0 e a metà di loro diamo una terapia A e all'altra metà una terapia B.
Alla fine gli chiediamo un giudizio di soddisfazione da 3 a 10, anche coi decimali, salvo il 10, valore massimo, e misuriamo i BMI.
Lo scatterplot rappresenta

in arancio l'esito della terapia A,

in blu quello della terapia B.

La terapia B ha dato mediamente più soddisfazione della A? Parrebbe di sì, ma non sarà forse una fluttuazione casuale?
La terapia B ha diminuito il BMI? Sembrerebbe un po' di sì, ma qua siamo molto più incerti. E la A? Ancora più incerti.

Ci servono FORMULE:

entrano numeri, esce sì/no.

Una tale formula è un *test statistico*.

38.3 Introduzione tecnica ai Test Statistici

Dalle determinazioni x_1,...,x_n di un campione aleatorio X_1,...,X_n tratto da una v.a. X di densità nota salvo un suo parametro a, vogliamo rispondere con "sì" o "no" , con "ragionevole certezza statistica" ovviamente, a una domanda sul parametro incognito.

In questa trattazione elementare a in generale è un numero incognito, come p che poi è la media μ in $B(1,p)$.
(Ma lasciamo aperta la possibilità, di livello superiore, che a sia una coppia di numeri come (μ,σ^2)).

La domanda potrebbe essere per esempio
$\mu > 0$? per una v.a. $N(\mu,\sigma^2)$, oppure
$p = \frac{1}{2}$? per una v.a. $B(1,p)$ che nella realtà sensibile può significare per esempio la regolarità di una moneta; da decidersi con ragionevolezza statistica dai risultati di molti lanci.

In Farmacia:
La glicemia è diminuita? (Cioè, così tanto da far ragionevolmente ipotizzare un effetto causale piuttosto che casuale).
Possiamo dire che il farmaco fa più del placebo?

Il test statistico si preordina – prima di avere i dati in mano ovvero prima di fare un esperimento nella realtà sensibile – formulando un'ipotesi statistica, indicata con H o H_0, *ipotesi nulla*, e una ipotesi *alternativa*, indicata con A o rispettivamente H_1, per esempio

$$H : p = \frac{1}{2} \qquad A : p \neq \frac{1}{2}\,.$$

Anticipiamo che l'ipotesi nulla H va identificata in generale col caso che si spera che non sia. ("Vogliamo A!").

Un esempio minimo potrebbe essere così: lanceremo 5 volte la moneta e rifiuteremo l'ipotesi di regolarità se viene testa 0 o 5 volte,

perché se la moneta è regolare quei risultati hanno complessivamente probabilità 1/16, un po' pochino.
In realtà la statistica usuale viene fatta "a 1/20", cioè al 5 ovvero 95%; ma largheggiando possiamo fare Statistica "al 90%" e allora respingeremmo l'ipotesi della regolarità con 0 o 5 teste su 5 lanci.

L'insieme $\{0, 5\}$ è la *regione critica* ovvero di rigetto dell'ipotesi (nulla). La regione critica di solito viene espressa come un sottoinsieme D di $\mathbb{R}$ in cui una certa funzione del campione aleatorio può cadere (e allora rifiutiamo H) o non cadere (non rifiutiamo H). In questi termini, potremmo porre la regione critica $\{0, 5\}$ e verificare se vi cade $X_1 + ... + X_5$ o più usualmente porre $D := \{0, 1\}$ e verificare se vi cade $\bar{X}_5$. In casi più significativi di questo microscopico esempio la regione critica di solito ha una forma del tipo $x > x_0$ (test unilatero) oppure $x < x_1 \vee x > x_2$ (test bilatero).
Vediamo un altro esempio. Sia I la variabile aleatoria che è la glicemia (iniziale) di un soggetto qualunque di un campione di 20 soggetti (persone iperglicemiche) e F la glicemia (finale) dopo la somministrazione di un certo farmaco. Ci potrebbe interessare se mediamente la glicemia diminuisce con quel farmaco cioè se la media (parametro incognito) della variabile aleatoria $X := I - F$ è > 0. (Iniziale grande, finale piccola).
Si formula l'ipotesi nulla: il farmaco non riduce la glicemia:

$$H : \mu \leq 0 \qquad \text{(finale grande come o più dell'iniziale)}$$

essendo μ la media di X.
In realtà spero che riduca la glicemia: ipotesi alternativa:

$$A : \mu > 0 \qquad \text{(finale più piccola dell'iniziale)}$$

Misuriamo la glicemia nei 20 soggetti (campione, o più precisamente determinazione $i_1, \ldots, i_{20}$ di un campione aleatorio $I_1, \ldots, I_{20}$). Diamo ai 20 soggetti il farmaco. Misuriamo di nuovo la glicemia dei 20 soggetti ottenendo così 20 determinazioni $f_1, \ldots, f_{20}$ della variabile aleatoria F. Facciamo 20 sottrazioni ottenendo 20 determinazioni

$x_1, \ldots, x_{20}$ della variabile aleatoria X, differenza *prima-dopo* ovvero iniziale-finale. Della variabile aleatoria X vogliamo sapere se la media (speranza matematica) è > 0, come speriamo, oppure no.

L'idea ingenua è fare la media aritmetica dei 20 numeri e concludere che se è > 0 il farmaco ha diminuito la glicemia.
Se fosse così in questo e analoghi casi, la statistica inferenziale non servirebbe, ma non è così: quella verifica non dice di per sè sostanzialmente nulla perché non distingue l'effetto del farmaco dalle inevitabili fluttuazioni casuali di X, che, non per niente, è da considerarsi una variabile *aleatoria*. (Non possiamo certo aspettarci un effetto *deterministico* del farmaco, che *sempre* riduca la glicemia).
È invece necessario applicare un opportuno test statistico, cioè di fatto applicheremo una non banale formula che ci potrà dire, nel caso che la media aritmetica degli x_i sia > 0, che quell'effetto con ragionevole plausibilità non è casuale. Se invece la media aritmetica viene negatitiva l'esperimento è andato male e la statistica non ci aiuta ulteriormente. Alla fine rifiutiamo o non rifiutiamo l'ipotesi nulla. Speriamo di rifiutarla.
Alcuni dicono "accettare" l'ipotesi nulla ma il modo corretto di vedere le cose è "non rifiutarla". Non abbiamo dimostrato che è vera: semplicemente non siamo riusciti a dimostrarla *verosimilmente falsa*.

Media degli X_i, differenza prima-dopo	
Statistica ingenua:	
La glicemia mediamente non è diminuita; il farmaco non funziona	La glicemia è diminuita; il farmaco funziona
0	
Statistica inferenziale:	
Non rifiutiamo l'ipotesi che la glicemia sia uguale o aumentata ovvero $\mu \leq 0$	Rifiutiamo l'ipotesi che la glicemia sia uguale o aumentata ovvero $\mu \leq 0$
0 *soglia*	È plausibile che il farmaco funzioni.

Insomma X deve essere mediamente ben > 0, non solo > 0, per escludere con ragionevole verosimiglianza la fluttuazione casuale.

Quanto > 0, lo dicono apposite formule che vedremo, che fanno uso dei quantili, di Student in questo caso.

I quantili si trovano e soprattutto si trovavano su tavole numeriche, che permettevano di ottenere la "ragionevolezza al 95%" o ancor meglio al 99%, e anche con altri *livelli di confidenza* tipici.
Diciamo subito che nella pratica si trova scritto indifferentemente "al 99%" o "all'1%" con lo stesso significato, e similmente con 95 e 5, eccetera: purtroppo c'è un'ambiguità terminologica.
Da adesso in questo paragrafo facciamo rifermento al valore "piccolo": non 0.95 ma 0.05, non 0.99 ma 0.01, eccetera.

38.4 Il test della signora del tè

Da Wikipedia, l'enciclopedia libera, un esperimento degli anni '30 del secolo scorso, alla base della moderna Statistica:

> The experiment is the original exposition of Fisher's notion of a null hypothesis (...) The lady in question (Muriel Bristol) claimed to be able to tell whether the tea or the milk was added first to a cup. Fisher proposed to give her eight cups, four of each variety, in random order (...) The experiment provides a subject with 8 randomly ordered cups of tea – 4 prepared by first pouring the tea, then adding milk, 4 prepared by first pouring the milk, then adding the tea. The subject has to select 4 cups prepared by one method (...) The null hypothesis is that the subject has no ability to distinguish the teas (...) The critical region for rejection of the null of no ability to distinguish was the single case of 4 successes of 4 possible, based on the conventional probability criterion $< 5\%$. This is the critical region because under the null of no ability to distinguish, 4 successes has 1 chance out of 70 ($\approx 1.4\% < 5\%$) of occurring (...) in the actual experiment the lady succeeded in identifying all eight cups correctly (...) in 70 (the combinations of 8 taken 4 at a time). (...) the famous case of the 'lady tasting tea' (...) one of the two supporting pillars ... of the randomization analysis of experimental data.

Il valore $\frac{1}{70}$ viene dalla Probabilità Combinatoria, essendo

$$\binom{8}{4} = \frac{8 \cdot 7 \cdot 6 \cdot 5}{1 \cdot 2 \cdot 3 \cdot 4} = 70$$

i modi di scegliere 4 elementi su 8 (combinazioni di 8 elementi a 4 a 4). L'indovinare solo 3 tazze su 4 invece è un evento molto probabile, $\approx 24.3\%$, nell'ipotesi nulla di tirare a caso ovvero di non avere la capacità affermata, e si calcola in modo non semplicissimo (distribuzione ipergeometrica) analogo al calcolo per il terno al lotto.

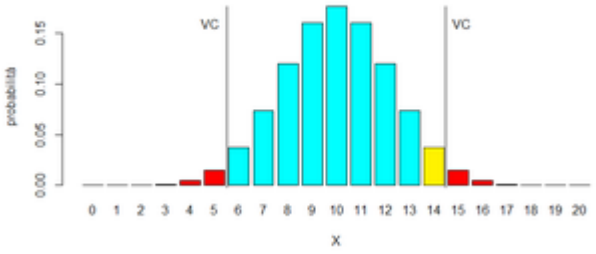


Figure 56: Rosso: regione di rifiuto del test a due code. Giallo: il numero 14 si aggiunge alla coda destra 15,...,20 dando – senza la coda sinistra – la regione di rifiuto del test a una coda. By Ppong.it in wikimedia.

38.5 Test a una coda e a due code

Wikipedia ci illustra un test per la regolarità della moneta molto meglio fatto di quello prima esposto con 5 lanci, e ne prevede 20. La regione critica è $\{0,1,2,3,4,5,15,16,17,18,19,20\}$, che ha probabilità 0.041 se la moneta è regolare, a cui corrisponde la *regione di accettazione* $\{6,7,8,9,10,11,12,13,14\}$, l'insieme complementare. Questo è un *test a due code*.
Il valore 0.41 è la probabilità di un grave allontanamento da 10, metà di 20, speranza matematica del numero di teste nell'ipotesi di regolarità, cioè speranza matematica di $B\left(20,\frac{1}{2}\right)$:

$$P(X \le 5 \vee X \ge 15) = P(X \le 5) + P(X \ge 15) =$$

$$= 1 - P(6 \le X \le 14) = 1 - \sum_{k=6}^{k=14} \binom{20}{k} \left(\frac{1}{2}\right)^k \left(1-\frac{1}{2}\right)^{n-k} =$$

$$= 1 - 2^{-20} \sum_{k=6}^{k=14} \binom{20}{k}$$

calcolabile con pazienza oppure con WolframAlpha che dà ≈ 0.4139.

Con un campione di 8 studenti possiamo ritenere maggioritarie le femmine se ne troveremo almeno 6:
WolframAlpha
Leggiamo in
https://it.wikipedia.org/wiki/Test_di_verifica_d%27ipotesi

> Tale valore p è la probabilità di ottenere un valore altrettanto o più estremo di quello osservato, ammesso che la

> moneta fosse effettivamente bilanciata. Nel nostro caso è pari a 0,041, ovvero del 4,1%. Giudicando bassa tale probabilità, rigettiamo l'ipotesi di bilanciamento della moneta in esame, ritenendo accettabilmente basso il rischio di compiere un errore di giudizio. La probabilità di rifiutare l'ipotesi sottoposta a verifica, nel caso questa fosse corretta, è pari al massimo valore-p che saremmo stati disposti ad accettare.
> (...) supponiamo che noi già prima di fare l'esperimento sospettassimo che fosse sbilanciata verso la testa, in tal caso potremmo dire che l'ipotesi nulla, che noi abbiamo intenzione di smentire, è che la probabilità che esca testa sia minore o uguale a 0,5, anziché necessariamente pari a 0,5. In tal modo evitiamo di rifiutare l'ipotesi nulla se otteniamo un numero di teste basso, ma se, al contrario, contiamo più di 10 teste, calcoliamo il valore-p senza tenere in considerazione i possibili risultati inferiori a 10. Come risultato, la regione di rifiuto perde gli elementi da 1 a 5, ma si allarga sulla destra includendo 14.

L'ipotesi e l'alternativa del test a una coda considerato sono

$$H : p \leq \frac{1}{2} \qquad A : p > \frac{1}{2}\,.$$

L'ipotesi e l'alternativa del test a due code considerato sono

$$H : p = \frac{1}{2} \qquad A : p \neq \frac{1}{2}\,.$$

38.6 Test a una coda facilmente fattibile con successo

Siamo abbastanza sicuri che le presente lezione di Farmacia nel 2022 sarà seguita in aula da più studentesse che studenti (succede ogni anno) ovvero detto altrimenti che il generico studente ha probabilità maggiore di essere di genere femminile piuttosto che maschile:

$$p := P(femmina) > \frac{1}{2} \qquad \text{(crediamo)}$$

Naturalmente potremmo contare e risolvere la questione, ma se la popolazione di riferimento fosse enorme o difficilmente indagabile nella sua totalità (per esempio se volessimo dimostrare che a Roma le gatte sono più numerose dei gatti, o simili cose su zecche o batteri o virus o quant'altro) non potremmo farlo. Faremo invece un test statistico su un campione di 20 soggetti presi in qualche modo a caso, per esempio in base alla prima lettera del cognome, che possiamo ragionevolmente ritenere indipendente dal genere del soggetto. **E la questione dell'indipendenza non sarà mai sottolineata abbastanza.**

Poniamo

$$H : p \leq \frac{1}{2} \qquad A : p > \frac{1}{2}$$

e speriamo di respingere l'ipotesi nulla. In base al paragrafo precedente, se troviamo 14 o più studentesse nel campione, ce l'abbiamo fatta, respingiamo l'ipotesi che siano meno o ugualmente numerose *in tutta la popolazione* considerata. Di solito 20 soggetti bastano perchè la sproporzione è notevole (non è affatto detto che sia così per i gatti di Roma), se fosse molto piccola ci vorrebbe un campione più numeroso per avere buone speranze di avere il desiderato successo col test.

38.7 A due code, facilmente fattibile, ma con insuccesso

È senz'altro vero che la probabilità che un soggetto sia nato in un giorno del mese pari non sia uguale a quella del dispari, ma la differenza è così piccola che quasi sicuramente con un campione di 20 soggetti non riusciremo a respingere l'ipotesi nulla

$$H : p = \frac{1}{2}$$

Comunque possiamo provare per vedere un po' come funziona "sul campo" la Statistica. Arruoliamo quindi 20 soggetti presi a caso, per esempio quelli più vicini al computer sulla cattedra, cosa che possiamo ragionevolmente ritenere indipendente dalla parità del

giorno della nascita. **E la questione dell'indipendenza non sarà mai sottolineata abbastanza.**
In base a quanto sopra detto in 38.5, se troviamo meno di 6 o più di 14 nati in data dispari, respingiamo l'ipotesi dell'equiprobabilità, e saremmo contenti (ma difficilmente succederà).
Rivediamo tutto con ordine.
Una persona si convince che non è ugualmente probabile nascere in un giorno con data pari o dispari. Di fatto questo è vero, ma la differenza di probabilità è talmente piccola che il nostro ben difficilmente riuscirà dimostrarlo con qualche decina di dati (verosimilmente riuscirebbe con qualche milione di soggetti).

– Ipotesi nulla che vuol respingere: è equiprobabile nascere in data pari o dispari.

– Pre-fissa il livello di significatività standard 0.05

– Esperimento: intervista 20 persone.

– **Supponiamo per esempio** che trovi 14 pari e 6 dispari (è contento, si illude)

– In un modo o nell'altro calcola $P(pari \leq 6 \vee pari \geq 14) \approx 0.115$

– Non riesce a respingere l'ipotesi nulla al 5% (e neppure al 10%).

Se ne trovava 15, riusciva a respingere l'ipotesi nulla.
Si noti che l'obiezione che lui fa, se non capisce la Statistica, "ho trovato più del doppio, 14, in una classe rispetto all'altra, 6", è irrilevante e fuorviante, su 20 soggetti; potrebbe bastare anche molto meno del +100%, anche solo il +1%, ma su un campione molto più numeroso.

38.8 Il p-value, in italiano valore p

Oggi i quantili vengono calcolati da numerosi software. L'uso di questi software ha permesso nei tempi moderni un passo ulteriore: trovare proprio la soglia discriminante, l'ultimo valore per il quale si passa dal non rifiutare al rifiutare l'ipotesi nulla, e questo valore soglia può ben essere diverso da 0.05 o 0.01, per esempio può essere 0.046, ed è il p-value ovvero valore p, e si trova spesso denotato P

sulle riviste scientifiche di Farmacia e Medicina.
Questo 0.046 è proprio il p-value dei 10 000 lanci di moneta con 5 100 teste prima considerato, come si potrebbe calcolare col computer: rifiutiamo l'ipotesi (nulla) di regolarità (per poco) perchè 0.046 è minore di 0.05 (che però a rigore dovevamo fissare prima dell'esperimento, e ciò è tanto più vero per valori di significatività diversi da quello classico 0.05). Ma al livello dell'1%, detto del 99% da altri Autori, non possiamo rifiutarla. (Stiamo cercando di fare un'affermazione troppo categorica ma l'esperimento non ce lo consente; ce lo consentirebbe se fossero venute molte più teste).

L'ideale è trovare un p-value piccolissimo, magari $< 10^{-6}$, ma almeno ≤ 0.05.

Definizione. Formalmente il p-value è definito come la probabilità di ottenere un valore uguale o più estremo di quello ottenuto, nell'ipotesi che sia vera l'*ipotesi nulla*.

Rivediamo questo fatto del valore più estremo nell'esempio:

– Vogliamo "dimostrare" che la moneta non è regolare

– Ipotesi nulla: moneta regolare ovvero $p = 0.5$, probabilità della testa (e questo p non c'entra nulla col p-value che stiamo per introdurre, è un'ambiguità notazionale).

– Pre-fissiamo il livello di significatività standard 0.05

– Esperimento: lanciamo la moneta 10 000 volte

– Vengono 5 100 teste

– In un modo o nell'altro calcoliamo $P(teste < 4\,900 \vee teste \geq 5\,100) \approx 0.046$

– Rifiutiamo l'ipotesi della regolarità con P 0.046

Nota finale. Il nostro *trial* è andato come speravamo.

Vengono pubblicati anche articolo scientifici con p value 0.1 ma apprezzatissimi sono p value come 10^{-3} o minori, anche 10^{-6} o più piccoli ancora.

Solo per fare un esempio, p value 0.0375, da un articolo[190] scientifico relativo al Sud Africa:

> The minimum distance between the source location and the medical facility was compared with the malaria case mortality rate to determine the relationship between the two factors (...) The correlation analysis illustrated a Spearman's rank correlation coefficient of 0.3816 with a p value of 0.0375.

Minor distanza minima dagli ambulatori, minor mortalità, appare. (Nota: in Italia il numero degli ospedali pubblici diminuisce da decenni, per *spending review*).

Solo per fare un esempio di p value molto piccoli, fra innumerevoli, il preprint[191] *Solar UV - B/A Radiation is Highly Effective in Inactivating SARS-CoV-2* trova per l'ipotesi nulla (sostanzialmente: che il covid sia indipendente dall'irraggiamento solare locale) dei p value dell'ordine di 10^{-11}. E un altro preprint[192] trova dei p value anche dell'ordine di 10^{-5} – 10^{-3} per l'ipotesi nulla che nella prima ondata la diffusione del covid sia stata indipendente dall'abbondanza locale di maiali d'allevamento, e perfino, tentativamente, pur nell'incertezza dei dati, 0.005 per l'indipendenza dall'abbondanza locale di cinghiali, per zone in cui è stato possibile trovare allo stesso livello di suddivisione geografica dati sul covid e sulla densità di cinghiali.

In casi semplici il p-value può essere calcolato a mano, per esempio con i 5 lanci di moneta prima considerati, per rifiutare l'ipotesi (nulla) della regolarità, il risultato "0 teste o 5 teste" ha p-value 1/16 = 0.0625 e l'ipotesi non viene rifiutata al livello del 5% ovvero 0.05 ovvero del 95% ovvero 0.95. Il valore $\frac{1}{16}$ viene da

$$= P(T,T,T,T,T \vee C,C,C,C,C) =$$

[190] Coetzer RH, Adeola AM. Assessing the Correlation between Malaria Case Mortality Rates and Access to Health Facilities in the Malaria Region of Vhembe District, South Africa. J Environ Public Health. 2020 Dec 2;2020:8973739. doi: 10.1155/2020/8973739. PMID: 33343669; PMCID: PMC7732409.

[191] DOI: https://doi.org/10.1101/2020.06.03.20121392

[192] DOI: 10.13140/RG.2.2.29852.31367

$$= P(T,T,T,T,T) + P(C,C,C,C,C) =$$

probabilità di 5 successi o 0 successi, casi ugualmente estremi, e di più estremi di quanto ottenuto non ce n'è, con $B(5,\frac{1}{2})$:

$$\left(\frac{1}{2}\right)^5 + \left(\frac{1}{2}\right)^5 = \frac{1}{32} + \frac{1}{32} = \frac{1}{16}\,.$$

Oppure rifiutiamo l'ipotesi della regolarità, al classico livello debole del 10%, invece di quello standard del 5%.

38.9 Buona Statistica: ipotesi prima dell'esperimento

Si capisce che la buona Statistica si fa formulando l'ipotesi prima dell'esperimento, mentre di minor valore è fare affermazioni su dati già noti; sicuramente a partire dai risultati di 10mila lanci, guardandoli con attenzione, si potranno trovare affermazioni strampalate del tipo "i lanci con numero ordinale che scritto in tagalog iniziano per vocale tendono a dare testa", e magari si riuscirà pure a dimostrarle con buon p value; ma se l'affermazione fosse stata fatta prima dei lanci, ben improbabilmente l'esperimento l'avrebbe confermato.

Si provi a formulare qualche ipotesi del genere su questo campione di 100 lanci virtuali di moneta tratto da WolframAlpha, e si capirà quanto delicata sia la questione della statistica. (H = head = testa = 1, T = tail = croce = 0)

T | H | T | H | T | H | T | H | T | H | T | T | T | T | H | H | H | H | H | T | T | T | T | T | T | H | H | H | T | T | T | T | H | T | H | H | H | T | H | T | T | H | T | H | H | T | H | H | H | H | H | T | T | H | T | T | T | T | T | H | T | T | H | T | H | H | T | H | H | H | H | H | T | H | T | T | H | H | T | T | H | H | T | T | T | H | T | H | T | H | T | H | H | T | T | H | T | T | H | T

Ecco per esempio un'affermazione che risulterà statisticamente confermata sul campione del passato, ma in generale non su uno del futuro, e si provi per credere:

in tutti i lanci dal secondo all'undicesimo la differenza assoluta fra teste è croci uscite (a partire dal primo lancio) è sempre minore

o uguale a 1.

Nello studio randomizzato controllato (Randomized controlled trial, RCT) l'ipotesi si formula prima. Per esempio che un vaccino

ridurrà i contagiati, oppure

ridurrà i casi sintomatici lievi, oppure

ridurrà i casi gravi, oppure

ridurrà i ricoveri in terapia intensiva, oppure

ridurrà la mortalià, oppure
ridurrà le ospedalizzazioni...

Non si può mica fare il *trial clinico* misurando tutte quelle cose, e poi facendo la statistica magari sull'unica andata bene, non dicendo nulla delle altre. Bisogna dire prima cosa si misurerà, e se tutto va bene si potrà respingere l'ipotesi nulla che non abbia funzionato per quell'*outcome*.

Per esempio per il famoso trial clinico del 2020 della Pfizer per ottenere l'autorizzazione in via emergenziale del vaccino anticovid, benchè il grande pubblico magari pensasse che il *main outcome* misurato fosse la morte (cioè che la statistica mostrasse una riduzione della mortalità da covid) in effetti si misurava la riduzione dei casi sintomatici lievi:

> Many may assume that successful phase 3 studies will mean we have a proven way of keeping people from getting very sick and dying from covid-19. And a robust way to interrupt viral transmission.
> Yet the current phase 3 trials are not actually set up to prove either, says Doshi.
> "None of the trials currently underway are designed to detect a reduction in any serious outcome such as hospitalisations, intensive care use, or deaths. Nor are the vaccines being studied to determine whether they can interrupt transmission of the virus," he writes.
> He explains that all ongoing phase 3 trials for which details have been released are evaluating mild, not severe, disease
> https://www.bmj.com/company/newsroom/covid-19-vaccine-trials-cannot-tell-us-if-they-will-save-lives/

38.10 Nota sul sentire statistico

La Statistica Inferenziale è poco insegnata a scuola, e ancor meno lo era nei decenni passati.
Per uno statistico, è sempre notevole constatare quanto poco presente sia la mentalità statistica a livello di popolazione italiana. Da una parte alcuni si attaccano al caso singolo senza rendersi conto di

quanto poco rappresentativo sia della situazione complessiva – in ciò magistralmente teleguidati dai media al soldo dei loro referenti – e dall'altra qualcuno ti obietta che "4 casi non fanno Statistica", credono che ci vogliano sempre migliaia di casi, e non è così, anzi Fisher, "mostro sacro" della Statistica, nel suo pionieristico lavoro fece proprio la famosa brillante esperienza di Statistica con 4 casi; dipende da vari fattori. Certo è vero che neppure migliaia di casi mal scelti bastano a fare buona Statistica, e questo è un errore sempre in agguato nella Statistica Medica e in particolare in Farmacia: fare statistiche solo sui soggetti che portano avanti un esperimento fino alle nostre misurazioni finali; come se fosse la stessa cosa che dei, diciamo, 500, che non completano il trial, siano morti 1 (che magari sarebbe la norma in quel lasso di tempo senza prendere il nostro prodotto sperimentale) oppure 50, oppure siano tutti andati alle Seichelles a godersi la guarigione: i *drop out* sono una catastrofe nella Statistica Medica, e semplicemente non considerarli è del tutto fuorviante.

Questo testo vorrebbe giungere a dare un po' di *forma mentis* statistica.

38.11 Errori di I e II Specie

Come detto, prima di eseguire un test statistico dobbiamo formulare un'ipotesi nulla H e la complementare ipotesi alternativa A.
Per esempio

$$H: \quad \mu \leq 0$$

$$A: \quad \mu > 0$$

dove μ potrebbe essere la media di una variabile aleatoria X di cui disporremo di un campione aleatorio $X_1, \ldots, X_n$. I ruoli delle 2 ipotesi non sono interscambiabili: in linea generale come ipotesi alternativa va fissata quella che speriamo vera (e come ipotesi nulla quella che ci avrebbe fatto perdere tempo).
Tratto ovvero prodotto ovvero rilevato il campione x_1,...,x_n, ne calcoliamo l'opportuna funzione che la statistica ci insegnerà a seconda del tipo di test, sia essa ora $g(x_1, .., x_n)$, per esempio la media $\bar{x}_n$, ci sono 4 casi relativamente alla regione critica D, a $g(x_1, .., x_n)$, e alla verità dell'ipotesi H:

1) $g(x_1, .., x_n) \in D$ ed è vera H:
male respingo ipotesi vera: errore di prima specie.
2) $g(x_1, .., x_n) \in D$ ed è vera A:
bene respingo ipotesi falsa: è il caso sperato. (α)
3) $g(x_1, .., x_n) \notin D$ ed è vera H:
non respingo ipotesi vera: ho perso tempo.
4) $g(x_1, .., x_n) \notin D$ ed è vera A:
male non respingo ipotesi falsa: errore di seconda specie.

Esempio di errore di seconda specie. L'errore di seconda specie è quello che si commise quella volta in cui si fece – per davvero – il test indicato in 38.6 trovando 13 femmine su 20 soggetti presi a caso, non riuscendo così a respingere l'ipotesi nulla (che volevamo respingere) che per un generico uditore della lezione fosse:
$P(maschio) \geq P(femmina)$
fatto comunque effettivamente falso, fattualmente: male non resp-

ingemmo ipotesi falsa, quella della minore o uguale probabilità del genere femminile. Che in aula ci fossero più femmine si vedeva conteggiando tutti i casi in aula, ma questo ovviamente non si sarebbe potuto fare nel caso, per esempio, fossimo stati interessati a dimostrare che in città, in strada, erano più numerose le gatte che i gatti.

Sulla gravità dell'errore di prima specie.
L'errore di prima specie è considerato molto più grave.
Ad esempio per un farmaco si può mettere l'ipotesi che non curi, sperando di falsificarla con l'esperimento.
La cosa peggiore è diffondere a milioni di persone un farmaco che non cura, solo perché in un *trial clinico* – necessariamente molto più limitato – ha dato buona prova di sè, causando appunto l'errore di prima specie – di fatto sempre possibile.
Diffondere il farmaco a milioni o miliardi di persone, affermandolo migliore del farmaco precedente (o di nessun farmaco) è un fatto gravissimo, magari la gente poteva curarsi col farmaco precedente con migliori risultati.

Leggiamo su Wikipedia, l'enciclopedia libera, in cui chiamano *null hypothesis H0* l'*ipotesi nulla* H_0, ma purtroppo *hypothesis* quella che in questa trattazione è chiamata *alternativa*:

> Hypothesis: "A patient's symptoms improve after treatment A more rapidly than after a placebo treatment."
>
> Null hypothesis (H0): "A patient's symptoms after treatment A are indistinguishable from a placebo."
>
> A Type I error would falsely indicate that treatment A is more effective than the placebo, whereas a Type II error would be a failure to demonstrate that treatment A is more effective than placebo even though it actually is more effective.

Stabiliamo cosa intendiamo per “il farmaco funziona".
Per esempio

aumenta la sopravvivenza a 5 anni rispetto al placebo
riduce la massa tumorale
riduce la glicemia negli iperglicemici.

I 4 casi in dettaglio sono questi.

1) e 3) è vera l'ipotesi H: il farmaco è inutile o dannoso (esempio: sopravvivenza a 5 anni uguale o diminuita).
2) e 4) è vera l'alternativa A: il farmaco è utile (esempio: sopravvivenza a 5 anni aumentata).

1) e 2) La sperimentazione dà esito buono.

1) Per caso i soggetti trattati sono vissuti a lungo e il farmaco dannoso fa bella figura e magari si diffonde: GRAVISSIMO.

2) I soggetti trattati sono vissuti a lungo grazie al farmaco che giustamente fa bella figura

3) e 4) La sperimentazione dà esito cattivo.

3) non respingo l'ipotesi che il farmaco sia inutile o dannoso; il farmaco inutile o nocivo è stato correttamente riconosciuto tale sperabilmente non verrà commercializzato. Ho perso tempo, speravo fosse utile.

4) Per caso i soggetti trattati sono vissuti poco ma il farmaco in generale funziona: il farmaco utile viene purtroppo abbandonato. Peccato.

Gli errori di prima e seconda specie hanno un analogo molto suggestivo negli errori giudiziari. L'ipotesi nulla è l'innocenza:

	innocente	colpevole
condannato	errore di I specie	ottimo
assolto	bene (ma perso tempo)	errore di II specie

Sull'assoluzione del colpevole, errore meno grave della condanna di un innocente, vale il classico: IN DUBIO PRO REO.

(Ovviamente lo scopo dell'apparato giudiziario non è assolvere gli innocenti ma condannare i colpevoli: se facesse questo e non altro realizzerebbe il suo compito, per questo è scritto "perso tempo").

In Medicina e Farmacia, gli errori di I e II specie si presentano anche nella considerazione dei test diagnostici, **ma con una seria duplicità ovvero ambiguità.**

In ambito medico alcuni considerano
ipotesi nulla: la malattia è presente
altri
ipotesi nulla: malattia non presente.

Con riferimento al secondo modo di impostare la questione, si ha allora questo schema:

	sano	malato
positivo	falso positivo: errore di I specie *male respingo ipotesi vera*	vero positivo: ottimo: trovato!
negativo	vero negativo bene ma perso tempo	falso negativo: errore di II specie (condizione non rilevata)

In quest'ottica, il peggio sarebbe curare un sano, il falso positivo a un test diagnostico, a rischio di danneggiarlo: era già sano! PRIMUM NON NOCERE.
Nell'altra ottica il peggio, l'errore di I specie, sarebbe il falso negativo, in pratica non avvertire della malattia il malato.
Attenzione a quest'ambiguità.

38.12 Nota sul parametro medico su cui si fa il test

Per un chemioterapico si può ipotizzare di testare la sopravvivenza a 5 anni (confrontando farmaco e placebo); oppure la riduzione della massa tumorale in un breve tempo; se poi in un tempo doppio il paziente muore, questo non rientra nella statistica. È ovvio che la sopravvivenza a 5 anni richiede una sperimentazione lunghissima, e costosa.
Ma per quanto di altissimo valore, anche quel solo parametro è poco, nella complessità della realtà: non si è tenuto conto della qualità della vita in seguito alla somministrazione del farmaco.

Essenzialmente si tratta di pre-decidere gli *outcomes* ricercati in un trial clinico, vedi https://en.wikipedia.org/wiki/Randomized_controlled_trial.

Tutt'altra questione: non si è tenuto conto neppure del costo economico: gli stessi soldi il Sistema Sanitario potrebbe spenderli con maggiore beneficio complessivo per la cura di altre malattie. (Resource Management, e in caso di sanità privata anche ROI, Return On Investment).

A questo proposito, su cosa si testa, leggiamo sul British Medical Journal, in https://www.bmj.com/company/newsroom/covid-19-vaccine-trials-cannot-tell-us-if-they-will-save-lives/, datato 21 ottobre 2020, e si noti che qua la prestigiosissima rivista scientifica tende a sostenere un punto di vista scientifico, non *differenti* interessi:

> None of the current trials are designed to detect a reduction in any serious outcome such as hospitalisations, intensive care use, or deaths
>
> Vaccines are being hailed as the solution to the covid-19 pandemic, but the vaccine trials currently underway are not designed to tell us if they will save lives, reports Peter

Doshi, Associate Editor at The BMJ today.

Several covid-19 vaccine trials are now in their most advanced (phase 3) stage, but what will it mean exactly when a vaccine is declared “effective"?

Many may assume that successful phase 3 studies will mean we have a proven way of keeping people from getting very sick and dying from covid-19. And a robust way to interrupt viral transmission.

Yet the current phase 3 trials are not actually set up to prove either, says Doshi.

“None of the trials currently underway are designed to detect a reduction in any serious outcome such as hospitalisations, intensive care use, or deaths. Nor are the vaccines being studied to determine whether they can interrupt transmission of the virus," he writes.[193]

[193] E continua:
He explains that all ongoing phase 3 trials for which details have been released are evaluating mild, not severe, disease - and they will be able to report final results once around 150 participants develop symptoms.

In Pfizer and Moderna’s trials, for example, individuals with only a cough and positive lab test would bring those trials one event closer to their completion.

Yet Doshi argues that vaccine manufacturers have done little to dispel the notion that severe covid-19 was what was being assessed.

Moderna, for example, called hospitalisations a “key secondary endpoint" in statements to the media. But Tal Zaks, Chief Medical Officer at Moderna, told The BMJ that their trial lacks adequate statistical power to assess that endpoint.

Part of the reason may be numbers, says Doshi. Because most people with symptomatic covid-19 infections experience only mild symptoms, even trials involving 30,000 or more patients would turn up relatively few cases of severe disease.

“Hospitalisations and deaths from covid-19 are simply too uncommon in the population being studied for an effective vaccine to demonstrate statistically significant differences in a trial of 30,000 people," he adds. “The same is true regarding whether it can save lives or prevent transmission: the trials are not designed to find out."

Zaks confirms that Moderna’s trial will not demonstrate prevention of hospitalisation because the size and duration of the trial would need to be vastly increased to collect the necessary data. “Neither of these I think are acceptable in the current public need for knowing expeditiously that a vaccine works," he told The BMJ.

Moderna’s trial is designed to find out if the vaccine can prevent covid-19 disease, says Zaks. Like Pfizer and Johnson and Johnson, Moderna has designed its study to detect a relative risk reduction of at least 30% in participants developing lab-confirmed covid-19, consistent with FDA and international guidance.

D'altra parte il testo sopra riportato è ormai (2021) superato, e in progresso di tempo molto è stato fatto per valutare i risultati mancanti negli studi iniziali, e ormai noi tutti sappiamo che i vaccini salvano delle vite.

38.13 Dopo gli errori di I e II specie, le assurdità

Gli errori di I e II specie sono essenzialmente sfortune, e, in prima approssimazione, non ha colpa chi fa il test: che colpa potrebbe avere se ha giudicato plausibilmente non regolare (rifiutando l'ipotesi di regolarità) una moneta pur regolare che però ha dato 19 teste su 20 lanci? Caso sfortunato. A un livello superiore rispetto a questa trattazione, comunque, si stabilisce come ridurre il rischio di quegli errori, dopodichè commetterli è parzialmente colpa del ricercatore.

Invece, bisogna fare attenzione – prevenire è meglio che curare – a veri e propri errori che può facilmente fare il non statistico addentro alle cose biomediche.

Supponiamo che 2 test statistici ben fatti (magari ciascuno è addirittura uno studio controllato randomizzato, ovvero studio clinico

Zaks also points to influenza vaccines, saying they protect against severe disease better than mild disease. "To Moderna, it's the same for covid-19: if their vaccine is shown to reduce symptomatic covid-19, they will feel confident it also protects against serious outcomes," Doshi writes.

But Doshi raises another important issue - that few or perhaps none of the current vaccine trials appear to be designed to find out whether there is a benefit in the elderly, despite their obvious vulnerability to covid-19.

If the frail elderly are not enrolled into vaccine trials in sufficient numbers to determine whether there is a reduction in cases in this population, "there can be little basis for assuming any benefit against hospitalisation or mortality," he warns.

Doshi says that we still have time to advocate for changes to ensure the ongoing trials address the questions that most need answering.

For example, why children, immunocompromised people, and pregnant women have largely been excluded; whether the right primary endpoint has been chosen; whether safety is being adequately evaluated; and whether gaps in our understanding of how our immune system responds to covid-19 are being addressed.

"The covid-19 vaccine trials may not have been designed with our input, but it is not too late to have our say and adjust their course. With stakes this high, we need all eyes on deck," he argues.

controllato randomizzato – RCT, dall'inglese randomized controlled trial), trovino con p value piccolissimi, che i soggetti con una certa caratteristica rimovibile, per esempio

- uso di eroina
- abbonamento al mensile *Il cacciatore di gatti con la fionda*

hanno maggior probabilità di fare un incidente stradale (per esempio nell'anno seguente), rispetto a chi manca di quella caratteristica.
Il profano di Statistica potrebbe pensare che, visto che i p value sono piccolissimi, forte indizio di non casualità delle correlazioni trovate, eliminare dalle persone le 2 caratteristiche rimovibili, l'eroina e l'abbonamento, riduca loro la probabilità fare un incidente stradale nell'anno successivo.
E infatti se riesce a togliergli l'eroina probabilmente gli ridurrà la probabilità di incidente e anche di morte nell'anno successivo (non ne siamo certi ma è verosimile).
Revocandogli l'abbonamento, però, probabilmente non otterrà molto.
Questo succede, molto approssimativamente, perchè la correlazione è cosa ben diversa dalla causalità, e il p value piccolissimo rende molto plausibile solo la correlazione delle vere variabili aleatorie retrostanti (abbonamento, una variabile aleatoria bernoulliana $X \sim B(1, p_1)$, e incidente, un'altra variabile aleatoria bernoulliana $Y \sim B(1, p_2)$) rendendo poco plausibile la casualità (dei valori empirici rilevati). È verosimile che gli abbonati a quel mensile siano molto meno intelligenti della media, ed è questo che causa la loro maggiore mortalità per incidenti stradali, non l'abbonamento. Revocato quello, non è che diventano subito molto più intelligenti; e continuano a fare incidenti.

Naturalmente sono possibili errori più sottili, proprio di tipo medico e non statistico, nell'interpretazione dei risultati degli esperimenti medici. Enfasi aggiunta:

> L'eroina(...) Venne risintetizzata nel 1897 da Felix Hoffmann, un chimico tedesco che lavorava per la Bayer; in quel periodo l'acetilazione era una tendenza diffusa per la ricerca di molecole più attive. Hoffmann realizzò l'acetilazione dell'acido salicilico, ottenendo l'Aspirina, e solamente 11 giorni dopo fece altrettanto con la morfina, producendo appunto l'eroina.
> L'intento era quello di ottenere una molecola più efficace della codeina nel sedare la tosse, la tubercolosi e le patologie respiratorie. Le effettive proprietà sedative sul centro del respiro (le stesse che portano a morte nell'overdose) furono inizialmente male interpretate, ritenendo che la riduzione del ritmo respiratorio dipendesse da una migliorata efficienza respiratoria.
> Fu battezzata commercialmente eroina (dal tedesco "heroisch", "eroico",

> giacché inizialmente la si credeva priva degli spiacevoli effetti collaterali di dipendenza e assuefazione palesati dalla morfina) e cominciò a essere venduta liberamente dalla multinazionale farmaceutica Bayer dal 1899. In breve tempo l'impiego terapeutico si ampliò alle più disparate patologie pneumologiche, ma anche neurologiche, ginecologiche, o a semplici dolori; si diffusero pertanto svariate preparazioni farmaceutiche acquistabili liberamente, e questo fece sì che l'eroina divenisse velocemente uno dei farmaci più venduti in assoluto.
>
> A parte qualche farmacologo controcorrente, non la si riteneva in grado di dare dipendenza.
> https://it.wikipedia.org/wiki/Eroina letto il 19 dicembre 2022.

Secondo questo testo – qua non possiamo discutere quanto esatto o inesatto – la riduzione del ritmo respiratorio (constatata senz'altro in un gran numero di casi) non era causata da una migliorata efficienza respiratoria.
E nello stesso soprariportato testo si noti che si ripropone la questione di cosa misuriamo in un trial clinico: solo la riduzione del dolore, in cui magari l'eroina riusciva bene, o anche l'aumento di mortalità – e comunque, in quali tempi? – o pure la dipendenza indotta?
Si noti anche l'ultima frase: "A parte qualche farmacologo controcorrente, non la si riteneva in grado di dare dipendenza". Un classico: *prima ti ignorano, poi ti deridono, poi ti combattono, infine dicono che loro lo hanno sempre detto.*

Nota. Di questa Lezione ci si dovrebbe aspettare che lo studente conosca a memoria la formula (α) oltre a tutto l'inquadramento teorico della questione.

ESERCIZI SULLA LEZIONE 38.11

ESERCIZIO$_{\mu 2018}$

$*$ Supponiamo che per un test statistico, con ipotesi (nulla) H e alternativa A, ad un certo livello α, la regione critica sia $[20.18, +\infty[$ e lo stimatore $T := g(X_1, ..., Xn)$ relativo al test abbia prodotto il valore 19.2, e che sia vera A. Quale delle seguenti affermazioni è vera?

- Si commette un errore di prima specie
- Era il caso in generale sperato
- Si commette un errore di seconda specie
- Si è sostanzialmente perso tempo
- Non è possibile rispondere perché non è specificato il quantile

SVOLGIMENTO

Lo stimatore T vale 19.2 che $\notin$ alla regione critica, e l'ipotesi (nulla) è falsa (perché è vera l'alternativa). Allora "male non respingo ipotesi falsa", cioè

Si commette un errore di seconda specie

ESERCIZIO$_{\mu 2018}$

$*$ Supponiamo che per un test statistico, con ipotesi (nulla) H, e alternativa A vera, al livello $\alpha = 0.1$, la regione critica sia $T > 734.66$ e il calcolo dello stimatore del test dia $T = g(x_1, ..., x_n) = 786.45$. Quale delle seguenti affermazioni è vera?

- Si commette un errore di prima specie
- Era il caso in generale sperato
- Si commette un errore di seconda specie
- Si è sostanzialmente perso tempo
- Non ha senso perché α deve essere ≤ 0.05 ossia 5%.

SVOLGIMENTO

Lo stimatore cade nella regione critica, perché $786.45 > 734.66$, e allora l'ipotesi (nulla) viene respinta, ed essa è falsa perché l'alternativa è vera. Siamo nel caso *"bene respingo ipotesi falsa"* che, come è ben noto, per i test statistici

Era il caso in generale sperato

Nota. Di questa Lezione ci si dovrebbe aspettare che lo studente conosca a memoria la formula (α) oltre a tutto l'inquadramento teorico della questione.