

Capitolo 2. Ipotesi e test statistici. Distribuzione t . Variabilità della media ed errore standard. Test t per campioni indipendenti e appaiati. Limiti fiduciali della media. Grandezza del campione.

I test statistici

Il termine test è sinonimo di prova, verifica, accertamento, ecc. Tutti i test (test di gravidanza, test elettorale, test di ammissione, test statistico, ecc.) si basano sulla verifica di una certa condizione ipotizzata. La verifica non avviene in modo diretto, ma attraverso la valutazione di fenomeni strettamente correlati con la condizione ipotizzata. Pertanto, poiché manca l'evidenza diretta, non avremo certezza ma solo una fiducia più o meno grande nel fatto che la condizione esista.

Le proposizioni di questo ragionamento sono normalmente sottintese, nascoste nelle pieghe della nostra mente. Ad esempio, dire 'test di gravidanza' vuol dire:

- verificare la condizione di gravidanza attraverso
- due ipotesi mutualmente esclusive: gravidanza sì / gravidanza no
- non in modo diretto, es. riconoscendo l'embrione, ma valutando un fenomeno strettamente correlato con la condizione di gravidanza (la positività di una reazione per le gonadotropine corioniche HCG nelle urine).

Quindi l'esito del test non dà certezza, ma solo una fiducia valutabile in termini di probabilità. Nell'esempio citato, il test per le HCG - per quanto affidabile - può essere influenzato dalle condizioni dei reagenti (es., mal conservati), dell'ambiente (es., temperatura fuori range), del campione biologico (es., alterato), ecc. Per cui potremo avere i seguenti risultati:

		Risultato del test	
		Positivo	Negativo
Condizione reale ignota	Non Gravidanza	falso-positivo	vero-negativo
	Gravidanza	vero-positivo	falso-negativo

Si definisce

- **sensibilità** o potenza di un test la frequenza di risultati veri-positivi
- **specificità** del test la frequenza di veri-negativi
- **a** la frequenza di falsi-positivi (detti anche errori del 1° tipo)
- **b** la frequenza di falsi-negativi (detti anche errori del 2° tipo)

e inoltre

- **ipotesi zero (H_0) o ipotesi nulla (H_N)**, l'ipotesi di non novità o di non variazione
- **ipotesi 1 (H_1) o ipotesi alternativa (H_A)**, l'ipotesi di novità o di variazione

Specificità e α sono complementari. Sono quindi complementari veri-negativi e falsi-positivi. Infatti se un test è sempre giustamente negativo sulle donne non gravide (100% di veri-negativi) non segnalerà mai positività sulle stesse donne per errore (0% di falsi-positivi).

Sensibilità e β sono complementari. Sono quindi complementari veri positivi e falsi negativi. Infatti se un test è sempre giustamente positivo sulle donne gravide (100% di veri-positivi) non segnalerà mai negatività sulle stesse donne per errore (0% di falsi negativi).

Quindi dire che un test è specifico è come dire che ha una bassa probabilità di falsi positivi, come anche che α è piccolo.

Dire che un test è sensibile è come dire che ha una bassa probabilità di falsi negativi, come anche che β è piccolo.

Riassumendo possiamo indicare:

		Risultato del test		totale
		Positivo T+	Negativo T-	
Condizione reale ignota	H0: Non Gravidanza G-	<i>quantifalsi positivi</i> errore α T+/G-	<i>quantiveri negativi</i> specificità T-/G-	100% dei casi di non gravidanza T+/G- + T-/G-
	H1: Gravidanza G+	<i>quantiveri positivi</i> sensibilità T+/G+	<i>quantifalsi negativi</i> errore β T-/G+	100% dei casi di gravidanza T+/G+ + T-/G+

Un test per essere buono deve possedere sia un'alta specificità che un'alta sensibilità. Non ha alcun senso un test altamente sensibile ma niente specifico (come ad esempio un test sempre positivo nel caso di gravidanza ma anche positivo nel caso di non gravidanza). Analogamente non ha alcun senso un test altamente specifico ma niente sensibile (ad esempio, un test sempre negativo in caso di non gravidanza, ma anche negativo in caso di gravidanza).

Cambiamo esempio ed immaginiamo il risultato di indagini di polizia a carico di un sospetto. Immaginiamo anche che le indagini raccolgano una serie di indizi ma non delle prove così sicure che rivelino con certezza la colpevolezza o l'innocenza dell'indagato (anche se il confine esatto tra indizio e prova resta soggettivo). Notare che in inglese il termine 'trial' significa sia processo giudiziario che esperimento controllato. Possiamo considerare anche in questo caso i quattro risultati:

		Risultato delle indagini	
		Indizi gravi Condannato	Indizi lievi Assolto
Condizione reale ignota	Innocente	falso-positivo	vero-negativo
	Colpevole	vero-positivo	falso-negativo

Le maggiori differenze rispetto all'esempio precedente riguardano:

- le conseguenze del risultato del test: qui si tratta di lasciare in libertà o mandare in prigione un individuo
- la ripetibilità del test: l'indagine di polizia non può essere ripetuta con disinvoltura, può durare mesi e costare molti soldi (mentre un test di gravidanza può essere ripetuto diverse volte senza eccessivo sforzo)

Questi ultimi aspetti impongono al giudice di considerare con estrema attenzione tutti i fatti prima di emettere il verdetto.

Consideriamo ora il fatto che gli indizi possono essere più o meno lievi o più o meno gravi. In altre parole gli indizi possono essere di qualsiasi genere. Possiamo quindi considerare una scala che rappresenti la gravità degli indizi. Ciò ci consente di analizzare meglio il caso del giudizio cosiddetto garantista e quello del giudizio cosiddetto sommario. Il giudizio garantista tende ad emettere condanna solo nel caso in cui esistano gravissimi indizi. Il giudizio sommario invece tende ad emettere condanna anche nei casi in cui gli indizi siano semplici sospetti. Queste diverse decisioni fanno variare la frequenza di falsi positivi (cioè innocenti condannati). Il giudizio garantista limita al massimo il rischio di condannare un innocente mentre il giudizio sommario non si preoccupa troppo di tale problema. In tal modo il giudizio garantista comporta un aumento di falsi negativi (cioè colpevoli assolti) mentre il giudizio sommario riduce tale rischio. Per decidere quale metodo sia il migliore occorre porsi il quesito: l'errore che si commette condannando un innocente è pari a quello che si commette assolvendo un colpevole? Tutte le persone di buon senso sono in grado di affermare che, tra le due, è meglio assolvere un colpevole che condannare un innocente.

		risultato delle indagini = gravità degli indizi 10 9 8 7 6 5 4 3 2 1 0 condanna  assoluzione soglia di decisione di un giudizio garantista	
Condizione reale ignota	Innocente	α piccolo	
	Colpevole		β grande

		risultato delle indagini = gravità degli indizi 10 9 8 7 6 5 4 3 2 1 0 condanna  assoluzione soglia di decisione di un giudizio sommario	
Condizione reale ignota	Innocente	α grande	
	Colpevole		β piccolo

Ora occorre considerare anche che i metodi per dimostrare la colpevolezza sono diversi dai metodi per dimostrare l'innocenza. Ad es., una impronta dimostra la colpevolezza, un alibi dimostra l'innocenza, ecc. ecc. Quindi, essendo diversi i metodi sono anche diversi gli errori α e β . Potremmo paradossalmente avere sia α che β grandi (se siamo un po' tonti e scegliamo dei metodi sbagliati) oppure sia α che β piccoli (se invece siamo bravi). Ecco perché α e β , cioè specificità e sensibilità, sono abbastanza indipendenti. In effetti quella linea verticale che separa le due colonne della tabella dovrebbe essere una linea di spessore variabile, che lascia libertà di avere α e β più o meno ampi.

E quindi bene utilizzare i migliori metodi che riducono sia l'errore di 1° tipo (α) che quello di 2° tipo (β). Per questo il giudice deve essere estremamente scrupoloso, attento e paziente nel valutare tutti gli elementi del processo. Tuttavia, al termine del dibattito, possono restare dei dubbi. Occorre quindi decidere quale tipo di errore sia più grave e quale livello di rischio si voglia accettare: più

garanzia per l'innocente può significare più rischio che un colpevole sia assolto, e viceversa. Questo è il difficile mestiere del giudice. Tutto è più semplice quando è queste valutazioni sono affrontate in forma quantitativa col supporto della statistica.

Analizziamo quindi i possibili risultati di un esperimento di laboratorio.

		Risultato dell'esperimento	
		Positivo	Negativo
Condizione reale ignota	Trattamento non efficace	falso-positivo <i>falsa scoperta o errore di 1° tipo</i>	vero-negativo <i>nessuna novità</i>
	Trattamento efficace	vero-positivo <i>vera scoperta</i>	falso-negativo <i>scoperta mancata o errore di 2° tipo</i>

La posta in gioco è il riconoscimento di una scoperta (ed eventuali finanziamenti, annunci a congressi, onori, ecc. ecc.).

Domanda: quando può si può riconoscere una scoperta ?

Risposta: il risultato positivo di un esperimento può essere accettato quando la probabilità che sia positivo per caso (fasullo o falso-positivo, valutato da α) è minore del 5%.

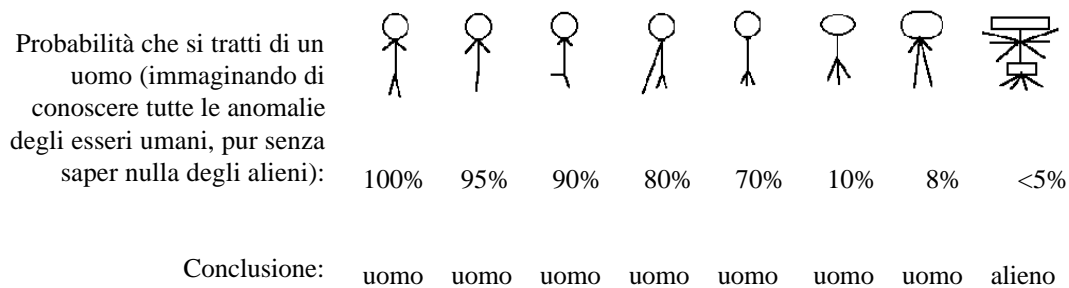
Se α è uguale o maggiore del 5% è opportuno rinunciare al riconoscimento della scoperta. Se poi il ricercatore crede nella sua ipotesi, potrà affinare le tecniche e migliorare l'esperimento in modo da giungere ad una conclusione più matura. Una scoperta fasulla non compromette solo la credibilità del ricercatore, ma comporta una serie di gravi conseguenze in termini di perdita di salute (es., nuovi farmaci che non curano), danari (investimenti per produrre i farmaci fasulli, poi riconosciuti e banditi dal commercio) e perdita di tempo (il lavoro di altri ricercatori ingannati dalle informazioni sbagliate).

Al proposito è interessante citare il caso di una trasmissione televisiva in cui si discuteva sulla natura umana o aliena di un cadavere mostrato steso su un letto in un vecchio filmato. Il cadavere mostrava caratteristiche anatomiche atipiche ma comunque riferibili a certe patologie genetiche. Nessun altro dato - biochimico, microscopico, ecc. - era disponibile oltre le immagini. Senza entrare nel merito del modo in cui la trasmissione era condotta, la situazione proposta in TV può essere affrontata con la logica di un test. Due sono le ipotesi:

- Ipotesi nulla: essere umano. Poiché sappiamo come è fatto un essere umano, possiamo esattamente valutare la probabilità che qualcosa sia un essere umano = possiamo valutare la probabilità dell'ipotesi nulla.

- Ipotesi alternativa: essere alieno non-umano. Ma poiché non conosciamo affatto gli esseri alieni (non abbiamo il modello del marziano, del venusiano, ecc.) non possiamo valutare direttamente la probabilità che qualcosa sia un essere alieno = non possiamo valutare direttamente le probabilità a favore dell'ipotesi alternativa.

Ora, stando a quando visto e quanto affermato da patologi e genetisti, esisteva qualche dubbio sulla causa delle anomalie presenti nel cadavere, ma non sul fatto che si trattasse di un uomo. Se avessero invitato un esperto di statistica (o semplicemente uno studente iscritto al 2° anno di Medicina) la questione si sarebbe risolta dicendo che non è lecito rifiutare l'ipotesi nulla e quindi credere al marziano sinché l'ipotesi nulla gode ancora di una probabilità superiore o uguale al 5%. Nel caso specifico la probabilità a favore dell'uomo era ben più alta, potrei dire oltre il 95%. Per cui c'è da suggerire agli ufologi di studiare un po' di statistica prima che vengano i marziani a insegnarla loro.



Nel caso del ricercatore è abbastanza semplice osservare i fenomeni attraverso mezzi e strumenti che forniscono misure quantitative oggettive. Questo consente l'analisi della distribuzione dei dati, dei parametri statistici e la valutazione dei falsi positivi e dei falsi negativi. Lo schema del test statistico è molto simile a quello sinora considerato. Anche in statistica conosciamo la distribuzione dei dati prevista dall'ipotesi nulla, mentre ignoriamo in parte o in tutto quella prevista dall'ipotesi alternativa.

Per il ricercatore medico o biologo,

- L'ipotesi nulla è l'ipotesi dello scettico, quella che nega il risultato, attribuendo le differenze osservate alla naturale variabilità dei fenomeni o al capriccio del campionamento. L'ipotesi nulla mantiene le attuali conoscenze negando la novità, la scoperta, il dato.
- L'ipotesi alternativa sostiene invece che le differenze esistono e non sono attribuibili alla naturale variabilità o al caso. L'ipotesi alternativa sostiene la novità.

		Risultato del test statistico a	
		se a < 5% H0 rifiutata Dato significativo	se a ≥ 5% H0 accettata Dato non significativo
Condizione reale ignota	H0	falso-positivo <i>errore I tipo</i> <i>valutato con a</i>	vero-negativo nessun errore
	H1	vero-positivo nessun errore	falso-negativo <i>errore II tipo</i> <i>valutato con b</i>

Occorre mantenere l'ipotesi nulla fino a che le prove o i dati in nostro possesso non siano tali da costringerci a rifiutarla (come ogni soggetto è ritenuto innocente sino a che non si dimostri il contrario). Concediamo quindi fiducia all'ipotesi nulla, rifiutandola solo quando l'evidenza dei risultati sia macroscopica, cioè quando la probabilità di falsi-positivi α sia minore del 5%. Quando $\alpha < 0.05$ il risultato del test è detto statisticamente significativo e si rifiuta l'ipotesi nulla. Meglio ancora se $\alpha < 0.01$, o $\alpha < 0.001$ ecc. In tal caso la probabilità di sbagliarci nel riconoscere la scoperta è inferiore a 1 caso su 100, o a 1 caso su 1000, ecc. Si parla di risultato altamente significativo.

I test sono sempre condotti sull'ipotesi nulla perché vogliamo privilegiare l'evidenza dei falsi-positivi rispetto ai falsi-negativi. Nessuna valore β (probabilità di falsi negativi) per quanto alto, consente di rifiutare l'ipotesi nulla quando α sia maggiore o uguale a 0.05.

Tuttavia quando si considera qualcosa in diretta relazione con la salute dell'uomo occorre prendere soprattutto considerazione il rischio di falsi-negativi e soglie diverse dal 5%. Consideriamo ad esempio il morbo della cosiddetta mucca pazza: non vi sono ancora dati statisticamente significativi circa il fatto che il morbo della mucca possa contagiare l'uomo (quindi, potremmo dire che $\alpha > 5\%$). Tuttavia, ci può essere un certo rischio che il dato negativo sia falso (es., $\beta > 50\%$). Pertanto occorre ragionare su due binari: quello scientifico che per ora non ha dimostrato la contagiosità del morbo e quello sanitario che, in attesa di dati scientifici più robusti, non consente di correre rischi.

Riassumendo:

Alla base di ogni test vi sono due ipotesi alternative: H_0 e H_1 .

Il test verte sulla probabilità di H_0 (**α**).
Per decidere se rifiutare o accettare H_0 occorre quindi valutare **α** .
A questo punto...

...ogni test calcola la sua specifica statistica, es. **t , r , F , q , z , t , χ^2** , ecc.

Il valore della statistica calcolata, confrontato con la sua distribuzione (vedi: tabella e valori critici) consente di ottenere **α** .

La statistica t di Student

In precedenza abbiamo imparato a

- stimare la variabilità dei dati mediante il parametro della deviazione standard
- stimare la variazione di un singolo dato rispetto alla media mediante la standardizzazione:

$$z = \frac{x - m}{s}$$

Parallelamente, a livello di medie, ora dobbiamo

- stimare la variabilità delle medie di vari campioni rispetto alla media vera della popolazione mediante il parametro della deviazione standard delle medie
- stimare la variazione della media di un singolo campione rispetto alla media vera della popolazione mediante la standardizzazione:

$$t = \frac{\text{media del campione} - \text{media vera della popolazione}}{\text{deviazione standard della media (del campione)}}$$

Variabilità delle medie ed errore standard

Se si avesse la possibilità di estrarre un certo numero di campioni da una stessa popolazione, si troverebbe che ogni campione ha una media diversa:



Queste medie, anche se diverse e ottenute da campioni di differente numerosità, sono tutte stime della stessa media della popolazione. Esiste pertanto una variabilità della media del campione (d'ora in poi detta media campionaria) attorno alla vera media della popolazione che è per lo più sconosciuta. Tale variabilità è stimabile come **deviazione standard della media** o **errore standard**, scritto col simbolo s_m (deviazione standard della media) o con le lettere ES (errore standard) o SE (standard error) o SEM (standard error of the mean). Bisogna assolutamente specificare deviazione standard **della media** per non confondersi con la deviazione standard delle osservazioni. Da questo punto di vista il termine di errore standard è meno ambiguo, anche se meno appropriato.

Disponendo di diversi campioni potremmo calcolare le rispettive medie e quindi stimare la loro deviazione standard, esattamente come potremmo calcolare la deviazione standard di un campione di dati. In pratica invece disponiamo quasi sempre di un solo campione, spesso anche piccolo, per cui sarebbe impossibile calcolare la deviazione standard della media in base alla variabilità di differenti medie. La media calcolata è anche l'unica di cui disponiamo. A questo punto ci

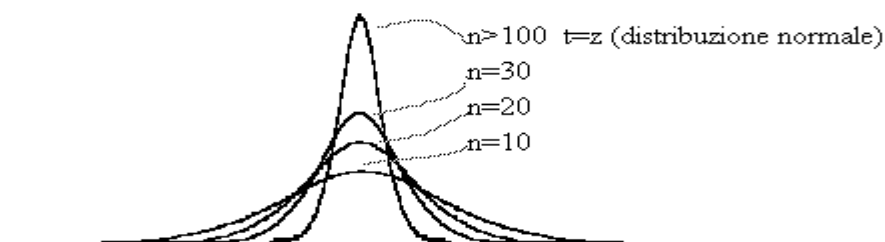
vengono in soccorso i matematici con una formula in grado di stimare la deviazione standard della media s_m in base ai dati di un solo campione:

$$s_m = \frac{s}{\sqrt{n}}$$

Dalla formula si ricava che la deviazione standard della media varia col variare della grandezza del campione (n). Questo si comprende bene considerando il fatto che una media ottenuta da 1,000,000 dati è senz'altro più affidabile di una ottenuta da 100 dati. In termini statistici, diciamo che la deviazione standard di una media ottenuta da un campione di 1,000,000 è molto più piccola della deviazione standard di una media ottenuta da un campione di soli 10 dati. A questo punto, distinguiamo bene tra

- la deviazione standard delle osservazioni del campione (s), che è una *caratteristica della popolazione*, invariante rispetto alla numerosità o grandezza del campione (n), e
- la deviazione standard della media (s_m), che è una *caratteristica del campione* che dipende dalla numerosità del campione (n).

Per questo, quando valutiamo una media dobbiamo sempre tener conto della dispersione della popolazione e della grandezza del campione da cui abbiamo tratto la media. Quindi, a seconda della numerosità del campione (es., $n=10, 20, 30, 100$) la media ha una diversa variabilità come indicato nel grafico:




Il test t di Student

Il test statistico che saggia la differenza tra due medie è il test t di Student.

Le ipotesi alla base del test sono due:

- l'ipotesi nulla, che sostiene che le due medie provengano da campioni estratti dalla stessa popolazione e quindi la loro differenza sia attribuibile a cause accidentali inerenti al campionamento e/o alle misurazioni.
- l'ipotesi alternativa, che sostiene che le due medie siano diverse in quanto rappresentano campioni provenienti da popolazioni diverse (naturali o sperimentali).

Come si valuta la probabilità di falsi-positivi? Occorre innanzi tutto un parametro in grado di valutare la variazione della media. Il fatto che s_m , la deviazione standard della media, vari in funzione di n , fa sì che anche t dipenda da n . Nei campioni numerosi ($n \geq 100$) t è distribuito in modo quasi normale ed ha quindi gli stessi valori critici di z (± 1.96 di ascissa includono il 95% dell'area). Nei campioni più piccoli i valori di t che includono la stessa area sono più grandi proprio perché la distribuzione è più dispersa. La tabella riporta i valori critici di t per i diversi gradi di libertà. I gradi di libertà di t corrispondono ai gradi di libertà del denominatore s_m . Nel nostro caso il denominatore ha $n-1$ gradi di libertà. Osservando la tabella, con 100 gradi di libertà abbiamo meno di 5 probabilità su 100 ($p < 0.05$) di ottenere una media campionaria tanto diversa dalla media della popolazione da produrre un t superiore a 1.96. In altre parole, un t superiore a 1.96 capita per caso meno di 5 volte su cento. Le probabilità scendono all'1% (0.01) per un t superiore a 2.58.

gradi di libertà	α (rischio di falsi positivi)					
	zona della non-significatività		soglia critica 	zona della significatività		
	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.819
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Notare che l'ultima riga della tabella ha gli stessi valori della distribuzione normale. Questo vuol dire che per infiniti gradi di libertà la distribuzione t è identica alla distribuzione normale z . Nella pratica t si considera normalmente distribuito quando i gradi di libertà sono ≥ 100 .

Ogniqualvolta si valuta t bisognerebbe indicarlo come $t_{p,gdl}$ dove p sta per il livello di probabilità scelto (in genere 0.05 o 0.01) e gdl sta per i gradi di libertà. Noi continueremo a scrivere comunque semplicemente $t = \dots$ intendendo però il valore di t che corrisponde ad un certo livello di probabilità con determinati gradi di libertà.

Il fatto di avere una tabella ad intervalli (se t supera... allora $p < \dots$) anziché il valore esatto di p per ogni valore di t calcolato dipende dal fatto che abbiamo tra le mani un foglio e non un computer. E' impossibile far stare in una tabella i valori di p per tutti i valori di t per tutti i gradi di libertà. Il computer è invece in grado di calcolare il valore esatto di p per ogni valore di t calcolato. Comunque la perdita di informazione derivante dall'uso della tabella è minima.

I° caso: medie di due campioni indipendenti

Si dicono campioni indipendenti (o non appaiati) quelli formati da individui diversi. Sono invece detti appaiati i campioni costituiti dagli stessi individui valutati o osservati in tempi diversi (prima e dopo una certa prova) o in condizioni diverse (con o senza un certo trattamento). Il disegno sperimentale che utilizza campioni appaiati è senz'altro più efficace di quello basato su campioni non appaiati o indipendenti. Tuttavia non sempre è possibile applicarlo, sia per problemi pratici, di fattibilità, sia per problemi etici connessi con la sperimentazione clinica su pazienti. Il test t per campioni indipendenti o non appaiati è dato dal rapporto:

$$t = \frac{\text{differenza tra due medie}}{\text{errore standard delle differenze tra le medie}} = \frac{m_a - m_b}{s_{m_a - m_b}} = \frac{m_a - m_b}{\sqrt{\frac{S_a + S_b}{n_a + n_b - 2} \cdot \frac{n_a + n_b}{n_a \cdot n_b}}}$$

L'espressione al numeratore non è tanto la differenza tra le due medie quanto la media delle differenze tra i dati dei due gruppi, presi a 2 a 2 (anche se le due espressioni danno risultati equivalenti). Allo stesso modo, l'espressione al denominatore non è tanto un pool dei due errori standard, ma piuttosto l'errore standard di questa media delle differenze tra i dati dei due gruppi.

La formula sfrutta la proprietà che la varianza delle (di tutte le possibili) differenze tra i dati di due popolazioni corrisponde alla somma delle due rispettive varianze:

$$s_{a-b}^2 = s_a^2 + s_b^2$$

Poiché nel nostro caso ci si riferisce a distribuzioni di medie:

$$s_{m_a - m_b}^2 = s_{m_a}^2 + s_{m_b}^2$$

da cui:

$$s_{m_a - m_b} = \sqrt{s_{m_a}^2 + s_{m_b}^2}$$

Dentro la radice possiamo sostituire i due termini ponendo

varianza della media = quadrato della deviazione standard della media = quadrato della (deviazione standard del campione diviso radice di n), cioè:

$$s_m^2 = (s_m)^2 = \left(\frac{s}{\sqrt{n}} \right)^2 = \frac{s^2}{n}$$

Il denominatore della formula del t pertanto diventa:

$$s_{m_a - m_b} = \sqrt{s_{m_a}^2 + s_{m_b}^2} = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$$

...e così va bene. E' meglio però osservare che secondo l'ipotesi nulla H_0 i due campioni provengono dalla stessa popolazione, per cui s_a^2 e s_b^2 sarebbero stime della stessa varianza dei dati della medesima popolazione. Pertanto, è meglio sostituire ciascuna delle due deviazioni standard con una unica stima combinata:

$$s_{\text{comb}}^2 = \frac{\text{somma devianze}}{\text{somma gradi di libertà}} = \frac{S_a + S_b}{n_a + n_b - 2}$$

Quindi, sostituendo e semplificando, il denominatore della formula del t diventa finalmente:

$$s_{m_a - m_b} = \sqrt{s_{m_a}^2 + s_{m_b}^2} = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}} = \sqrt{\frac{s_{comb}^2}{n_a} + \frac{s_{comb}^2}{n_b}} = \sqrt{\frac{s_{comb}^2 \cdot (n_a + n_b)}{n_a \cdot n_b}} = \sqrt{\frac{S_a + S_b}{n_a + n_b - 2} \cdot \frac{n_a + n_b}{n_a \cdot n_b}}$$

La formula si semplifica molto se i campioni sono **bilanciati** (quando $n_a = n_b$).
Quella esposta è preferibile perché generalizzata. Le ipotesi del test t per campioni indipendenti sono:

H0: differenza tra le due medie = 0

H1: differenza tra le due medie \neq 0

Si entra in tabella con $n_a + n_b - 2$ gradi di libertà.

Poiché abbiamo deciso di rifiutare l'ipotesi nulla che sostiene che le due medie provengano dalla stessa popolazione e ad accogliere l'ipotesi alternativa solo quando il rischio di falso-positivo è minore del 5%, riterremo la differenza non significativa se il valore di t non sarà superiore a quello tabulato per $\alpha=0.05$ (detto soglia critica di significatività). Solo quando il valore di t supererà tale valore la differenza sarà ritenuta significativa. A questo punto la probabilità di sbagliarci, nell'accettare l'ipotesi alternativa, è minore del 5%.

test t per campioni non-appaiati (anche non bilanciati)	
H0: differenza tra medie=0	
frequenza del battito cardiaco in gruppi di animali diversi (dati di pura fantasia)	
topi bianchi	topi neri
56	79
75	73
65	85
60	82
76	73
78	-
n=6	n=5
m=68.33	m=78.40
S=429.33	S=115.20
$t=2.138$	
gdl=9	
p=0.061 (non significativo)	

$$t = \frac{m_{\text{topi neri}} - m_{\text{topi bianchi}}}{\sqrt{\frac{S_{\text{topi neri}} + S_{\text{topi bianchi}}}{n_{\text{topi neri}} + n_{\text{topi bianchi}} - 2} \cdot \frac{n_{\text{topi neri}} + n_{\text{topi bianchi}}}{n_{\text{topi neri}} \cdot n_{\text{topi bianchi}}}}} = \frac{68.33 - 78.4}{\sqrt{\frac{429.33 + 115.2}{6 + 5 - 2} \cdot \frac{6 + 5}{6 \cdot 5}}} = -\frac{10.07}{\sqrt{60.50 \cdot 0.3666}} = -2.138$$

Il valore di $p=0.061$ è stato fornito dal calcolatore. In mancanza di calcolatore, si confronta il t calcolato (valore assoluto) con quello tabulato per il livello minimo di significatività del 95%

($\alpha=0.05$) con 9 gradi di libertà, che è pari a 2.262. Poiché il t calcolato non supera il t tabulato si conclude che i valori di frequenza di battito cardiaco nei topi bianchi e neri provengono dalla stessa popolazione e che la differenza riscontrata è attribuibile alle normali fluttuazioni dei campioni.

II° caso: medie di due campioni appaiati

Il test t per campioni appaiati è dato dal rapporto:

$$t = \frac{\text{differenza media}}{\text{errore standard della differenza media}}$$

L'errore standard delle differenze si calcola come abitualmente. Le ipotesi del test t per campioni appaiati sono:

H_0 : differenza media = 0

H_1 : differenza media \neq 0

Si entra in tabella con $n-1$ gradi di libertà, ove n è il numero di coppie di dati.

test t per campioni appaiati (necessariamente bilanciati)		
H_0 : differenza media=0		
frequenza del battito cardiaco negli stessi atleti (dati di pura fantasia)		
prima di una corsa	dopo una corsa	differenze
66	69	-3
70	77	-7
65	95	-30
76	89	-13
70	78	-8
65	74	-9
n=6 coppie di dati		
		m=-11.67
		$s_m=3.896$
		$t=-2.995$
		gdl=5
		p=0.03 (significativo)

$$t = \frac{\text{media delle differenze}}{\text{errore standard delle differenze}} = \frac{-11.67}{3.896} = -2.995$$

Il valore di $p=0.03$ è stato fornito dal calcolatore. In mancanza di calcolatore, si confronta il t calcolato con quello tabulato per il livello minimo di significatività del 95% ($\alpha=0.05$) con 5 gradi di libertà, che è pari a 2.571. Poiché il t calcolato supera il t tabulato si conclude che la corsa ha modificato la distribuzione dei valori di frequenza di battito cardiaco, determinandone un significativo incremento.

Esercizi

Test t per campioni non-appaiati

gruppo A		gruppo B	
x_A	$(x_A - \text{media}_A)^2$	x_B	$(x_B - \text{media}_B)^2$
$n_A =$	$S_A =$	$n_B =$	$S_B =$
$m_A =$		$m_B =$	
GDL = $n_A + n_B - 2 =$			

$$t = \frac{\text{differenza tra due medie}}{\text{errore standard delle differenze tra le medie}} = \frac{m_a - m_b}{\sqrt{\frac{S_a + S_b}{n_a + n_b - 2} \times \frac{n_a + n_b}{n_a \times n_b}}}$$

$$t = \frac{\text{differenza tra due medie}}{\sqrt{\frac{S_a + S_b}{n_a + n_b - 2} \times \frac{n_a + n_b}{n_a \times n_b}}} =$$

$$t =$$

H_0 : differenza tra le medie = 0

$p =$ (vedi tabella)

risultato significativo ?

Test t per campioni appaiati

con / prima	senza / dopo	differenze (d)	$(d - \text{media}_d)^2$
n (coppie di dati) =			
		$m_d =$	$S_d =$
			$s_d^2 =$
			$s_d =$
			$s_{m_d} =$

$$t = \frac{\text{media delle differenze}}{\text{errore standard delle differenze}} = \text{—————} =$$

$$t =$$

H0: differenza media = 0

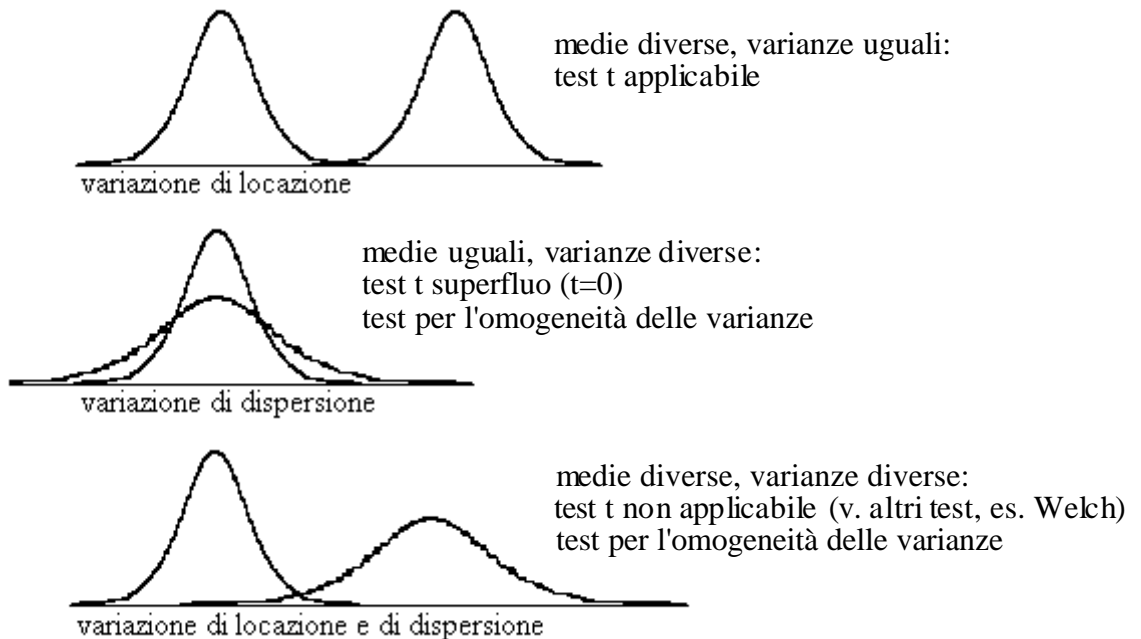
p = (vedi tabella)

risultato signifivativo ?

Ultime considerazioni

Nel test t l'ordine delle medie (a-b, b-a) è irrilevante in quanto la distribuzione t è simmetrica. Pertanto si prescinde dal segno. Nella tabella del t sono riportati solo i valori assoluti.

Poiché il test t valuta differenze tra medie, è applicabile solo a condizione che i dati siano distribuiti normalmente ed abbiano varianze uguali. Tuttavia si dice anche che il test t è **robusto**, è cioè in grado di reggere anche in caso di piccole irregolarità rispetto a queste norme. Lo schema seguente esemplifica diverse situazioni:



Se...

...le varianze sono diverse ma le distribuzioni sono comunque normali (condizione detta di eteroschedasticità) si deve applicare un test t modificato che passa sotto il nome di test di Welch (o test t per varianze diseguali)

...le distribuzioni non sono normali si deve sempre applicare un test non-parametrico, es. il test di Wilcoxon (vedi cap. 8).

L'adozione del corretto tipo di test è fondamentale. Anche Excel (Microsoft) comprende tra le sue applicazioni., oltre il test t di Student, il test di Welch ed il test di Wilcoxon. Controllate.

Un'ultima avvertenza. Se nell'ambito dello stesso studio si effettuano diversi test t tra diverse medie, il rischio globale di falsi positivi aumenta. Per cui il test t non è adatto quando si pianifica un esperimento con molti gruppi o trattamenti da confrontare tra loro. Per questo tipo di analisi esistono adeguati test che mantengono per tutti i confronti pianificati un valore globale di $\alpha < 0.05$. L'argomento sarà trattato nel prossimo capitolo.

Limiti fiduciali della media

Avendo calcolato la media di un campione di n osservazioni, ci interessa sapere quanto questa media può differire dalla media non nota della popolazione, per un certo livello di probabilità scelto da noi. Per far questo cerchiamo nella tabella delle distribuzioni t la riga per gli $n-1$ gradi di libertà e la colonna per il livello di probabilità p . In corrispondenza di $n-1$ e p leggeremo un valore di t , che possiamo indicare con $t_{p,n-1}$. A questo punto possiamo dire che, ad un certo livello di probabilità p , la media della popolazione è compresa nell'intervallo media campionaria $\pm t \cdot s_m$. Questo è anche detto intervallo fiduciale o intervallo di confidenza. Attenzione: il livello di probabilità si riferisce alla probabilità di star fuori dai limiti fiduciali (la probabilità di trovare un valore di t uguale o maggiore...). Ad es. il livello di probabilità 0.05 (5%) consente di calcolare i limiti fiduciali entro i quali si troverà la media vera nel 95% dei casi, mentre nel 5% dei casi essa sarà fuori dell'intervallo. p in altre parole è il rischio che si vuole correre. In definitiva i limiti fiduciali (LF) sono quindi:

$$LF = m \pm t \cdot s_m$$

Esempio, avendo una media $m=40$, con $s_m=5$ e $n=21$, scelto il livello di probabilità di 0.05, si trova in tabella il corrispondente valore di $t=2.09$. Quindi i limiti fiduciali della media saranno:

$$LF = 40 \pm 2.09 \cdot 5 = \begin{cases} 50.45 \\ 29.55 \end{cases}$$

Con una probabilità di sbagliare 5 volte su 100, diremo che la media vera è compresa tra 29.55 e 50.45.

Grandezza del campione (metodo parametrico)

La numerosità del campione è spesso anche detta grandezza del campione (per dire numerosità, in inglese si usa il termine 'size' che in italiano si traduce con 'grandezza'). Quanti dati occorre prendere per ottenere una buona media, cioè una media rappresentativa? Spesso il problema deriva dal fatto di avere troppi o troppo pochi dati da processare. Nel primo caso, perderemmo un sacco di tempo a valutarli tutti; nel secondo caso, un sacco di soldi per viaggiare o per fare costosi esperimenti supplementari. Il problema può essere adeguatamente affrontato mediante la distribuzione t . Se infatti si parte dal presupposto di voler ottenere una media i cui limiti fiduciali siano contenuti entro ad es. il 5% del valore stesso della media, possiamo trovare un'eguaglianza tra l'espressione generale:

$$LF = m \pm t \cdot s_m$$

e la nostra opzione che i LF siano pari al 5% della media (sopra e sotto):

$$LF = m \pm 0.05 \cdot \text{media}$$

Dalle due espressioni si evidenzia che

$$t \cdot s_m = 0.05 \cdot m$$

Ma poiché sappiamo che:

$$s_m = \frac{s}{\sqrt{n}}$$

possiamo scrivere:

$$t \cdot \frac{s}{\sqrt{n}} = 0.05 \cdot m$$

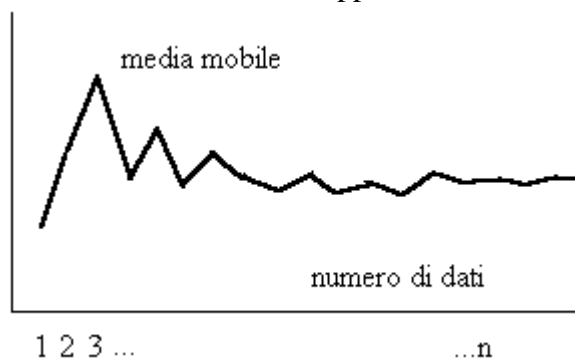
Da cui otteniamo finalmente:

$$n = \frac{t^2 \cdot s^2}{0.05^2 \cdot m^2}$$

Ovviamente, tale procedimento implica l'esecuzione di uno studio pilota per stimare in prima approssimazione la media e la deviazione standard.

Grandezza del campione (metodo grafico)

Un metodo alternativo è quello di fare uno studio pilota, controllando il comportamento della media al crescere del numero di dati. La media è di volta in volta ricalcolata (si dice appunto **media fluttuante** o **media aggiornata** o **media mobile**) dopo 2, 3, 4, 5 ... ecc. osservazioni. Succede che la media calcolata per poche osservazioni varia bruscamente. Poi si stabilizza per un semplice fatto di inerzia. A questo punto dobbiamo stabilire quale è la numerosità che consente di ottenere una media affidabile. Per questo consideriamo il primo punto del grafico oltre il quale le oscillazioni sono appena un decimo (o un ventesimo, o ancora meno) dell'intera banda di oscillazione registrata all'inizio del tracciato. Oppure consideriamo il primo punto del grafico oltre il quale le oscillazioni non superano del 10% o del 5% la media stabilizzata con la massima numerosità (i due criteri si equivalgono). Rintracciato questo punto, troviamo in ascissa il valore di n che utilizzeremo nelle successive applicazioni.



Esercizio

Costruire il grafico della media aggiornata della seguente serie di dati:

4, 15, 6, 3, 7, 8, 2, 5, 8, 16, 3, 7, 6, 4, 9, 12, 5, 6, 8, 10, 9, 3, 12, 8, 9, 6, 10



Una volta definita la media stabile utilizzando tutti i dati,

- tracciare la banda di confidenza del 10% sopra e sotto la media stabile
- valutare quale è il numero minimo di dati la cui media non esca dalla banda di confidenza tracciata