Il test di Student

F. Scotti

23 maggio 2006

Indice

1 Introduzione ai test statistici di significatività: ipotesi nulla e livello di significatività del test

Tutti i test statistici di significatività assumono inizialmente la cosiddetta *ipotesi zero*, anche chiamata *ipotesi nulla*. Quando si effettua il confronto fra due o più gruppi di dati (campioni), l'ipotesi zero prevede sempre che non esista alcuna differenza tra i gruppi riguardo al parametro considerato. Quindi se l'ipotesi nulla è valida, i campioni vengono dalla stessa popolazione e le eventuali differenze osservate nei campioni (il parametro considerato) vanno attribuite al solo caso.

Una decisione di respingere l'ipotesi nulla (presa sulla base del test statistico) è probabilmente giusta, ma potrebbe anche essere errata. La probabilità di commettere questo errore si chiama *livello di significatività* del test.

Il livello di significatività di un test può essere scelto a piacere dallo sperimentatore. Questa probabilità, chiamata anche *valore* P, di solito viene fissata ai valori tipi di 0.05 (5%) o di 0.01 (1%). Ricordiamo quindi che questa probabilità rappresenta una stima quantitativa della probabilità che le differenze osservate siano dovute al caso.

Gli statistici definiscono questo concetto in un modo più preciso dicendo:

> il valore P è la probabilità di ottenere un risultato altrettanto estremo o più estremo di quello osservato se la diversità è interamente dovuta alla sola variabilità campionaria, assumendo quindi che l'ipotesi iniziale nulla sia vera.

2 Il test di di Student

Se il parametro che andiamo a considerare dei due campioni è la media e vogliamo sapere se la eventuale differenza fra medie è significativa useremo il test di Student.

2.1 Introduzione al test di Student

Vediamo ora il contesto nel quale è nato il test di Student (Alias William Sealy Gosset, 1876-1937). La preoccupazione di Gosset era questa: "Posso usare le mie nozioni di statistica applicabili agli esperimenti con grandi campioni (quindi usare la distribuzione normale nei conti) anche con campioni di dimensioni ridotte, oppure la distribuzione che meglio descrive i fenomeni è un'altra?". Gosset era un chimico inglese assunto dalla famosa birreria Guinness di Dublino ed eseguiva analisi statistiche su campioni dei prodotti per la mansione che oggi verrebbe chiamata controllo di qualità. In generale, rilevare un campione costa sempre *tempo* e *denaro*. Per questo motivo, spesso Gosset era costretto ad usare per le sue indagini statistiche un numero ridotto di campioni.

Similmente al contesto lavorativo di Gosset, anche durante il nostro normale lavoro in laboratorio, assai raramente la mole di dati rilevati permette di inferire ipotesi sulla media campionaria $\hat{X}$ di un campione avendo la sua varianza σ^2.[1] Di solito infatti, quando non e' nota la media di una popolazione, nemmeno la sua varianza σ^2 lo è. Solitamente, quello che può essere fatto, è semplicemente sostituire nei conti la varianza vera della *popolazione* (σ^2) con quello del *campione* (S^2).

Effettuando questa sostituire occorre tenere presente che la distribuzione delle probabilità non è più fornita dalla distribuzione normale, ma da quella del t, detta *t di Student*. Pertanto è questa la distribuzione che dovremo usare se vogliamo che i nostri test sulle medie di due serie di campioni siano corretti.

2.2 La variabile aleatoria t e la forma della distribuzione di Student

Gosset si rispose da solo dimostrando che per piccoli campioni, lo scarto tra le medie dei campioni estratti dalla stessa popolazione e la media dell'universo, in rapporto all'errore standard, non è distribuito da una normale (come accadrebbe per campioni di grandezza infinita) ma bensì da una distribuzione diversa. Usando le formule al posto delle parole, possiamo scrivere la variabile casuale studiata da Gosset in questo modo:

$$t = \frac{MediaDelCampione - MediaDellaPopolazione}{ErroreStandardDelCampione} = \frac{(\hat{X} - \mu)}{S/\sqrt{n}}, \tag{1}$$

dove il numero campioni n viene chiamato da Gosset con il termine *gradi di libertà*.

La distribuzione di probabilità della variabile aleatoria t è mostrata nella figura 1. Essa è a forma di campana e simmetrica, ma con una dispersione che diminuisce con il numero di gradi di libertà. Pertanto non vi è una sola curva t ma, a differenza di quanto avviene per la distribuzione Normale, esiste una intera famiglia di distribuzioni t, una per ogni grado di libertà. Questo significa che la variabile aleatoria t ha una sua distribuzione per gli esperimenti con 5 campioni ($n = 5$), per 6 ($n = 6$), ecc. Aumentando il numero di campioni n, la distribuzione di Student tende a una distribuzione Normale (disegnata in Figura 1 con $n = \infty$).

2.3 Ipotesi per la validità del test di Student e le 4 modalità di impiego

Quando la distribuzione t di Student è applicata a test di verifica delle ipotesi, è necessario rispettare le seguenti ipotesi:

- la distribuzione dei dati deve essree una distribuzione *Normale*;

[1] Confrontare due medie avendo la loro varianza sarebbe piuttosto diretto. Ad esempio è possibile calcolare le due deviazioni standard (la radice quadrata della varianza) e controllare se le due medie distano più della metà della somma delle due deviazioni standard.

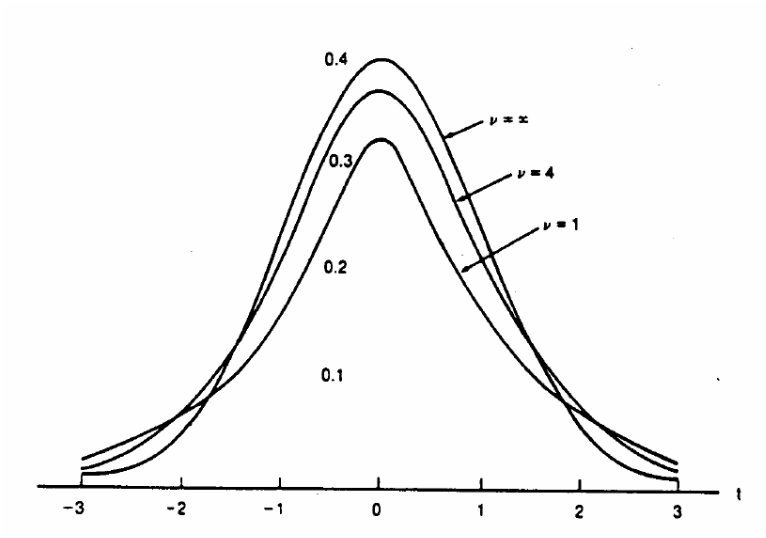

Figura 1: Distribuzione di Student. Sull'asse x abbiamo la variabile aleatoria t, mentre sull'asse y abbiamo la densità di probabilità. Le tre curve si riferiscono agli esperimenti con numero di gradi di libertà pari a 1, 4 ed ∞.

- le osservazioni devono essere raccolte in modo *indipendente*.

La seconda condizione dipende dalla modalità di organizzazione della raccolta dei dati. Ad esempio, le osservazioni non sono indipendenti se entro un gruppo di persone delle quali si misura il peso esistono più fratelli. Similmente l'indipendenza dei campioni non è rispettata se, in un esperimento sulla conducibilità elettrica di un metallo a temperature diverse, si utilizzano campioni di metallo diversi ma un campione è misurato più volte. Rispetto alla condizione di normalità la distribuzione t è robusta, ovvero rimane approssimativamente valida, anche quando le distribuzioni di dati non rispettano esattamente la condizione di normalità.

Nelle attività tecniche e di laboratorio, il test di Student è impiegato in quattro casi, ovvero per il confronto tra:

1. la media di un campione e la media delluniverso o una generica media attesa;
2. un singolo dato e la media di un campione, per verificare se possono appartenere alla stessa popolazione;
3. la media delle differenze di due campioni dipendenti con una differenza attesa;
4. le medie di due campioni indipendenti.

3 Confronto fra le medie di due campioni

Fra le 4 modalità di impiego del test di Student che abbiamo elencato, il confronto fra le medie di due campioni è la modalità più frequente nella attività di laboratorio.

3.1 Un esperimento tipico

Immaginiamo una ditta che produce birra impiegando una miscela di malti che chiameremo A. Vogliamo verificare se una nuova miscela di malti B usando gli stessi procedimenti produttivi produce una birra più gradevole. Per ottenere questa risposta occorre creare un *esperimento statistico.*

Facciamo assaggiare a N assaggiatori la birra prodotta con la miscela di malti A, poi chiederemo ad ogni assaggiatore di dare un voto alla birra da 0 a 10. A questo punto, prendiamo altri M assaggiatori e gli facciamo assaggiare la birra prodotta con la miscela di malti B.

Nell'ipotesi che gli assaggiatori ben rappresentino la popolazione di consumatori di birra, allora è ovvio che metteremo in produzione la birra fatta con la miscela con il voto medio più alto. Attenzione però, lo faremo solo quando questa media sarà significativamente diversa! La riposta verrà proprio dalla applicazione del test di Student.

Il problema è che usando pochi assaggiatori, le medie sono sempre un poco diverse. Questo accadrebbe addirittura anche nel caso nel quale faremmo assaggiare al gruppo di assaggiatori M e N la stessa miscela di malti! Il test di Student ci dice quanto questa differenza fra le medie è significativa, oppure dipende solo da fluttuazioni casuali dovute al fatto che usiamo dei campioni di dimensioni ridotte (ovvero non un numero infinito di assaggiatori).

3.2 L'ipotesi nulla nel caso in esame

In questo caso l'*ipotesi nulla* del test presuppone che *la birra fatta con i malti di tipo A o di tipo B sia mediamente buona uguale.* Esprimendo la stessa affermazione, uno statistico direbbe che l'ipotesi nulla presuppone che le due medie a confronto siano estratte dalla *stessa* popolazione (quindi si presuppone che siano identiche) Facciamo chiarezza: lo statistico intende che se potesse fare assaggiare la birra fatta con i malti A a infinite persone otterrebbe un voto medio μ_1 e se potesse far assaggiare la birra fatta con i malti B a infinite persone otterrebbe un voto medio μ_2. Se la birra fatta con i malti A ed i malti B è mediamente buona uguale allora deve essere $\mu_1 = \mu_2$, e questa, per lui, è l'ipotesi nulla.

Nel caso dell'ipotesi nulla, le differenze effettivamente riscontrate nelle medie campionarie $\overline{X}_1$ e $\overline{X}_2$ sarebbero imputabili a variazioni casuali, come effetti dovuti al campionamento, cioè alla estrazione casuale di alcuni dati da un universo teoricamente infinito, formato da valori tra loro diversi e con una distribuzione normale intorno alla loro media.

Usando il test di Student possiamo calcolare la probabilità che l'ipotesi nulla non sia vera (cioè che le medie μ_1 e μ_2 in realtà sono diverse e quindi le medie campionarie $\overline{X}_1$ e $\overline{X}_2$ sono diversi non solo per fattori casuali dovuti alla estrazione del campione). Di solito si considera il test come significativo se la probabilità è minore del 5%.

3.3 La variabile aleatoria da usare in questo caso

Nel confronto fra due medie di campioni (indipendenti e distribuiti come una normale, dicono le ipotesi del test di Student) si usa le distribuzione della variabile casuale t definita invece in questo modo

$$t = \frac{DifferenzaFraLeMedieCampionarie}{ErroreStandardDellaDifferenzaFraLeMedieCampionarie}. \tag{2}$$

La forma della distribuzione è sempre uguale a quella generale mostrata in Figura 1. Usando questa curva con i parametri corretti (che possiamo estrarre dai dati degli assaggiatori) è possibile trovare la nostra risposta alla domanda sulla significatività della differenza fra le medie.

3.4 Test di Student ad una coda ed a due code

Il test di Student parte dalla analisi della distribuzione mostrata in Figura 1 e può essere applicato in due modi:

Vogliamo sapere se le medie sono uguali o diverse → In questo caso si una il test di Student nella forma *bilaterale* (anche detta *a due code*).

Vogliamo sapere se una media è maggiore dell'altra (e posso escludere che sia minore) → In questo caso si una il test di Student nella forma *unilaterale* (anche detta *a una coda*).

Tenendo conto di questa distinzione, nel caso della birra fatta con i due malti diversi dovremo quindi usare il test di Student a due code. Anche quando ***non*** sono ***sicuro*** di poter affermare che una media possa essere solo uguale o maggiore dell'altra, ma non minore, uso sempre il test a due code.

3.5 Applichiamo il test di Student per il confronto di due medie

Immaginiamo di aver rilevato i voti degli assaggiatori delle birre ottenute con i malti A e di tipo B in una tabella come questa

```
#   A B
-------
01  7 7
02  8 8
03  9 7
04  8 8
05  8 9
06  7 8
07  6 8
08  7 9
09  8 7
10  7 6
11  7 8
12  8 6
13  8 9
14  7 8
15  8 9
16  6 8
17  8 8
18  7 7
19  6
20  7
```

Partendo dai dati è possibile effettuare il test di Student usando delle tabelle, oppure utilizzando uno strumento informatico. Noi scegliamo questa seconda possibilità.

Usando la funzione copia ed incolla possiamo incollare questi dati in un foglio Excel. A questo punto sotto la voce di menù *Tools* possiamo cliccare su *Data analysis* e scegliere a questo punto *t-Test: Two-Sample Assuming equal variances* (ovvero il test a due code). Notiamo che usando questa funzione stiamo assumento che anche che la varianza dei voti dei gradimenti degli assaggiatori abbia la stessa varianza con le due birre.

Così facendo appare una finestra di dialogo nella quale dobbiamo immettere:

- Il `range1` di celle del primo campione (la colonna dei voti del malto A: 7, 8, 9, 8, 8,...);
- Il `range2` di celle del secondo campione (la colonna dei voti del malto B: 7, 8, 7, 8, 9, ...);
- `Alpha` (il livello di significatività del test che vogliamo, ad esempio il 5%, ovvero 0,05);
- `Output range` (a partire da quale cella vogliamo stampare i risultati del test).

Il test di Student permette di confrontare medie anche di gruppi il cui numero dei campioni è diverso, proprio come nel nostro caso: nel gruppo A abbiamo 20 voti, nel gruppo B abbiamo 18 voti. La stampa dei risultati che otteniamo è piuttosto interessante:

```
t-Test: Two-Sample Assuming Equal Variances

Variable 1 Variable 2
Mean 7,35 7,777777778
Variance 0,660526316 0,888888889
Observations 20 18
```

```
Pooled Variance 0,768364198
Hypothesized Mean Difference 0
df 36
t Stat -1,502083626
P(T<=t) one-tail 0,070898582
t Critical one-tail 1,688297694
P(T<=t) two-tail 0,141797165
t Critical two-tail 2,028093987
```

Le medie sono diverse ed apparentemente ci dicono che la miscela di malti B produce una media sui voti più alta (7,78 contro 7,35). I risultati ci dicono però che non era vero che le *varianze dei dati* erano uguali (0,66 contro 0,89). Quindi riapplichiamo il test usando questa volta la funzione più corretta: *t-Test: Two-Sample Assuming Unequal variances.* Immettendo ancora gli stessi parametri `range1`, `range2`, `Alpha` e `Output range` questa volta otteniamo:

```
t-Test: Two-Sample Assuming Unequal Variances

Variable 1 Variable 2
Mean 7,35 7,777777778
Variance 0,660526316 0,888888889
Observations 20 18
Hypothesized Mean Difference 0
df 34
t Stat -1,490152837
P(T<=t) one-tail 0,072701698
t Critical one-tail 1,690924198
P(T<=t) two-tail 0,145403397
t Critical two-tail 2,032244498
```

Il parametro che a noi interessa ($P(T <= t)$) è la probabilità della ipotesi nulla (ovvero della ipotesi "*le medie sono uguali*"), ovvero 0,14. Questo valore è troppo alto, di solito si considera come accettabile un valore almeno sotto 0,05.

Cosa possiamo dire quindi delle due miscele di birre? Possiamo dire che esiste una differenza delle medie che sembra favorire la miscela di malti B, ma il test di Student considera questa differenza fra le medie come *non significativa*. Questo significa che la differenza dipende da fattori casuali e non è dimostrabile che dipenda dal fatto che la miscela di malti B sia meglio di quella A.

3.5.1 Un comune fraintendimento di cosa sia P

Molti tecnici fraintendono cosa sia veramente questa probabilità P. Se la probabilità P vale 0.03 significa che *c'e' il 3% delle possibilità di osservare una differenza fra le medie maggiore a quella trovata nei dati quando le medie delle popolazioni sono uguali (ovvero se fossimo nel caso delle birre fatte con gli stessi malti e con lo stesso procedimento)*(Affermazione A).

Si è tentati quindi di dire (sbagliando) che *c'è quindi il 97% di probabilità che la differenza fra le medie misurata rifletta una differenza reale fra le medie e che sia del 3% la probabilità che possa dipendere solo dal caso* (Affermazione B). Questa affermazione è sbagliata!

L'unica cosa che invece potremmo dire avendo trovato $P = 0.03$ è solo l'affermazione A.

3.5.2 Cosa fare quando il test è non significativo?

L'unica cosa che è possibile fare è ripetere il test aumentando il numero di campioni (ossia aumentando il numero degli assaggi). A questo punto si controlla di nuovo se le medie sono diverse e se la probabilità $P(T <= t) < 0,05$.

Negli esercizi del corso possiamo regolarci in questo modo: se il numero di dati totali supera le due centinaia in totale e la probabilità $P(T <= t)$ continua a non scendere sotto 0,05, si considerano

	Medie campionarie molto simili	Medie campionarie molto diverse
$P > 0.05$	Aggiungere Campioni (test non significativo)	Aggiungere Campioni (test non significativo)
$P \leq 0.05$	Test significativo, ma il farmaco influisce poco	Test significativo ed il farmaco influisce molto

Tabella 1: Esempio di impiego del test di Student

le medie come non significativamente diverse. L'esperienza mostra invece che, se le medie sono significativamente diverse, bastano poche decine di dati per vedere la probabilità $P(T <= t)$ andare ben al di sotto del valore 0,05 che noi abbiamo scelto come soglia. La tabella 1 riassume come potremmo comportarci in un caso pratico.

3.6 Conclusioni

Per tutti questi motivi, quando durante la nostra attività lavorativa dovremo confrontare fra loro due medie di valori ottenuti da esperimenti, non guarderemo semplicemente se le medie sono diverse, ma applicheremo sempre anche il test di Student per vedere se la differenza è significativa o meno.