ALCUNI ELEMENTI DI VERIFICA DI IPOTESI STATISTICHE

Con applicazioni nell'ambiente statistico R

Vittorio Colagrande

Altro problema dell'inferenza è quello della verifica di ipotesi: si ipotizza su una caratteristica di una popolazione oggetto di studio e "si verifica", sulla base di osservazioni campionarie, se l'ipotesi fatta va confermata o smentita.

L'ipotesi formulata in relazione al problema è detta ***ipotesi nulla*** e si indica H_0; le situazioni non contemplate da H_0 costituiscono l'***ipotesi alternativa***, usualmente indicata con H_1.

L'ipotesi nulla deve essere formulata in termini quantitativi e la strategia di verifica, facendo riferimento ad una opportuna *statistica campionaria*, porta a determinare la probabilità che, supposta vera l'ipotesi, il risultato campionario che si ottiene si sia prodotto "per effetto del caso".

Esempio. Si vuole saggiare l'ipotesi (H_o) che la distribuzione della statura degli studenti maschi dell'Università di Chieti abbia media $\mu = 175$ cm. Si calcola la media $\bar{x}$ delle stature di un campione di studenti (per es. 700) e si confrontano $\bar{x}$ e 175. All'uopo si calcola il valore della statistica adatta a descrivere la distribuzione dei dati (la media campionaria) e si associa ad esso un valore di probabilità, attraverso il quale si decide se respingere o meno l'ipotesi nulla.

Esempio illustrativo. Si vuole verificare la maggiore efficacia di un nuovo farmaco nel trattamento della cefaléa contro un altro da tempo impiegato. Per misurare l'efficacia si può far riferimento al rapporto tra numero di pazienti migliorati con ciascuno dei due farmaci ed il numero totale di pazienti trattati. Se è noto (attraverso dati storici, ad esempio) che il farmaco impiegato da tempo ha una percentuale di migliorati p_o, l'obiettivo dello studio è quello di verificare se la percentuale p di migliorati con il nuovo farmaco è maggiore di p_o. *L'ipotesi nulla* H_0 è che la percentuale dei pazienti migliorati col nuovo farmaco *non sia diversa* da quella dei pazienti migliorati col vecchio farmaco. L'*ipotesi alternativa* H_1 è che il nuovo farmaco è più efficace, in sostanza: $p > p_o$. Si progetta, allora, un esperimento per saggiare l'ipotesi nulla su un campione casuale della popolazione oggetto di studio; la statistica campionaria è definita dal rapporto **p** tra il numero di migliorati e il numero di elementi campionari ed è nota la sua distribuzione nel caso sia vera l'ipotesi nulla $p = p_o$. La logica della verifica si basa sul fatto che se H_0 fosse vera, risulta *poco probabile* che il rapporto p sia *molto più grande* di un particolare valore "soglia" (critico) p_c. Se, tuttavia, l'esperimento campionario fa ottenere un risultato $p \geq p_c$, allora si è portati a rifiutare l'ipotesi nulla a favore di quella alternativa. Si tratta naturalmente di fissare il valore soglia p_c e ciò viene fatto in riferimento alla probabilità $Prob(p \geq p_c)$ calcolata sulla base della distribuzione nota della statistica campionaria **P**. Tale probabilità esprime l'errore che si è disposti a commettere rifiutando l'ipotesi nulla e, di conseguenza, accettando come vera l'ipotesi alternativa di maggiore efficacia del nuovo farmaco rispetto al preesistente.

Una strategia per la verifica di ipotesi viene comunemente indicata come *test di ipotesi* e conduce alle alternative: si rifiuta l'ipotesi nulla o non la si rifiuta. In genere, l'ipotesi nulla H_0 fa riferimento ad una situazione che si vorrebbe negare (assenza di efficacia di un trattamento, assenza di un cambiamento prima e dopo un intervento, assenza di differenza tra due gruppi di individui), mentre la sua alternativa, l'ipotesi H_1, è proprio quella circostanza che si vorrebbe fosse vera. Si pensi al precedente esempio: l'ipotesi nulla è la non maggiore efficacia del nuovo farmaco rispetto al vecchio, mentre tale maggiore efficacia è contemplata dall'ipotesi alternativa.

Nel prendere la decisione intorno all'ipotesi H_0 è possibile incorrere in due tipi di errore:

- l'ipotesi nulla è vera, ma viene rifiutata (***errore di I tipo***)
- l'ipotesi nulla è falsa, ma non è rifiutata (***errore di II tipo***).

Tali errori sono legati al fatto che le decisioni derivate dalla verifica si basano su dati campionari; così, ad esempio, una grande differenza riscontrata tra quanto ipotizzato e il risultato campionario potrebbe essere un puro effetto del caso e, quindi, il campione estratto essere uno di quelli "sfavorevoli".

Si comprende allora che le decisioni relative alla verifica sono soggette ad incertezza e che è opportuno esprimere tale incertezza con valori di probabilità.

In realtà la *probabilità di commettere un errore di I tipo* è espressa dal livello di probabilità:

$$\alpha = \text{Probabilità(rifiutare } H_o\text{, supposto } H_0 \text{ vera)} = \text{Prob(rifiuto } H_0 | \ H_0 \text{ vera)}$$

La *probabilità di commettere un errore di II tipo* è data da:

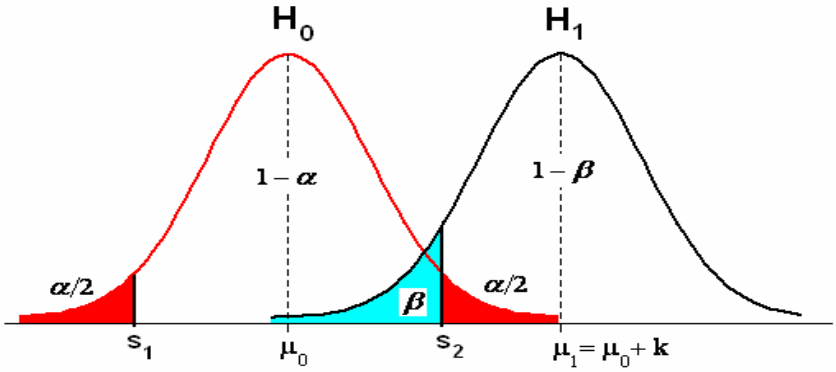
$$\beta = \text{Probabilità(non rifiutare } H_o\text{, supposto } H_1 \text{ vera)} = \text{Prob(non rifiuto } H_0 | \ H_1 \text{ vera)}.$$

Le opzioni sono illustrate nella tabella seguente:

DECISIONE STATISTICA DOPO IL TEST	IPOTESI VERA	
	H_0	H_1
Accettazione (non rifiuto) di H_0	Conclusione esatta Probabilità = $1-\alpha$	Errore di II tipo Probabilità = β
Rifiuto di H_0	Errore di I tipo Probabilità = α	Conclusione esatta Probabilità = $1-\beta$

La probabilità α, detta ***livello di significatività***, deve essere un valore fissato a priori per non essere tentati di far prevalere la tesi più "comoda". In genere si fa riferimento a valori "molto piccoli" di α, pari al 5%, all'1% o all'1‰. Un livello di significatività del 5%, ad esempio, indica che la probabilità di rigettare l'H_0, quando è vera, a causa di fluttuazioni casuali del campionamento è del 5%, ovvero ci sono 5 possibilità su cento di prendere una decisione errata. Se poi si ritiene molto importante non respingere come falsa un'ipotesi vera, si può considerare una probabilità dell'1% o dell'1‰.

I livelli di probabilità α e β vanno determinati in riferimento alla distribuzione campionaria della *statistica S* presa in esame; ad esempio se la verifica riguarda una media si può considerare la *statistica-test* media campionaria $\overline{X}$, se si tratta di proporzioni la *statistica-test* è la proporzione campionaria *P*. La statistica S presenta opportune distribuzioni campionarie, sia supponendo vera l'ipotesi nulla H_o che supponendo vera l'ipotesi alternativa H_1. La figura che segue rappresenta le distribuzioni campionarie della media sotto (cioè supposta vera) l'ipotesi nulla H_o: media = μ_0 e sotto quella alternativa H_1: media = μ_0+k. In essa sono indicate anche delle aree che misurano le probabilità degli errori di I e II tipo; in particolare l'area α (somma delle due aree indicate con $\alpha/2$) è relativa al rifiuto dell'ipotesi nulla: questa ***non*** è accettata se il valore della media su un campione casuale preso in esame risulta inferiore al valore s_1 oppure superiore a s_2, che rappresentano i valori *critici* della statistica-test S.

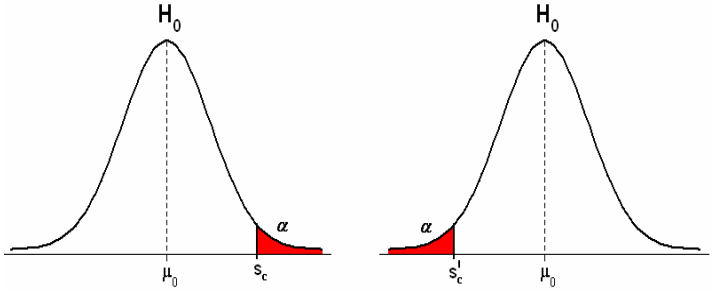


Il livello di significatività α, allora, divide l'area della distribuzione campionaria relativa ad H_0 in due regioni: la ***regione di rigetto*** e la ***regione di accettazione*** dell'ipotesi nulla. Nell'esempio di figura la

regione di rigetto è relativa alle aree $\alpha/2$ e quella di accettazione all'area restante $1-\alpha$ al di sotto della curva di distribuzione della media campionaria, ottenuta sotto l'ipotesi nulla.

La figura precedente illustra la situazione relativa ad un test *bidirezionale*: la regione di rifiuto dell'ipotesi nulla è individuata dai valori della statistica (media campionaria) maggiori o uguali a s_2 oppure minori o uguali a s_1, essendo i due valori critici individuati dalle: $Prob(S \geq s_2|H_0\ vera)=\alpha/2$ e $Prob(S \leq s_1|H_0\ vera) = \alpha/2$.

L'esempio illustrativo iniziale, invece, si riferisce ad un test *unidirezionale*: la regione di rifiuto di H_0 è definita dai valori della statistica (proporzione campionaria) maggiori o uguali al valore critico s_c (nell'esempio p_c), che viene individuato sulla base della $Prob(S \geq s_c|H_0\ vera)=\alpha$.
Naturalmente un test unidirezionale può portare anche a regioni di rifiuto in cui la statistica-test S è minore o uguale ad un dato valore critico $s_c^{|}$: $Prob(S \leq s_c^{|}\ |H_0\ vera)=\alpha$. La figura a lato rappresenta le zone di rifiuto dell'ipotesi nulla per test unidirezionali relativi alla media.



Nel processo di verifica è importante stabilire preliminarmente se il test deve essere unidirezionale (ad una coda: *one-sided*) o bidirezionale (a due code: *two-sided*).

Altro concetto importante è quello di ***potenza di un test***, cioè la probabilità $1-\beta$ di respingere l'ipotesi H_0 quando è falsa. Quando si è in grado di minimizzare l'errore di II tipo, allora si può dire che un test è più potente e un buon test dovrebbe avere contemporaneamente un valore basso di α e uno basso di β. Tuttavia, come si può intuire visivamente anche dalla prima figura, non è possibile ridurre contemporaneamente i due errori, in quanto essi sono correlati.

Nell'inferenza statistica si può disporre di due tipi di test: i ***parametrici*** ed i ***non parametrici***.

I primi sono test che, basandosi sull'uso della curva normale, della t di Student, etc, richiedono essenzialmente che la distribuzione del carattere analizzato nella popolazione di riferimento sia normale.

I secondi, invece, sono utilizzati quando non è possibile supporre una data forma per la distribuzione della popolazione e/o ci si trova in presenza di campioni di piccole dimensioni.

Prima di procedere vengono ora puntualizzati i passi da seguire per porre in atto la strategia della verifica di ipotesi:

- Stabilire quale deve essere l'ipotesi nulla e quale quella alternativa.
- Scegliere la statistica-test.
- Determinare la distribuzione campionaria della statistica-test, nel caso sia vera l'ipotesi nulla; questa permette di calcolare la probabilità di un dato risultato campionario.
- Si fissa il livello di significatività e si individua la regione di rifiuto dell'ipotesi nulla; più è piccola tale regione e minore è il rischio che si corre nel respingere Ho. La regione di rifiuto è formata da tutti i risultati che hanno la probabilità di verificarsi non maggiore di α, qualora fosse vera l'ipotesi H_0. Quindi il livello α determina un'area in cui cadono i risultati poco probabili e difficilmente riscontrabili nelle realtà, sempre che si supponga vera H_0.
- Si estrae un campione casuale dalla popolazione analizzata.
- Si calcola il valore della statistica-test sui dati campionari e si confronta con il valore critico del test (che individua la regione di rifiuto di H_0) in relazione al livello di significatività α prescelto. Generalmente i valori critici del test sono tabulati (vedi Tavole della distribuzione normale e della distribuzione t di Student).

Si decide di respingere l'H_0 e accettare H_1 quando il valore della statistica-test calcolato sui dati campionari cade nella zona di rifiuto dell'ipotesi H_0; in caso contrario l'ipotesi H_0 non viene rigettata. Se si respinge l'ipotesi nulla si dice che *il test è significativo al livello α.*

È necessario a questo punto far presente che per la significatività statistica spesso, nella pratica, si fa riferimento ad un valore di probabilità, detto ***p-value***, che, in qualche modo, quantifica la "forza dell'evidenza" contro l'ipotesi nulla H_0 (a favore dell'alternativa) espressa dai dati osservati su un campione.

In sostanza il ***p-value*** esprime quanto sia plausibile che i dati osservati si ottengano essendo vera l'ipotesi nulla: un p-value grande esprime evidenza sperimentale a favore dell'ipotesi nulla, mentre un suo valore piccolo un'evidenza a favore dell'ipotesi alternativa.

Ad esempio, se la verifica fa riferimento ad una data statistica S e tale statistica assume il valore s_{calc} sui dati di un campione casuale, il p-value può essere determinato attraverso le: $Prob(S \geq s_{calc} \mid H_0 \text{ vera})$ oppure $Prob(S \leq s_{calc} \mid H_0 \text{ vera})$ per test unidirezionali, mentre per test bidirezionali va calcolata la probabilità $Prob(S \leq -s_{calc} \text{ oppure } S \geq s_{calc} \mid H_0 \text{ vera})$.

A scanso di equivoci, va subito specificato che un p-value basso ***non*** significa che è bassa la probabilità che l'ipotesi nulla sia vera, ma soltanto che è più ragionevole ammettere che i dati osservati siano stati ottenuti essendo vera l'ipotesi alternativa piuttosto che l'ipotesi nulla.

Sulla base del p-value si decide sulla significatività del test. In particolare, valori inferiori al 10% ma superiori o uguali al 5% indicano una debole evidenza sperimentale contro l'ipotesi nulla e a favore dell'alternativa; valori inferiori al 5% e superiori o uguali all'1% portano a ritenere il test significativo; si parla di test abbastanza significativo in presenza di un p-value tra l'1% e l'1$^{o}/_{oo}$, mentre il test è considerato molto significativo per probabilità inferiori all'1$^{o}/_{oo}$. Per tutti gli altri valori non c'è evidenza per il rifiuto dell'ipotesi nulla. Tuttavia si fa semplicemente rilevare che nell'analisi statistica relativa ad un dato parametro è opportuno affiancare al p-value un *intervallo di confidenza* del parametro stesso, per decidere più adeguatamente in merito al rifiuto dell'ipotesi nulla o sull'opportunità di ulteriori approfondimenti dello studio che si sta conducendo.

I test di verifica di ipotesi possono essere applicati ad un solo campione oppure a più campioni.

I primi sono, in realtà di scarsa utilità perché spesso non si conosce il valore della media μ o della proporzione π della popolazione di riferimento.

Quando, invece, si pongono a confronto due o più campioni, si verifica, ad esempio, la provenienza di due campioni da un'unica popolazione oppure si confronta un gruppo di controllo con un gruppo sperimentale.

I test di verifica di ipotesi su un campione sono tuttavia utili per introdurre alcune caratteristiche comuni a tutti i tipi di test e possono costituire, quindi, il presupposto per lo studio dei confronti tra due e più campioni.

Verifica parametrica di ipotesi su un campione

La trattazione verrà sviluppata in riferimento alla verifica di ipotesi per medie e proporzioni.

Se è nota la deviazione standard della popolazione, è possibile ricorrere alle proprietà della distribuzione normale sia quando la dimensione **n** del campione è "abbastanza grande" (anche per n>30), sia quando si ha un campione di numerosità **n** qualunque, purché il carattere preso in esame si distribuisca nella popolazione secondo una distribuzione normale. In tal caso per la verifica di ipotesi si utilizza il ***TEST Z.***

Se non si conosca la varianza della popolazione e si ha un campione di piccole dimensioni, si sceglierà il ***TEST t DI STUDENT.***

Esempio 1. Si consideri la popolazione degli studenti fumatori di una facoltà di Medicina. Si estragga un campione di n = 100 fumatori e si supponga di rilevare che i maschi fumatori sono 60 (percentuale p=0.60). Si verifichi, al livello di significatività α=0.05, se la percentuale di maschi fumatori è superiore a quella di femmine fumatrici.

L'ipotesi nulla è H_0: $\pi = 0.50$ e l'ipotesi alternativa H_1: $\pi > 0.50$ (test unidirezionale).

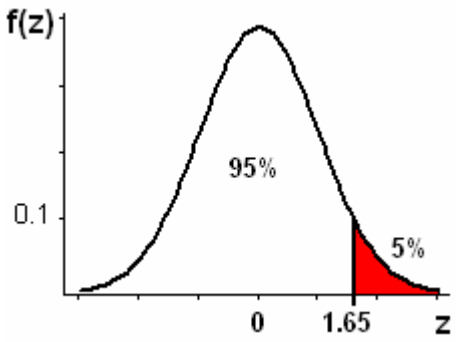
Poiché nel caso di n > 30 la distribuzione di una proporzione è approssimativamente normale, si può utilizzare la statistica-test Z:

$$Z = \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}},$$

che per i dati campionari assume il valore: $z = (0.6 - 0.5)/\sqrt{0.0025} = 2$.

Il valore critico z_c si determina sulla *Tavola (Tavola 1) della normale standard* (vedere Appendice *Dispensa Distribuzione Normale*) in corrispondenza alla probabilità Prob($Z \geq z_c$) = 5%:

$$\text{Prob}(Z \leq z_c) = 1 - \text{Prob}(Z > z_c) = 1 - 0.05 = 0.95 \quad \Rightarrow \quad z_c = 1.65.$$



Poiché z = 2 calcolato è maggiore del valore critico z_c, si decide di rifiutare l'ipotesi nulla: nella popolazione degli studenti di Medicina, sulla base dei risultati campionari, non è verosimile che il sesso *non* influisca sull'essere fumatori.

Esempio 2. Si supponga di voler verificare se il livello di colesterolemia, riscontrato su un campione casuale di n=25 soggetti, sia significativamente diverso (maggiore) dal livello medio in soggetti normali pari a μ_0 =210 mg/dl. È noto che nella popolazione di riferimento il livello di colesterolemia è distribuito secondo la curva normale. Nel campione il valore medio di colesterolemia è risultato pari a $\bar{x}$ =270 mg/dl e la deviazione standard di s = 79 mg/ml. Verificare, al livello di significatività α = 1%, se la differenza del campione sia dovuta al caso o a significative differenze sistematiche.

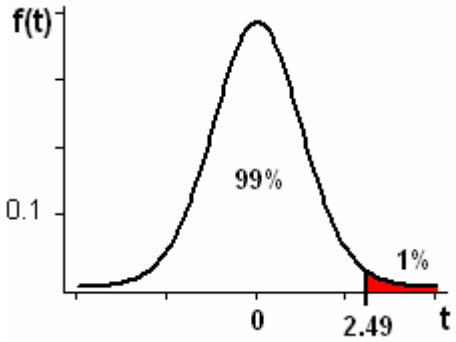
L'ipotesi nulla è data da $H_0: \ \mu = \mu_0$ e quella alternativa $H1: \ \mu > \mu_0$ (test unidirezionale).

Poiché la popolazione si distribuisce normalmente, il campione è estratto casualmente, σ^2 è ignoto e n<30 si sceglie il test t di Student. Il numero di gradi di libertà (ν) sono determinati dalla numerosità del campione meno 1: $\nu = n-1$ e, nell'esempio, ci sono (25−1) = 24 gradi di libertà.

Nella *Tavola (Tavola 2) della distribuzione t* (vedere Appendice *Dispensa sul Campionamento*), in corrispondenza di ν=24 e per un α = 0.01 si trova il valore critico $t_c = t_{\alpha,\nu}$ = 2.49 che delimita l'area di rigetto.

Il valore di t sul campione è dato da:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{60}{15.8} = 3.79.$$



Poiché il valore empirico di t = 3.79 > 2.49, con la probabilità dell'1% di commettere un errore di I tipo, si decide di respingere l'ipotesi nulla e di concludere che i soggetti del campione appartengono ad una popolazione di individui con valori di colesterolemia superiori alla norma.

Per il calcolo del ***p-value*** si può far ricorso all'ambiente statistico **R**:

```
> pt(3.79,24,lower.tail=F)
  [1]  0.0004471033
```

e, dunque, *p-value* = Prob($t \geq 3.79$, supposto vera H_0) =0.00045. Come si vede, il *p-value* è molto piccolo (decisamente inferiore a 0.001), pertanto si è in presenza di una molto forte evidenza sperimentale contro l'ipotesi nulla, a favore dell'ipotesi alternativa e il test è molto significativo.

Esempio 3. In una popolazione di donne in gravidanza è noto che il livello di glicemia nel sangue si distribuisce approssimativamente secondo una curva normale con media μ_0 = 78.3 mg/100ml. In un campione di n = 100 donne gravide si rileva un livello medio di glicemia $\bar{x}$ = 83.5 mg/100ml con una deviazione standard s = 13.5 mg/100ml. Ci si chiede se la differenza riscontrata tra campione e popolazione è dovuta "semplicemente al caso" (ovvero è stato considerato casualmente un campione con livelli di glicemia più elevati), oppure le donne appartenenti al campione presentano "scompensi" glicemici dovuti a cause diverse dall'errore di campionamento e, quindi, non appartengono alla popolazione di riferimento.

L'ipotesi nulla è H_0: $\mu = \mu_0$ e l'ipotesi alternativa H1: $\mu \neq \mu_0$ (si decide di scegliere un test bidirezionale).

Poiché la popolazione si distribuisce normalmente, il campione è estratto casualmente e la sua numerosità è elevata, per la verifica si può far riferimento alla distribuzione normale.

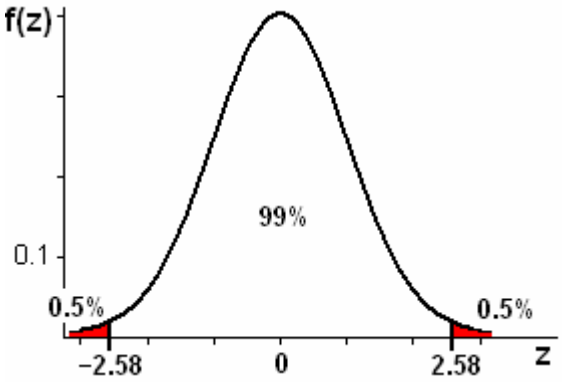
Assumendo un livello di significatività α = 1%, sulla *Tavola 1* va ricercato quel valore critico z_c per il quale:

$$\text{Prob}(Z \leq z_c) = 1-\alpha/2 = 1-0.01/2 = 0.995 \Rightarrow z_c = 2.58.$$

e la regione di rifiuto dell'ipotesi nulla è individuata dai valori $z \geq z_c$ oppure dai valori $z \leq -z_c$.

Il valore di z calcolato sul campione è dato da:

$$z = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{6.2}{1.35} = 4.59$$



Poiché il valore empirico di z = 4.59 > z_c, con una probabilità dell'1% di commettere un errore di I tipo, si decide di respingere l'ipotesi nulla e di concludere che le donne del campione appartengono ad una popolazione con valori di glicemia diversi dalla popolazione presa in esame.

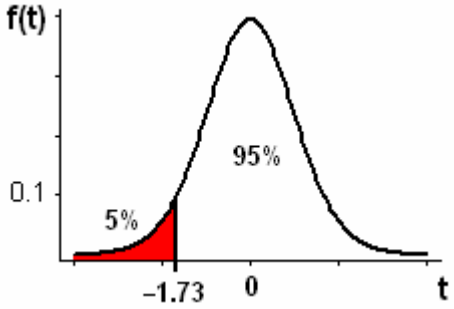
Il ***p-value*** si ottiene in **R** attraverso: *> 2*pnorm(4.59,lower.tail=F)*, ottenendo un valore pari a $4.4 \cdot 10^{-6}$ che conferma la significatività (molto forte) del test.

Esempio 4. Si consideri una popolazione di N= 500 ipertesi, per la quale le pressioni arteriose massime sono distribuite normalmente con una media pari a μ = 180 mmHg. Si consideri un campione estratto casualmente di n =20 pazienti a cui è somministrato un nuovo preparato contro l'ipertensione, ottenendo una media di pressione arteriosa pari a $\bar{x}$ = 160 mmHg, con una deviazione standard s = 40 mmHg. Si verifichi l'ipotesi che il preparato non sia efficace, contro l'ipotesi alternativa che lo sia, al livello di α= 0.05.

L'ipotesi nulla è H_0: $\bar{x} = \mu$ e l'ipotesi alternativa H_1: $\bar{x} < \mu$ (test unidirezionale).

Poiché σ è ignoto, il campionamento è casuale (senza reimbussolamento) e n < 30 si sceglie il test t di Student:

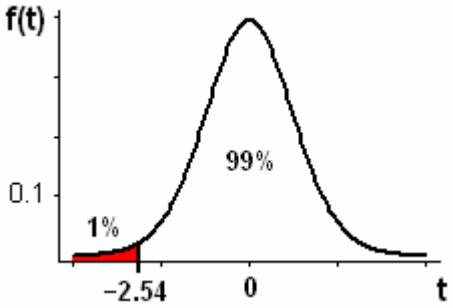
$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} = \frac{160-180}{\frac{40}{\sqrt{20}}\sqrt{\frac{500-20}{499}}} = -2.28$$



Si leggono i valori critici di t sulla *Tavola 2*. In corrispondenza di α = 0.05 e con ν =19 gradi di libertà, risulta un t_c = 1.73, per cui la zona di rifiuto è $t \leq -1.73$.

Essendo il t calcolato sul campione ($t = -2.28$) inferiore al valore critico $t_c = -1.73$, si decide di respingere l'ipotesi nulla al livello del 5% e di concludere che, sulla base dei risultati campionari, non è verosimile che gli ipertesi trattati abbiano in media una pressione arteriosa uguale a quella degli ipertesi non trattati. Calcolando direttamente il *p-value* = Prob($t \leq -2.28$, supposto vera H_0) = 0.02 si è portati al rifiuto dell'ipotesi nulla.

Se si considerasse, invece, un livello di significatività $\alpha = 0.01$, si avrebbe un valore critico $t_c = 2.54$ e una regione di rifiuto data da $t \leq -2.54$; dal momento che il valore calcolato del test t non appartiene alla regione di rifiuto, non si ritiene di avere elementi per asserire, su base statistica, che gli ipertesi trattati hanno in media una pressione arteriosa inferiore a quella degli ipertesi non trattati. Il *p-value* di 0.02, superiore a 0.01, conferma quanto detto in merito alla *non* significatività del test.



Verifica parametrica di ipotesi su due campioni

Si affronterà la verifica di ipotesi per la media.

Le situazioni più ricorrenti non riguardano il confronto tra media campionaria e media della popolazione, bensì il confronto tra due medie campionarie $\bar{x}_1$ e $\bar{x}_2$.

L'ipotesi nulla è data da:

H_0: $\mu_1 = \mu_2$

cioè i due campioni appartengono effettivamente alla stessa popolazione con media $\mu = \mu_1 = \mu_2$, oppure i due campioni sono diversi, nelle medie campionarie, soltanto per differenze casuali.

L'ipotesi alternativa può essere formulata come:

H_1: $\mu_1 \neq \mu_2$ (test bidirezionale), oppure H_1: $\mu_1 > \mu_2$ ($\mu_1 < \mu_2$) (test unidirezionale),

supponendo l'esistenza di una differenza reale tra le due medie μ_1 e μ_2.

La direzionalità del confronto (test unidirezionale o bidirezionale) è insita nella natura dell'esperimento considerato, poiché da essa dipende le distribuzioni delle probabilità che portano alla verifica di ipotesi.

La verifica, anche in questo caso, viene effettuata facendo ricorso al test *z* o il test ***t*** di Student.

Nel caso di ***due campioni indipendenti*** si consideri il seguente esempio.

Esempio 5. Ad un esame di statistica medica un campione di 30 studenti che hanno frequentato le lezioni, fa rilevare un voto medio di 27, un altro campione di 20 studenti che non hanno frequentato, evidenzia come voto medio 23; le varianze sono rispettivamente 9 e 8.5. Si verifichi l'ipotesi che la partecipazione alle lezioni non influisce sul voto.

Si indichino con μ_1 e μ_2 i valori medi (incogniti) dei voti, rispettivamente, degli studenti che hanno frequentato e di quelli che lo non hanno fatto.

L'ipotesi nulla è:

H_0: $\mu_1 = \mu_2$

cioè la frequenza alle lezioni non influisce sul voto.

L'ipotesi alternativa è che la frequenza influisca positivamente sul voto, ossia

H_1 : $\mu_1 > \mu_2$

Si consideri che la distribuzione dei voti sia normale. La statistica test da utilizzare è

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

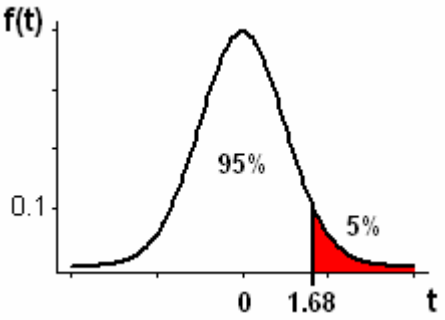
essendo la deviazione standard s_p dei due campioni raggruppati (*pooled*) data da:

$$S_p^2 = \frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{n_1 + n_2 - 2}.$$

La statistica considerata si distribuisce secondo una t di Student con n_1+n_2-2 gradi di libertà. Poiché l'ipotesi alternativa prevede che il voto dei frequentanti sia maggiore di quello dei non frequentanti, il test dovrà essere condotto sulla coda di destra: la regione critica sarà quella in cui t assume valori superiori al valore critico con $\alpha = 0.05$. Con l'utilizzo della *Tavola 2* si ottiene che il valore critico della t in corrispondenza di $\alpha = 0.05$ e 30+20−2 = 48 gradi di libertà è pari a $t_c = 1.68$.

Per i due campioni presi in esame si ha $s_p^2 = \frac{9 \cdot 29 + 8.5 \cdot 19}{48} = 8.80$ da cui:

$$t = \frac{27 - 23}{2.96\sqrt{\frac{1}{30} + \frac{1}{20}}} = 4.67,$$



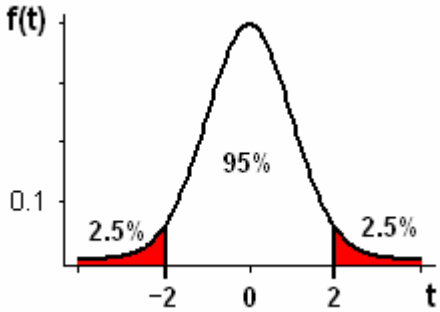
valore superiore a 1.68 e che, pertanto, cade nella zona di rifiuto dell'ipotesi nulla: la frequenza alle lezioni sembra portare, con un livello di probabilità di errore del 5%, a voti medi più altri rispetto alla non frequenza. Il calcolo del *p-value*= Prob($t \geq 4.67$, supposto vera H_0) $=1.2 \cdot 10^{-5}$ conferma pienamente quanto detto.

Esempio 6. Ad un campione di $n_1 = 30$ donne diabetiche viene somministrato un nuovo farmaco A a base di insulina e ad un secondo campione di $n_2 = 25$ donne affette dalla stessa patologia un farmaco B da tempo impiegato per ridurre i livelli glicemici nel sangue. Dopo i trattamenti, i livelli medi di glicemia sono risultati pari a $\overline{x}_1 = 83.5$ mg/100ml per il primo campione e a $\overline{x}_2 = 95.6$ mg/100ml per il secondo; le deviazioni standard risultano pari a $s_1 = 13.5$ mg/100ml e $s_2 = 17.0$ mg/100ml. Ci si chiede se la differenza riscontrata fra i livelli medi di glicemia nei due campioni sia attribuibile al diverso tipo di farmaco oppure al "caso". In sostanza si vuole verificare se le popolazioni da cui provengono i due campioni hanno gli stessi livelli medi di glicemia oppure se tale ipotesi va rigettata.

L'ipotesi nulla è: H_0: $\mu_1 = \mu_2$ e come alternative si opta per un test bidirezionale: H_1: $\mu_1 \neq \mu_2$.

La statistica test è la stessa dell'esercizio precedente e si considera un livello di significatività $\alpha = 0.05$. L'ipotesi alternativa prevede che le medie siano diverse e, quindi, il test è a due code: la regione di rifiuto dell'ipotesi nulla è individuata dai valori $t \geq t_c$ oppure dai valori $t \leq -t_c$, essendo t_c quel valore critico per il quale:

$$\text{Prob}(t \geq t_c) = \alpha/2 = 0.025 \Rightarrow t_c = 2.0.$$



come risulta dalla *Tavola 2* in corrispondenza del livello $\alpha/2 = 0.025$ e di $\nu = 30+25-2 = 53$ gradi di libertà.

A questo punto va calcolata la statistica t sui valori campionari:

$$s_p^2 = (182.25 \cdot 29 + 289 \cdot 24)/53 = 230.6 \quad \text{e} \quad t = (83.5 - 95.6)\Big/(15.2 \cdot \sqrt{1/30 + 1/25}) = -2.94$$

e, dato che il valore calcolato di t appartiene alla regione di rigetto dell'ipotesi nulla, si è portati a concludere che, sulla base dei risultati campionari, l'effetto dei due farmaci sembra diverso. Risultato confermato dal *p-value* Prob($t \leq -2.94$ oppure $t \geq 2.94$, supposta vera H_0) = 0.005.

Nel caso di ***due campioni dipendenti*** i dati sono appaiati:

- ogni osservazione di un campione è accoppiata con una e una sola osservazione dell'altro campione;
- i due gruppi hanno sempre lo stesso numero di dati;
- si mira a creare il massimo di omogeneità entro ogni coppia e il massimo di eterogeneità tra le coppie.

Si possono avere anche dati auto-appaiati: ogni soggetto serve come controllo di se stesso e i dati sono ricavati dagli stessi individui in momenti diversi (es. confronto tra i livelli di pressione rilevati nello stesso gruppo di individui sia in condizioni normali che dopo uno stress, confronti prima-dopo riferiti agli stessi individui).

Tecnicamente il confronto è semplice: l'analisi è ridotta alla sola serie delle differenze tra le misure di ciascuna coppia.

L'ipotesi nulla è data da:

H_0 : la media delle differenze è 0 ($\delta = 0$);

l'ipotesi alternativa H_1 può essere:

H_1: $\delta \neq 0$ (test bidirezionale), oppure H_1: $\delta > 0$ ($\delta < 0$) (test unidirezionale).

La statistica-test è definita dalla:

$$t = \frac{\overline{D} - \delta}{S/\sqrt{n}},$$

essendo $\overline{D}$ la media delle differenze, δ la differenza attesa (spesso, ma non necessariamente 0), S la deviazione standard delle differenze, n il numero di coppie di dati, corrispondente al numero delle differenze e $S/\sqrt{n}$ l'errore standard della media delle differenze.

Esempio 7. In un campione di 13 pazienti affetti da virus di Epatite C è stata misurata la transaminasi GPT (U/L) prima e dopo un trattamento con interferone.

Pz.	prima	dopo	Differenza d
1	56	68	12
2	310	25	-285
3	172	90	-82
4	457	29	-428
5	74	50	-24
6	66	37	-29
7	45	50	5
8	71	27	-44
9	42	44	2
10	321	41	-280
11	96	34	-62
12	42	44	2
13	61	22	-39
Media			-96.31
Deviazione Standard			140.79

Ci si chiede se il trattamento abbia significativamente determinato le differenze di valori di GPT.

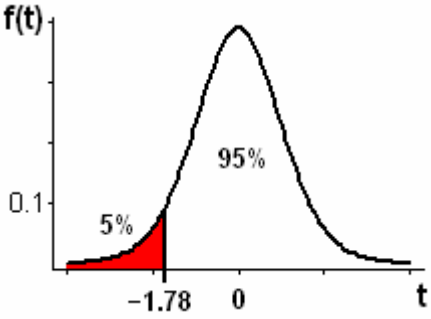
Le ipotesi sono:

$$H_0: \delta = 0; \qquad H_1: \delta \neq 0$$

Ad un livello di significatività $\alpha = 0.05$, il valore critico della distribuzione t di Student (*Tavola 2*) per $\nu = 13-1=12$ è dato da $t_c = 2.18$. Il valore calcolato di t:

$$t = \frac{-96.31}{140.79/\sqrt{13}} = -2.47$$

appartiene alla regione di rifiuto dell'ipotesi nulla (zona $t \leq -2.18$) e quindi la probabilità che la differenza riscontrata sia casuale è $p < 0.05$ (*p-value*=0.03). Si è in presenza di una "moderata" indicazione contro l'H_0: si conclude, con una probabilità di errore del 5%, che il trattamento ha determinato una variazione statisticamente significativa dei valori di AFT. Va osservato, poi, che per una verifica più attenta andrebbe considerata l'ipotesi alternativa H_1: $\delta < 0$, procedendo con un test a una coda. Sempre al livello $\alpha = 0.05$, il valore critico della t di Studente per $\nu = 12$ è pari a $t_c = 1.78$ ed essendo il valore calcolato $t = -2.47 < -1.78$, si è maggiormente portati a rifiutare l'ipotesi di non efficacia del trattamento.



Di seguito si riporta una sessione di analisi dei dati dell'esempio con l'utilizzo dell'ambiente **R**.

```
> x=c(56,310,172,457,74,66,45,71,42,321,96,42,61)
> y=c(68,25,90,29,50,37,50,27,44,41,34,44,22)
> t.test(y,x,mu=0,paired=T,alternative="two.sided")

  Paired t-test
    data:  y and x
     t = -2.4664, df = 12, p-value = 0.02969
     alternative hypothesis: true difference in means is not equal to 0

> t.test(y,x,mu=0,paired=T,alternative="less")

  Paired t-test
    data:  y and x
     t = -2.4664, df = 12, p-value = 0.01484
     alternative hypothesis: true difference in means is less than 0.
```

Verifica non parametrica di ipotesi su due campioni [1]

Il test Z e quello t, come altri test parametrici, presuppongono che i campioni si riferiscano a caratteri con distribuzione tendente a quella normale. Quando questo assunto non è rispettato o anche la numerosità campionaria è particolarmente ridotta (cosa molto frequente in ambito biomedico) si fa ricorso ai test non parametrici. Questi vanno utilizzati quando sussistono seri dubbi sull'attendibilità di quelli parametrici: una strategia corretta è quella di confrontare i risultati di un test parametrico con quelli del rispettivo non parametrico.

[1] Quanto esposto in questo paragrafo **non costituisce oggetto della prova d'esame**, tuttavia viene inserito come stimolo di approfondimento in quanto argomento importante nell'ambito della verifica di ipotesi per dati biomedici.

Le tecniche non parametriche richiedono un minor numero di assunti sulla natura delle distribuzioni originali e, pertanto, sono denominati metodi indipendenti dalla distribuzione. Comunque i test di ipotesi non parametrici si basano sostanzialmente sulla stessa strategia di analisi dei test parametrici.

Gran parte dei test non parametrici fa riferimento ai *ranghi*, cioè ai numeri (naturali) che vengono impiegati per sostituire i valori del carattere preso in esame, in modo tale che al più piccolo valore del carattere si assegni rango 1, al valore immediatamente superiore il rango 2 e così via. Ad esempio, se la variabile è la pressione sistolica con valori 145, 123, 145, 130, 150, occorre anzitutto ordinare tali valori (mm/Hg): 123, 130, 145 (*), 145 (*), 150 e, quindi, assegnare i ranghi: 1, 2, 3.5 (*), 3.5 (*), 5. A dati identici (*) viene attribuita la media aritmetica dei ranghi che sarebbero assegnati a tali valori nel caso fossero stati diversi: nel caso in esame il valore 3.5 è originato dalla media dei ranghi che sarebbero stati assegnati ai due valori 145 (ranghi 3 e 4), ovvero: (3+4)/2 = 3.5.

L'utilità del ricorso ai ranghi si fonda sul fatto che, anche se non è nota la legge di distribuzione del carattere esaminato, la variabile rango è "dominabile" statisticamente e permette, quindi, di effettuare verifiche sulla validità o meno di ipotesi statistiche.

Di seguito si illustreranno brevemente il test di Mann-Whitney e quello di Wilcoxon.

Test U di Mann-Whitney

Il test permette di verificare la significatività della differenza tra le mediane ***di due campioni indipendenti***; in certo senso può essere considerato il corrispettivo non parametrico del test t di Student per dati indipendenti.

L'ipotesi nulla è che i due insiemi di dati appartengano alla stessa popolazione o a due popolazioni con la stessa **mediana:**

$$H_0: \ Me_1 = Me_2$$

e l'ipotesi alternativa: $H_1: Me_1 \neq Me_2$, oppure $H_1: Me_1 < Me_2$ o ancora $H_1: Me_1 > Me_2$,

essendo Me_1 ed Me_2, rispettivamente la mediana del I gruppo e quella del II gruppo.

Per presentare la metodica della verifica si consideri l'esempio seguente.

Esempio 8. Su due campioni di pazienti (A e B) di numerosità n_A=11 e n_B=13 sono stati misurati i livelli di CD4 (cellule/mm^3). I risultati sono riportati in tabella.

Gruppo A	619	600	490	1076	654	955	563	955	827	873	1253		
Gruppo B	346	507	598	228	576	338	1153	354	560	517	381	415	626

In primo luogo i dati vanno uniti in un unico insieme e posti in ordine crescente secondo il loro valore algebrico e va assegnato ad essi il rango:

CD4	228	338	346	354	381	415	490	507	517	560	563	576
Gruppo	B	B	B	B	B	B	A	B	B	B	A	B
Rango	1	2	3	4	5	6	7	8	9	10	11	12

CD4	598	600	619	626	654	827	873	955	955	1076	1153	1253
Gruppo	B	A	A	B	A	A	A	A	A	A	B	A
Rango	13	14	15	16	17	18	19	20.5	20.5	22	23	24

Vanno considerate poi la somma R_A dei ranghi del gruppo A e quella R_B dei ranghi del gruppo B e calcolate le quantità:

$$W_A = R_A - \frac{n_A(n_A+1)}{2} \quad \text{e} \quad W_B = R_B - \frac{n_B(n_B+1)}{2};$$

la statistica di riferimento è ottenuta prendendo il minimo tra i due precedenti valori:

$$U = \min(W_A, W_B).$$

Per l'esempio risulta:

$$R_A = 7+11+14+15+17+18+19+20.5+20.5+22+24 = 188 \quad e \quad W_A = 188 - 66 = 122$$
$$R_B = 1+2+3+4+5+6+8+9+10+12+13+16+23 = 112 \quad e \quad W_B = 112 - 91 = 21$$

dunque:

$$U = 21.$$

Si fa osservare che, in realtà, tra i due valori W sussiste la relazione: $W_A+W_B = n_A \cdot n_B$.

Nel caso che sia vera l'ipotesi nulla di uguaglianza delle due mediane, i dati dei due campioni dovranno essere casualmente "mescolati" e la statistica U tenderà ad un valore medio dato da:

$$\mu_U = \frac{n_A \cdot n_B}{2}.$$

La significatività della differenza tra le due mediana può essere testata facendo riferimento alla distribuzione di U sotto l'ipotesi nulla e confrontando il suo valore calcolato sul campione con il valore atteso μ_U. Senza entrare negli aspetti teorici della problematica, ai fini applicativi, per valutare la significatività e, in particolare, calcolare il *p-value*, è opportuno utilizzare un software statistico.

Assumendo come ipotesi alternativa H_1: $Me_1 \neq Me_2$, di seguito si riporta una sessione di lavoro con **R** relativa all'esempio del CD4:

```
> A=c(619, 600, 490, 1076, 654, 955, 563, 955, 827, 873,1253)
> B=c(346, 507, 598, 228, 576, 338, 1153, 354, 560, 517, 381, 415, 626)

> wilcox.test(A,B,alternative="two.sided",exact=F,correct=T)

  Wilcoxon rank sum test with continuity correction

  data:  A and B
  W = 122, p-value = 0.003762
  alternative hypothesis: true location shift is not equal to 0

> wilcox.test(A,B,alternative="two.sided",exact=F,correct=T)

  Wilcoxon rank sum test with continuity correction

  data:  B and A
  W = 21, p-value = 0.003762
  alternative hypothesis: true location shift is not equal to 0
```

Essendo il *p-value* inferiore all'1%, si rifiuta l'ipotesi nulla e si accetta quella alternativa di diversità (test a due code) dei valori di CD4 nei due gruppi di pazienti considerati.

Nel caso che le numerosità campionarie siano superiori a 15 (almeno una di esse), considerata la deviazione standard:

$$\sigma_U = \sqrt{\frac{n_A \cdot n_B \cdot (n_A + n_B + 1)}{12}} \quad [2],$$

la quantità:

[2] In presenza di più dati ripetuti (*ties*) la formula andrebbe opportunamente modificata.

$$z = \frac{U - \mu_U}{\sigma_U}$$

si distribuisce secondo la curva normale standard e, quindi, per la significatività, si può utilizzare la relativa Tavola (*Tavola 1*). L'approssimazione con la normale è accettabile anche per numerosità appena superiori a 10 e per l'esempio analizzato risulta:

$$\mu_U = 71.5, \ \sigma_U = 17.26 \qquad z = (21-71.5)/17.26 = -2.93$$

con un *p-value* pari a:

```
> 2*pnorm(abs(-2.93),lower.tail=F)
[1] 0.00338962
```

indicativo di una forte evidenza espressa dai dati osservati contro l'ipotesi nulla, in favore di quella alternativa. Si fa osservare, per inciso, che si ottiene significatività statistica (*p-value*<0.01) anche utilizzando un t-test di Student per campioni indipendenti, ma naturalmente vanno verificati preliminarmente gli assunti di applicabilità del test.

Test di Wilcoxon

Il test permette di verificare la significatività delle differenze tra le **n** coppie di dati ***di due campioni dipendenti***; può essere considerato il corrispettivo non parametrico del test t di Student per dati appaiati.

L'ipotesi nulla è che la mediana delle differenze tra coppie di dati è pari a zero:

$$H_0: \ \delta = 0$$

mentre l'ipotesi alternativa: $H_1: \ \delta \neq 0$, oppure $H_1: \ \delta < 0$ o ancora $H_1: \ \delta > 0$.

La metodologia di verifica sarà illustrata considerando l'esempio del GPT (esempio 7).

Nella tabella che segue sono riportati, dalla seconda alla quinta colonna, i dati prima/dopo della transaminasi, le loro differenze e le differenze in valore assoluto.

Pz.	prima	dopo	Differenze	Val. Ass. Diff.	Rango	Rango (con segno)
1	56	68	12	12	4	+ 4
2	310	25	-285	285	12	−12
3	172	90	-82	82	10	−10
4	457	29	-428	428	13	−13
5	74	50	-24	24	5	− 5
6	66	37	-29	29	6	− 6
7	45	50	5	5	3	+ 3
8	71	27	-44	44	8	− 8
9	42	44	2	2	1.5	+1.5
10	321	41	-280	280	11	−11
11	96	34	-62	62	9	− 9
12	42	44	2	2	1.5	+1.5
13	61	22	-39	39	7	− 7

Le eventuali differenze nulle vanno eliminate per l'analisi successiva. La sesta e settima colonna della tabella evidenziano, rispettivamente, il rango delle differenze in valore assoluto ed il rango munito del segno della differenza corrispondente.

Successivamente si sommano i ranghi aventi lo stesso segno:

S_+ = somma ranghi positivi = 4+3+1.5+1.5 = 10,

S_- = somma ranghi negativi = 12+10+13+5+6+8+11+9+7 = 81,

e la statistica di riferimento si ottiene dal minimo delle due somme:

$$V = \min (S_+, S_-).$$

È utile osservare che tra le due somme sussiste la relazione: $S_- = n\cdot(n+1)/2 - S_+$. Per l'esempio considerato si ottiene V = 10.

Sotto l'ipotesi nulla, la mediana delle differenze tra le due serie di **n** coppie dovrebbe essere uguale a zero. Pertanto la somma dei ranghi con segno positivo (+) dovrebbe essere uguale a quella dei ranghi con segno meno (−) e, quindi, V dovrebbe tendere al valore medio atteso:

$$\mu_V = \frac{n \cdot (n+1)}{4}.$$

La significatività della mediana delle differenze viene analizzata testando la significatività della differenza tra V e μ_V e, nello specifico, facendo riferimento alla distribuzione della statistica V sotto l'ipotesi nulla. Ai fini applicativi si può utilizzare un software statistico.

Con l'ipotesi alternativa H_1: $\delta \neq 0$, l'ambiente **R** produce i seguenti risultati:

```
> wilcox.test(prima,dopo,paired=T,alternative="two.sided",exact=F)

  Wilcoxon signed rank test with continuity correction

  data:  prima and dopo
  V = 81, p-value = 0.01442
  alternative hypothesis: true location shift is not equal to 0

> wilcox.test(dopo,prima,paired=T,alternative="two.sided",exact=F)

  Wilcoxon signed rank test with continuity correction

  data:  dopo and prima
  V = 10, p-value = 0.01442
  alternative hypothesis: true location shift is not equal to 0
```

C'è, allora, indicazione per un rifiuto dell'ipotesi nulla e l'accettazione dell'ipotesi alternativa di differenza significativa nel valore mediano di transaminasi GPT.

Se si assume l'ipotesi alternativa H_1: $\delta < 0$, si ottengono, sempre in **R**, i risultati:

```
> wilcox.test(dopo,prima,paired=T,alternative="less",exact=F)

  Wilcoxon signed rank test with continuity correction

  data:  dopo and prima
  V = 10, p-value = 0.007208
  alternative hypothesis: true location shift is less than 0
```

con una indicazione maggiore ad un rifiuto dell'ipotesi H_0.

Per campioni con numerosità n > 25, la statistica

$$z = \frac{V - \mu_V}{\sigma_V}$$

si distribuisce approssimativamente secondo la curva normale standard, con deviazione standard data da[3]:

$$\sigma_V = \sqrt{\frac{n \cdot (n+1) \cdot (2 \cdot n+1)}{24}}\ .$$

In realtà la distribuzione normale risulta una buona approssimazione già quando n è intorno a 13÷15 coppie. Allora, per i dati dell'esempio 7 considerato risulta:

$$\mu_V = 45.5, \quad \sigma_V = 14.31 \quad \text{e} \quad z = (10–45.5)/14.31 = –2.48$$

con un *p-value* a due code:

```
> 2*pnorm(abs(-2.48),lower.tail=F)
[1] 0.01313824
```

che porta a ritenere il risultato dello studio statisticamente significativo.

IL TEST CHI-QUADRATO

Analisi di frequenze

Esempio 9. Una farmacia ha 4 marche di prodotti dietetici: A, B, C e D. Si vuole studiare la preferenza degli utenti analizzandone un campione casuale. Considerato un campione di n = 400 clienti, si supponga di aver osservata la distribuzione di vendite di tabella seguente:

Prodotto	A	B	C	D
Frequenza osservata (n° vendite)	40	80	120	160

Ci si chiede: le scelte riscontrate sono state effettuate "a caso" o dipendono, ad esempio, dalla propaganda commerciale, dal prezzo di vendita, dalla bontà del prodotto o, in generale, "non da scelte casuali"?

Affinché ci sia scelta casuale deve essere:

$$\text{Probabilità (scelta di una marca)} = \frac{1}{4}$$

e, quindi, la *frequenza attesa* (*teorica*) di scelta di una marca è data da: $\frac{1}{4} \cdot 400 = 100$. Pertanto risulta:

Prodotto	A	B	C	D
Frequenza osservata	40	80	120	160
Frequenza attesa	100	100	100	100

È intuitivo che differenze "piccole" tra frequenze osservate e frequenze attese possono essere ritenute casuali e, quindi, non in grado di negare un sostanziale accordo tra quanto osservato e quanto atteso; differenze "grandi" portano a supporre la presenza di fattori effettivi che originano tali differenze.

[3] Nel caso di pochi punteggi uguali (ties); altrimenti si impone una correzione per la deviazione standard.

Come misura di "distanza" tra le frequenze osservate f_i e quelle attese f_i^* si considera l'indice *chi-quadrato* χ^2 (vedere *Appunti di Statistica Descrittiva_2*):

$$\chi^2 = \sum_i \frac{(f_i - f_i^*)^2}{f_i^*}$$

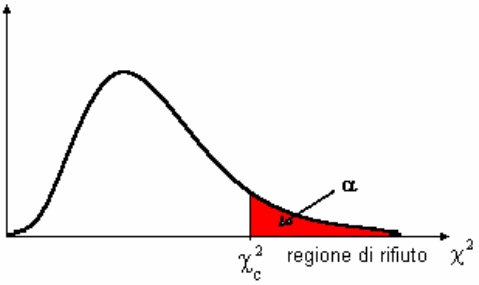
che, per i dati considerati, assume il valore:

$$\chi^2 = \frac{(40-100)^2}{100} + \frac{(80-100)^2}{100} + \frac{(120-100)^2}{100} + \frac{(160-100)^2}{100} = 80\,.$$

È evidente che il chi-quadrato aumenta con l'aumentare delle differenze tra frequenze osservate e quelle attese. Se esso supera certi opportuni valori, la differenza viene ritenuta *significativa*; in caso contrario, non si può affermare l'esistenza di una significativa differenza tra dati osservati e frequenze teoriche.

Come si può stabilire, allora, se la distribuzione delle frequenze osservate è *significativamente differente,* in termini statistici, da quella delle frequenze attese ?

Si tratta di effettuare una verifica di ipotesi sulle frequenze utilizzando la ***statistica*** χ^2 (vedere *Dispensa sul Campionamento*). Si ipotizza, ed è l'ipotesi nulla H_0, che le scelte siano state effettuate "a caso", mentre l'ipotesi alternativa H_1 suppone una distribuzione osservata realmente diversa da quella attesa. Fissato, allora, un livello di significatività α, bisogna determinare quel valore $\chi_c^2 = \chi_{\alpha,\nu}^2$ della distribuzione χ^2, con $\nu = k-1$ gradi di libertà, per il quale $\text{Prob}\left(\chi^2 \geq \chi_c^2 | H_0\right) = \alpha$. Se il valore del chi-quadrato calcolato sui dati (χ_{calc}^2) risulta superiore o uguale a χ_c^2, si rifiuta l'ipotesi nulla accettando l'ipotesi alternativa, in caso contrario si afferma che le differenze riscontrate non sono significative. Tale metodica di verifica si basa sul fatto che a valori alti del χ^2 si associano probabilità piccole che le differenze tra osservato e atteso siano dovute al caso; mentre valori bassi della statistica portano ad elevati valori della probabilità che tali differenze siano imputabili al caso.

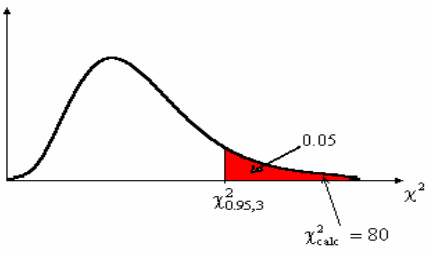


Il valore χ_c^2 può essere individuato utilizzando, ad esempio, la *Tavola* (*Tavola 3*) *della Distribuzione chi-quadrato* (Appendice *Dispensa sul Campionamento*). Per l'esempio in esame, in corrispondenza ad un livello di significatività α = 5% e ν = k–1 = 4–1 = 3 gradi di libertà, si ottiene il valore $\chi_c^2 = \chi_{0.05,3}^2 = 7.815$.

Se, in alternativa, si utilizza l'ambiente **R**, si ottiene:

```
> qchisq(0.05,3,lower.tail=F)
[1] 7.814728
```

Poiché $\chi_{calc}^2 = 80$ è decisamente superiore a quello tabulato (7.815), si può affermare che la differenza tra le frequenze osservate e quelle teoriche è statisticamente significativa al livello di errore del 5%. In altre parole: ammettendo che i 4 prodotti abbiano pari probabilità di scelta, ovvero siano scelti casualmente dai clienti, e ripetendo l'esperimento infinite volte, si potrebbe osservare piuttosto raramente (ossia 5 volte su 100 o meno!) dati simili a quelli ottenuti oppure con frequenze osservate ancora più "diverse" da quelle attese.



In conclusione l'affermazione "le scelte non sono dovute al caso" ha il 95% probabilità o più di essere vera e, allora, si respinge l'ipotesi nulla della non influenza della propaganda fatta dalle ditte produttrici o di altre cause sulle scelte dei clienti.

Come già osservato nella pratica si fa spesso riferimento al *p-value* $p = \text{Prob}\left(\chi^2 \geq \chi^2_{calc} \middle| H_0\right)$; così per l'esempio dei prodotti dietetici si consideri la seguente procedura in **R** (***df*** indica il numero di gradi di libertà):

```
> freq<-c(40,80,120,160)
> prob<-c(1,1,1,1)/4
> chisq.test(x=freq,p=prob)

      Chi-squared test for given probabilities
 data:  freq
 X-squared = 80, df = 3, p-value < 2.2e-16.
```

Come si vede, il *p-value* è molto piccolo (decisamente inferiore a 0.001), pertanto l'ipotesi di scelta casuale dei prodotti dietetici va rifiutata.

Test chi-quadrato per l'indipendenza

Esempio 10. Nell'ambito dello studio sull'efficacia di un farmaco di nuova sintesi, 100 pazienti vengono curati mediante somministrazione del farmaco; ad altri 100 pazienti, affetti dalla stessa patologia, viene somministrato del placebo. A conclusione del trattamento si rilevano i seguenti risultati:

Terapia	Esito		Totale
	Guarito	Non guarito	
Farmaco	75	25	100
Placebo	65	35	100
Totale	140	60	200

Ci si chiede se, in base ai risultati rilevati, si può ritenere il farmaco efficace, vale a dire se le differenze di esito riscontrate tra i trattati ed i non trattati sono dovute al caso oppure all'effetto del farmaco.

Il test statistico che ci permette di dare una risposta al problema è il test χ^2**:**

$$\chi^2 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}\left(f_{ij} - f^*_{ij}\right)^2}{f^*_{ij}},$$

dove f_{ij} sono le frequenze assolute osservate, f^*_{ij} le frequenze assolute teoriche in caso di indipendenza fra le due variabili, **r** è il numero di righe e **c** il numero di colonne della *tabella di contingenza* sopra riportata [(4)].

Le frequenze teoriche si ottengono ipotizzando, appunto, l'indipendenza tra i due caratteri presi in esame (nell'esempio Terapia ed Esito) e, precisamente, attraverso le

---

[(4)] L'approssimazione della formula scritta sopra alla distribuzione teorica del χ^2 è basata sull'assunto che la dimensione del campione sia sufficientemente elevata. In caso contrario per verificare l'ipotesi che le frequenze osservate non sono significativamente diverse da quelle attese è necessario utilizzare altri test statistici. Quando ci si trova di fronte a tabelle 2 x 2 con frequenze piccole è consigliabile usare per il calcolo del χ^2 la correzione di Yates per la continuità, opportuna in quanto si passa da una distribuzione discontinua a quella continua del χ^2. La correzione si effettua sottraendo 0.5 al valore assoluta della differenza tra le frequenze osservate e quelle attese. Il calcolo del χ^2 viene così modificato: $\chi^2 = \frac{\sum\left(\left|f - f^*\right| - 0{,}5\right)^2}{f^*}$.

$$f^*_{ij} = \frac{f_{i,\bullet} \cdot f_{\bullet,j}}{n}, \qquad \text{per } i = 1,2,\ldots,r \text{ e } j = 1,2,\ldots,c,$$

essendo n il totale dei dati, $f_{i,\bullet}$ il totale della riga i e $f_{\bullet,j}$ il totale della colonna j.

L'ipotesi nulla H_0 presuppone che tra terapia e guarigione non ci sia relazione, cioè che i due caratteri siano indipendenti; *l'ipotesi alternativa H_1* presuppone una effettiva dipendenza.

Sotto l'ipotesi H_0 si sarebbe dovuto riscontrare nei due gruppi la stessa percentuale di guarigioni, in altre parole valori di frequenze osservate "vicine" a quelle teoriche:

$$f^*_{11} = (100 \cdot 140)/200 = 70, \quad f^*_{12} = (100 \cdot 60)/200 = 30, \quad f^*_{21} = (100 \cdot 140)/200 = 70, \quad f^*_{22} = (100 \cdot 60)/200 = 30$$

ottenendo i risultati di tabella:

Terapia	Esito		Totale
	Guarito	Non guarito	
Farmaco	70	30	100
Placebo	70	30	100
Totale	140	60	200

Il calcolo del chi-quadrato campionario è dato da:

$$\chi^2_{calc} = \frac{(75-70)^2}{70} + \frac{(25-30)^2}{30} + \frac{(65-70)^2}{70} + \frac{(35-30)^2}{30} = 2.38\,.$$

Questo valore va confrontato con il valore critico χ^2_c tabulato sulla *Tavola 3* in corrispondenza di un certo *livello di significatività* α e di $\nu =$ (r−1)·(c−1) *gradi di libertà*. Le tavole indicano i valori critici oltre i quali le differenze riscontrate tra le frequenze osservate e le teoriche in caso di indipendenza non sono attribuibili al caso, ma a fattori sistematici.

Se il valore χ^2_{calc} calcolato sul campione risulta minore di quello tabulato, si dice che il test non è significativo, ovvero le differenze riscontrate tra frequenze osservate e teoriche sono attribuibili, con un certo livello di significatività, al caso e, quindi, non si rifiuta l'ipotesi H_0 di indipendenza tra le variabili. Se invece χ^2_{calc} è superiore o uguale a χ^2_c, si dice che il test è significativo e si rifiuta H_0, concludendo che tra i caratteri presi in esame esiste una relazione statisticamente significativa (ipotesi alternativa H_1).

Nell'esempio del farmaco, poiché il χ^2_{calc} =2.38 è minore del valore critico χ^2_c = 3.84, ottenuto dalla *Tavola 3* in corrispondenza ad un livello di significatività del 5% e ν = 1 gradi di libertà, si può affermare che il test non è significativo, non si rifiuta l'ipotesi H_0 di indipendenza e si conclude che il farmaco, alla luce dell'esperimento campionario, non risulta statisticamente efficace e le differenze riscontrate tra valori osservati e valori attesi sono imputabili a variazioni casuali di campionamento.

Di seguito è riportata una procedura per effettuare il test nell'ambiente **R** (è riportato anche il calcolo con la correzione di Yates):

```
> Terapia=c("Farmaco","Placebo")
> esito=c("Guarito","non guarito")
> tab=as.table(matrix(c(75,65,25,35),2,2,dimnames=list(Terapia,esito)))

> summary(tab)
  Number of cases in table: 200
```

Number of factors: 2

Test for independence of all factors:
Chisq = 2.381, df = 1, p-value = 0.1228

> chisq.test(tab)
Pearson's Chi-squared test with Yates' continuity correction
data: tab
X-squared = 1.9286, df = 1, p-value = 0.1649.

Si osservi che, sia senza correzione sia con la correzione di Yates, i *p-value* danno indicazione per una non significatività del test.

Esempio 11. Nascite di bambini nelle regioni di England e Walles nel 1956

Sesso	Vitalità		Totale
	Nati vivi	Nati morti	
Maschi	359881	8609	368490
Femmine	340454	7796	348250
Totale	700335	16405	716740

Si vuole verificare se il sesso ha influenza sulla mortalità del neonato, cioè se il Sesso e la Vitalità dei neonati sono in relazione. L'ipotesi nulla H_0 suppone l'indipendenza tra i due caratteri. Il numero di gradi di libertà è pari a $\nu = (2-1)(2-1) = 1$ e si assuma un livello di significatività del 5%.

Sotto H_0, la frequenza attesa che un neonato sia, ad esempio, maschio e nato vivo, è data da: $f^*_{11} = (368490 \cdot 700335)/716740 = 360056$; in modo analogo si calcolano le altre frequenze teoriche.

La tabella doppia di frequenza in caso di indipendenza fra caratteri è la seguente:

Sesso	Vitalità		Totale
	Nati vivi	Nati morti	
Maschi	360056	8424	368490
Femmine	340279	7971	348250
Totale	700335	16405	716740

La differenza tra frequenza osservata e frequenza teorica vale sempre 175 e quindi:

$$\chi^2_{calc} = \frac{175^2}{360056} + \frac{175^2}{340279} + \frac{175^2}{8434} + \frac{175^2}{7971} = 7.65$$

Poiché il χ^2_{calc} =7.65 è maggiore del valore critico χ^2_c = 3.84, determinato sulla *Tavola 3* per un livello di significatività del 5% e $\nu = 1$ gradi di libertà, il test è significativo; si rigetta l'ipotesi H_0 di indipendenza, concludendo che, con un errore del 5%, nei neonati tra Sesso e Vitalità non vi è indipendenza.

Di seguito è riportata una procedura di verifica in **R**:

> Sesso=c("maschi","femmine")
> vitalità=c("nati vivi","nati morti")
> nascite=as.table(matrix(c(359881,340454,8609,7796),2,2,dimnames=list(Sesso,vitalità)))

> chisq.test(nascite)
Pearson's Chi-squared test with Yates' continuity correction
data: nascite
X-squared = 7.5933, df = 1, p-value = 0.005859.

che fa rilevare una forte evidenza espressa dai dati campionari contro l'ipotesi nulla di non dipendenza tra Sesso e Vitalità.

Esempio 12. Si vuole stabilire l'esistenza di una relazione tra il valore della Glicemia e la Razza di appartenenza di individui. In un campione casuale si rilevano le frequenze riportate in tabella.

Glicemia	Razza		Totale
	Negri	Bianchi	
Alta ≥ 110	300	400	700
Border-line 90-110	400	550	950
Normale ≤ 90	150	200	350
Totale	850	1150	2000

L'ipotesi nulla H_0 concerne l'inesistenza di una relazione statisticamente significativa tra Razza e livello di Glicemia; quella alternativa H_1 ipotizza l'esistenza di associazione tra i due caratteri. Anche per questo esempio va considerata la distribuzione campionaria del chi-quadrato: si osservi ancora che il valore del χ^2 aumenta con l'aumentare della discrepanza tra le frequenze osservate e quelle teoriche (ottenute sotto l'ipotesi di indipendenza dei due caratteri), mentre assume valore minimo nel caso di vicinanza tra i due tipi di frequenze. Si consideri un livello di significatività α = 1%; allora in corrispondenza di un α=0.01 e di $\nu=(2-1)\cdot(3-1) = 2$ gradi di libertà, dalla *Tavola 3* si ottiene un valore $\chi^2_c = \chi^2_{0.01,2} = 9.210$. Come anche per le altre situazioni analizzate, l'ipotesi H_0 può essere "rifiutata" al livello dell'1% se il valore del chi-quadrato calcolato (χ^2_{calc}) supera il valore 9.210: in tal caso, infatti, il *p-value* $\text{Prob}\left(\chi^2 \geq \chi^2_{calc} \middle| H_0\right) < 0.01$.

Costruendo la tabella di indipendenza:

Glicemia	Razza		Totale
	Negri	Bianchi	
Alta ≥ 110	297.5	402.5	700
Border-line 90-110	403.75	546.25	950
Normale ≤ 90	148.75	201.25	350
Totale	850	1150	2000

Si calcola il chi-quadrato sui dati campionari:

$$\chi^2_{calc} = \frac{(300-297.5)^2}{297.5} + \frac{(400-402.5)^2}{402.5} + \frac{(400-403.75)^2}{403.75} + \frac{(550-546.25)^2}{546.25} + \frac{(150-148.75)^2}{148.75} + \frac{(200-201.25)^2}{201.25} = 0.12$$

Tale valore è inferiore a χ^2_c=9.210, pertanto si conclude, con una probabilità dell'1% di commettere un errore del I tipo, per un *non rifiuto* dell'ipotesi nulla; in sostanza non si può ammettere l'esistenza di associazione tra Razza e Glicemia.

Anche per questo esempio si effettua la procedura in **R**, che porta ancora a concludere per la non significatività del test:

```
> Glicemia<-c("Alta","Border-line","Normale")
> Razza<-c("negri","bianchi")
> RazzaGlic<-as.table(matrix(c(300,400,150,400,550,200),3,2,dimnames=list(Glicemia,Razza)))

> chisq.test(RazzaGlic)
  Pearson's Chi-squared test
  data:  RazzaGlic
  X-squared = 0.1154, df = 2, p-value = 0.944.
```

QUESITI

1) I livelli di acido urico in una popolazione maschile assai numerosa si distribuiscono normalmente con media pari a 6mg/100ml. Un campione di 101 maschi estratto con campionamento casuale semplice dalla popolazione di tutti i maschi affetti da diabete mellito fa ottenere una media = 6.5mg/100ml, con deviazione standard s = 2mg/100ml. Si verifichi l'ipotesi che l'universo dei maschi affetti da diabete mellito abbia una media pari a 6mg/10ml, contro l'ipotesi alternativa che sia maggiore, al livello $\alpha = 0.01$.

2) In un campione di 300 individui si è rilevato che 90 maschi e 100 femmine hanno occhi scuri, 40 maschi e 70 femmine occhi chiari. Verificare, ad un livello di significatività del 5%, se il sesso influenza il colore degli occhi.

3) La parti orientale e occidentale di una città ricevono l'acqua potabile da 2 acquedotti diversi. L'acquedotto che fornisce le abitazioni della zona orientale è stato contaminato per diversi anni dagli scarichi tossici provenienti da una fabbrica, e si ipotizza che tale contaminazione abbia determinato un aumento dei casi di tumore allo stomaco. Per testare tale ipotesi, vengono confrontati due campioni casuali di 614 e 512 individui provenienti dalle due zone (Est e Ovest rispettivamente), nei quali si riscontrano 28 e 16 casi di tumore. Cosa si può concludere dopo aver analizzato i dati con un test statistico appropriato?

4) In un campione di 80 pazienti si sono riscontrati i seguenti esiti in riferimento ad un dato trattamento medico:

	Guarito	Non guarito	Totale
Trattato	15	15	30
Non trattato	20	30	50
Totale	35	45	80

Verificare, ad un livello di significatività del 5%, se c'è associazione tra trattamento e guarigione.

5) In uno studio sulle relazioni tra gruppi sanguigni e malattie, tre campioni molto ampi di malati di ulcera peptica, tumore allo stomaco e controlli senza queste due malattie, sono stati tipizzati per il sistema ABO. Questi sono i risultati (gli individui AB, essendo in numero molto esiguo, non vengono considerati):

Gruppo sanguigno	N. malati di ulcera	N. malati di tumore	N. pazienti sani
O	983	383	2892
A	679	416	2625
B	134	84	570

Testare l'ipotesi nulla che la distribuzione delle malattie è indipendente dal gruppo sanguigno.