



# Cenni introduttivi di statistica inferenziale con

per il modulo SM3 della SSSMT di Locarno

Dr Giorgio Pioda\*

Ultima compilazione il 12 settembre, 2022

Quest'opera è distribuita con licenza [Creative Commons](https://creativecommons.org/licenses/by-nc-sa/4.0/) “Attribuzione – Non commerciale – Condividi allo stesso modo 4.0 Internazionale”.



# Indice

<b>1</b>	<b>Introduzione</b>	<b>3</b>
<b>2</b>	<b>Distribuzioni di campionamento e stimatori</b>	<b>3</b>
2.1	Popolazione e campione . . . . .	3
2.2	Proprietà delle distribuzioni di campionamento . . . . .	3
2.3	Deviazione standard e deviazione standard corretta . . . . .	4
2.4	Lavorando su un solo campione . . . . .	6
<b>3</b>	<b>Test statistici di significatività</b>	<b>8</b>
3.1	Rapporti critici . . . . .	8
<b>4</b>	<b>Significatività per i grandi campioni (approssimazione di Gauss)</b>	<b>8</b>
4.1	La logica dei test di significatività . . . . .	8
4.2	Esempi di determinazione della significatività . . . . .	10
4.3	Due tipi di errori $\alpha$ e $\beta$ . . . . .	11
4.4	Sensibilità, specificità e valori predittivi . . . . .	14
<b>5</b>	<b>La significatività per i piccoli campioni e la curva di Student</b>	<b>15</b>
5.1	Quando $\sigma$ della popolazione non è noto... . . . . .	15
<b>6</b>	<b>Gli intervalli di confidenza e l'inferenza sulle medie</b>	<b>16</b>
<b>7</b>	<b>Confronto tra medie</b>	<b>16</b>
7.1	Medie non indipendenti . . . . .	16
7.1.1	Tre metodi equivalenti per l'inferenza . . . . .	18
7.1.2	Test con R . . . . .	18
7.2	Medie indipendenti . . . . .	18
<b>8</b>	<b>Verifica dei prerequisiti per i test <math>t</math></b>	<b>21</b>
8.1	Verifica della normalità . . . . .	21
8.2	Test per l'omoschedasticità . . . . .	22
8.2.1	Test F . . . . .	23
8.2.2	Test di Levene . . . . .	23
<b>9</b>	<b>Test equivalenti <i>non parametrici</i></b>	<b>24</b>
9.1	Il test Wilcoxon per dati appaiati. . . . .	24
9.2	Il test U di Wilcoxon-Mann-Whitney . . . . .	25
<b>10</b>	<b>Stima della numerosità di un campione</b>	<b>26</b>
10.1	Con i limiti di confidenza per una ipotesi . . . . .	27
10.2	Con ipotesi nulla $H_0$ e ipotesi alternativa $H_A$ (errori $\alpha$ e $\beta$ ) . . . . .	27
<b>11</b>	<b>Inferenza con le proporzioni (distribuzione binomiale)</b>	<b>28</b>
11.1	Sensibilità, specificità, & co (reloaded) . . . . .	30
11.2	Proporzioni indipendenti (test binomiale) . . . . .	31
11.3	Proporzioni indipendenti (test $\chi^2$ ) . . . . .	32
11.4	Proporzioni dipendenti con il McNemar test . . . . .	33
<b>12</b>	<b>Tabelle</b>	<b>35</b>
12.1	Integrali della curva di Gauss . . . . .	35
12.2	Valori critici di Student . . . . .	36
12.3	Valori critici per il $\chi^2$ . . . . .	37
12.4	Valori critici $T$ di Wilcoxon (paired oppure sign rank test) . . . . .	38
12.5	Valori critici $U$ di Wilcoxon . . . . .	39

## 1 Introduzione

L'inferenza statistica è il processo per mezzo del quale dai risultati osservati in un campione si possono trarre conclusioni riguardanti la popolazione. Nelle pagine seguenti l'interesse è accentrato sui dati quantitativi e sulle inferenze riguardanti le medie delle popolazioni. Inizialmente si considererà una singola media e quindi il confronto tra due medie usando la distribuzione normale e quella di Student. Si introdurranno successivamente anche alcuni metodi non parametrici (robusti). Infine si aprirà una parentesi sulla comparazioni delle proporzioni usando la distribuzione binomiale

## 2 Distribuzioni di campionamento e stimatori

### 2.1 Popolazione e campione

In statistica vi sono due tipi di distribuzioni: distribuzioni della **popolazione** e distribuzioni **campionarie**. Il primo dei due termini si riferisce alla distribuzione delle caratteristiche in definite popolazioni; per esempio, livelli di acido urico in soggetti di sesso maschile o tempi di sopravvivenza di pazienti affetti da cancro trattati con un farmaco. Il secondo termine si riferisce alle distribuzioni che si derivano dalle distribuzioni della popolazione estraendo dalla popolazione campioni con una certa dimensione  $n$ . Di rado si incontrano intere distribuzioni di campionamento; però è necessario studiarle e conoscere le loro proprietà al fine di eseguire l'inferenza statistica. Infatti le distribuzioni di campionamento costituiscono il collegamento tra le distribuzioni delle popolazioni e le informazioni riguardanti un singolo campione in esame, che è ciò che normalmente si ha a disposizione sperimentalmente.

Le distribuzioni di campionamento delle medie si possono generare nel seguente modo:

- 1) Si estrae un campione di  $n$  osservazioni, selezionate in modo completamente casuale da un'ampia popolazione. Si determina la media del campione. Per quanto possa sembrare contro intuitivo è necessario che il campionamento avvenga con re immissione dei dati<sup>1</sup>. Tale procedimento evita di dover apportare correzioni ai valori ottenuti. Questo procedimento si chiama **campionamento Bernoulliano**.
- 2) Si ripeta all'infinito la procedura di campionamento, calcolando la media del campione casuale di grandezza  $n$  ciascuna volta..
- 3) Il risultato è un insieme di **medie di campioni** di grandezza  $n$  ( $n$  è la numerosità di ogni singolo campione). Se ciascuna media dell'insieme è ora considerata come un'osservazione individuale si determina la distribuzione di campionamento delle medie dei campioni di dimensione  $n$ . Le proprietà di questa distribuzione ci permettono di legare i dati di un singolo campione ai dati della popolazione da cui essi derivano.

**Esempio 2.1.** Se si ottengono ripetuti campioni di 25 unità da un'ampia popolazione di maschi, si determinano i livelli di acido urico serico in ciascun campione, si sostituiscono le 25 osservazioni ogni volta e quindi si dispone la media in una distribuzione., Si genera così la distribuzione di campionamento delle medie dei livelli di acido urico serico dei campioni di grandezza  $n = 25$ .

*Nota .* Si noti che già al livello della modalità di estrarre un campione da una popolazione esistono diversi approcci. Non vogliamo in questo testo entrare nel merito di questa questione e da ora in avanti si considererà il campione sempre estratto con reimmissione (campionamento Bernoulliano). Le formule che troverete da qui in avanti valgono solamente per questo tipo di campionamento; esistono tuttavia formule che permettono di trasformare medie e scarti quadratici da un tipo di campionamento all'altro che per motivi di spazio ed opportunità non sono riportate in questo riassunto.

### 2.2 Proprietà delle distribuzioni di campionamento

In questo testo si è usata la convenzione di utilizzare lettere greche per denotare i valori della popolazione e lettere latine per denotare i valori del campione. Quindi,  $\mu$  e  $\sigma$ , rispettivamente, sono la media e la deviazione standard (DS, chiamato talvolta scarto quadratico medio) della popolazione, mentre  $\bar{x}$  ed  $s$ , rispettivamente, denotano la media e la DS in un singolo campione di grandezza  $n$ . Da una popolazione

<sup>1</sup>[https://en.wikipedia.org/wiki/Bernoulli\\_sampling](https://en.wikipedia.org/wiki/Bernoulli_sampling). Si veda anche l'help della funzione `R *sample()*` e la sua opzione `replace`.

con media  $\mu$  e DS,  $\sigma$ , la distribuzione di campionamento di medie campionarie ripetute di dimensione  $n$  (cioè ripetute  $\bar{x}$ ) ha tre proprietà estremamente importanti:

- 1) La media della distribuzione di campionamento delle medie è uguale alla media della popolazione  $\mu = M(\bar{x})$ . In questo caso si dice che la media campionaria è uno **stimatore corretto** (non distorto) della media della popolazione.
- 2) La DS della distribuzione di campionamento, chiamata anche **errore standard** (ES) delle medie è  $\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}$  (e la varianza della distribuzione di campionamento è  $\sigma^2(\bar{x}) = \frac{\sigma^2}{n}$ ).
- 3) La forma della **distribuzione di campionamento delle medie è approssimativamente normale**, indipendentemente dalla forma della distribuzione della popolazione e presupposto che  $n$  sia **sufficientemente grande**:  $n$  per questi casi deve essere più grande di 30 o di 25 a seconda degli autori. Questo è in realtà il risultato di una famiglia di teoremi matematici chiamati **teoremi centrali del limite**<sup>2</sup>.

**Esempio 2.2.** La metà superiore della Figura 1 riporta una distribuzione di popolazione empirica: l'età alla comparsa della malattia per tutti i casi di poliomielite nel Massachusetts nel 1949<sup>3</sup>. La media di questa distribuzione è  $\mu = 15.86$  anni con  $\sigma = 12.29$  anni. Questa distribuzione di popolazione è chiaramente **non normale**. Da questa popolazione vengono estratti 1000 campioni casuali di  $n = 36$  osservazioni ciascuno.

La media delle 1000 medie è  $M(\bar{x}) = 15.9$  anni e l'ES delle medie è  $\sigma(\bar{x}) = 1.983$  anni.

La distribuzione delle 1000 medie è illustrata nella parte inferiore della Figura 1; questa è in buon accordo con la Proprietà 1 e 2 dei teoremi centrali del limite; la media delle medie è infatti  $M(\bar{x}) = 15.9$ , e per l'ES si ha teoricamente  $ES(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{15.86}{\sqrt{36}} = 2.048$ , valore da confrontare con il valore empirico della DS delle 1000 medie campionarie che è pari a  $\sigma(\bar{x}) = 1.983$ .

Da notare inoltre, per la Proprietà 3, la differenza nella forma delle due distribuzioni: quella della popolazione decisamente non normale e quella delle medie campionarie sostanzialmente normale.

### 2.3 Deviazione standard e deviazione standard corretta

L'operazione fatta ora per le medie dei campioni può essere fatta per qualsiasi parametro statistico prelevato dai campioni (mediana, varianza, scarto quadratico, campo di variazione, differenza interquartile, ecc), ottenendo in molti casi formule analoghe.

Vale la pena osservare che cosa accade alla distribuzione delle **DS dei singoli campioni**. Cioè di ciascun campione viene determinato anche la DS e dalla relativa distribuzione delle DS se ne può ancora calcolare la media.

In questo caso la media delle DS dei campioni  $M(s)$  non è un parametro stimatore corretto (non distorto) della DS della popolazione sottostante, cioè  $M(s) \neq \sigma$  (e lo stesso vale per la varianza  $M(s^2) \neq \sigma^2$ ). Le DS dei singoli campioni sottostimano sistematicamente la DS della popolazione.

Invece la DS corretta  $\hat{s}$  è uno stimatore non distorto (corretto) della DS della popolazione sottostante. Tale valore si calcola nel seguente modo ( $s = \sigma$  nelle formule stanno ad indicare che di per sé le formule possono essere usate sia su campioni sia su popolazioni):

$$s = \sigma = \sqrt{\frac{\sum_i (X_i - \bar{X})^2}{n}} \quad (1)$$

$$\hat{s} = \hat{\sigma} = \sqrt{\frac{\sum_i (X_i - \bar{X})^2}{n - 1}} \quad (2)$$

Quindi valgono le seguenti relazioni tra la media delle DS campionarie corrette  $\hat{s}$  e la DS della sottostante popolazione  $\sigma$ :

<sup>2</sup>[https://it.wikipedia.org/wiki/Teoremi\\_centrali\\_del\\_limite](https://it.wikipedia.org/wiki/Teoremi_centrali_del_limite)

<sup>3</sup>Dati simulati in modo verosimile. Dati originali non disponibili.

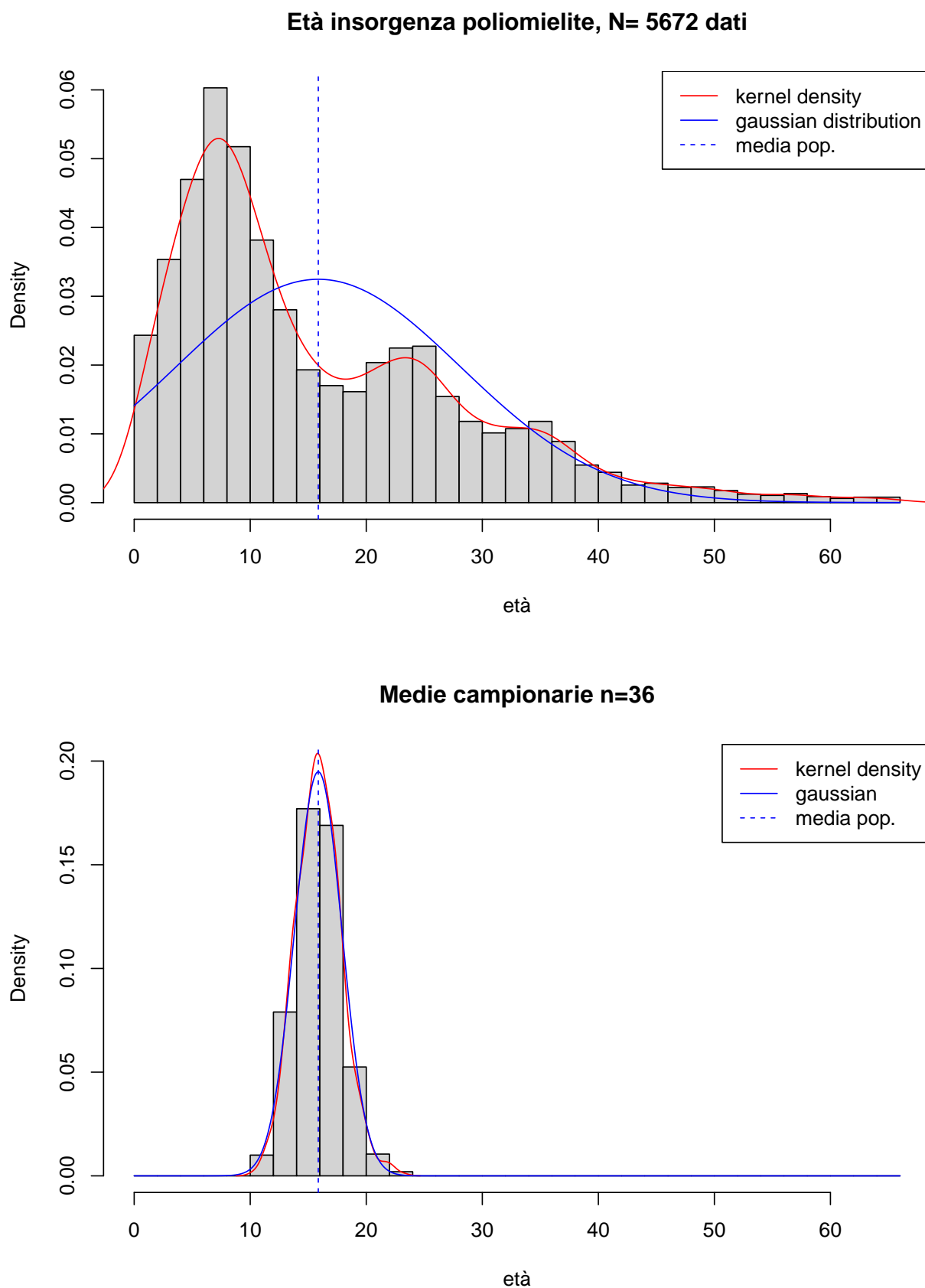


Figura 1: Popolazione e distribuzione di campionamento per i casi di poliomielite

## Deviazioni standard campionarie n=36

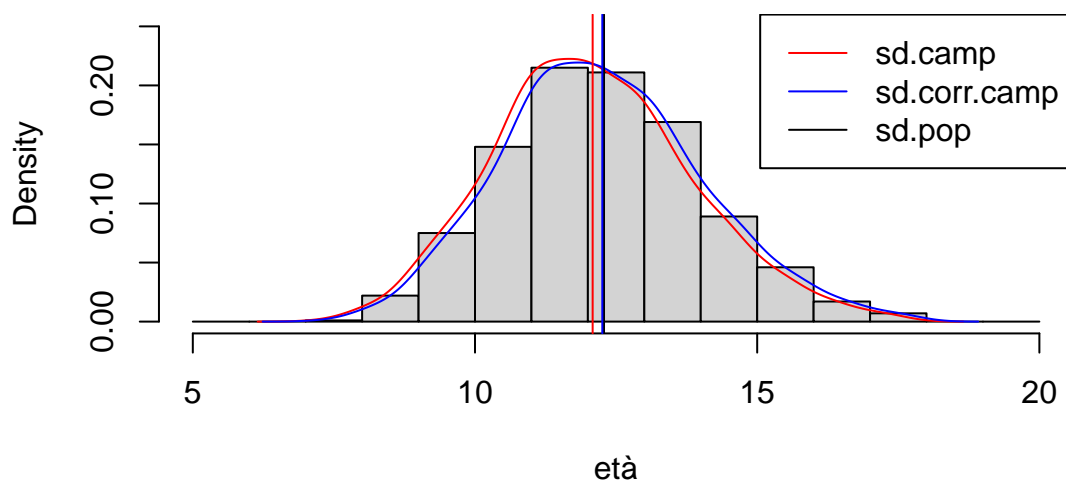


Figura 2: Andamento delle deviazioni standard e delle deviazioni standard corrette

$$M(\hat{s}) = \sigma \quad (3)$$

$$M(\hat{s}^2) = \sigma^2 \quad (4)$$

Inoltre si lascia dimostrare anche che:

$$\hat{\sigma} = \sqrt{\frac{n}{n-1} \cdot \sigma^2} \quad (5)$$

$$\hat{s}^2 = \frac{n}{n-1} \cdot \sigma^2 \quad (6)$$

*Nota* . Spesso si crea confusione con le DS in quanto nella loro determinazione non viene specificato se si tratta della formula campionaria o di quella per la popolazione. Alcuni autori usano per semplicità unicamente lo scarto quadratico medio corretto, anche quando questa grandezza non sarebbe appropriata (cioè al di fuori di studi di campionamento); di per sé questa scelta ha senso in quanto per le popolazioni ci si aspetta  $n$  molto alto e quindi i due valori differiscono di poco. Per inciso la funzione `sd()` di **R** calcola la DS campionaria. Non è disponibile nel pacchetto base una funzione per la DS della popolazione.

### 2.4 Lavorando su un solo campione

Nella maggior parte dei casi si è confrontati con un ristretto insieme di dati: uno studio della distribuzione di campionamento non risulta praticabile. Non si può determinare  $\sigma$  estrapolandola dalla distribuzione di campionamento. Se  $\sigma$  della popolazione non è noto, si può comunque usare la DS corretta. Riassumendo di può dire che la media del campione sia uno stimatore corretto della media della popolazione, mentre per stimare la DS e la varianza della popolazione si usa la DS corretta. Qui sotto una tabella riassuntiva delle varie grandezze in gioco.

Tabella 1: Riassunto delle proprietà collegate ai teoremi del limite centrale

	Media	Scarto quadratico
Popolazione (tutto l'universo, spesso infinito, spesso unicamente ipotizzato)	$\mu$	$\sigma$
Distribuzione di frequenza dei campioni (più campioni ciascuno di ampiezza $n$ )	$M(\bar{x}) = \mu$ $DS(\bar{x}) = \frac{\sigma}{\sqrt{n}} = ES$	$M(s) \neq \sigma$ $M(\hat{s}) = \sigma$
Un campione di $n$ elementi	$\bar{x}$	$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$ $\hat{s} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$

### 3 Test statistici di significatività

#### 3.1 Rapporti critici

Tenendo conto di quanto detto prima a proposito di media e scarto quadratico di popolazione e distribuzione di campioni si possono usare le unità standard  $z$  per analizzare i campioni e la loro probabilità di realizzazione. In pratica si ha la seguente formula di standardizzazione.

$$z_{stat} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad (7)$$

Notare come al denominatore l'espressione  $\frac{\sigma}{\sqrt{n}}$  non sia altro che l'ES precedentemente discusso.

**Esempio 3.1.** Da una popolazione data dalla misurazione dell'acido urico serico su pazienti maschi si estrae un campione di 25 soggetti. Dato che la media della popolazione è  $5.4[\frac{mg}{100ml}]$  e  $\sigma = 1[\frac{mg}{100ml}]$  calcolare:

- Quale proporzione dei campioni ha media pari a  $5.9[\frac{mg}{100ml}]$  o maggiore? [0.621%]
- Quale valore lascia alla sua destra il 5% della distribuzione delle medie campionarie di dimensione 25? [5.729]
- Quale deve essere la dimensione del campione perché il 5% delle medie dei campioni di tale dimensione superi la media della popolazione di  $0.2[\frac{mg}{100ml}]$  mg per 100 ml o più? [69]

**Esempio 3.2.** La popolazione consiste nei pazienti affetti da cancro con una sopravvivenza media  $\mu = 38.3$  mesi e deviazione standard  $\sigma = 43.3$  mesi. In questa popolazione il tempo di sopravvivenza non è distribuito in modo normale dato che  $\mu - \sigma$  fornisce un valore negativo. Per avere una distribuzione ragionevolmente normale la media dovrebbe essere almeno pari al triplo della DS. La dimensione campionaria di 100 è certamente grande abbastanza da assicurare che la distribuzione di campionamento delle medie sia normale.

- Tra i campioni di grandezza  $n = 100$  estratti da questa popolazione quale proporzione avrà una sopravvivenza media sia di 46.9 mesi e più sia di 29.7 mesi e meno? Gli scostamenti sono in ambedue le direzioni dalla media e sono anche simmetrici. [4.702%]
- Quali limiti comprendono il 95% delle medie dei campioni di dimensione  $n = 100$  estratti da questa popolazione? [29.81; 46.79]
- Quale dimensione  $n$  del campione di pazienti cancerosi è necessaria affinché il 95% delle medie dei campioni di questa dimensione sia compreso in un intervallo di  $\pm 6$  mesi attorno alla media della popolazione? [200]

### 4 Significatività per i grandi campioni (approssimazione di Gauss)

#### 4.1 La logica dei test di significatività

In questo paragrafo è presentato il procedimento logico con cui viene eseguito un test statistico. Gli strumenti di calcolo sono quelli già visti nei precedenti paragrafi. Si tratta ora di costruire un percorso che può portare ad accettare o a rifiutare un'ipotesi statistica.

Solitamente si parte con una considerazione a proposito della popolazione: si definisce o si recupera dalla letteratura il valore di  $\mu$  e anche quello di  $\sigma$ . Per il valore di  $\mu$  in particolare potrebbe trattarsi di un'informazione nota oppure anche di una ipotesi di lavoro.

Per i dati dei pazienti dell'esempio 3.2 affetti da tumore e trattati con il farmaco si potrebbe affermare che la sopravvivenza media nella popolazione è pari a quella dei pazienti non trattati, cioè 38.3 mesi. Per i dati dell'esempio 3.1 si potrebbe stabilire che il livello medio di acido urico serico di una popolazione maschile con una certa malattia, per esempio il diabete mellito, è pari a quella della popolazione e cioè  $5.4[\frac{mg}{100ml}]$ . In tali affermazioni,  $\mu_0$  è solitamente il valore della media della popolazione. Questa affermazione è chiamata tecnicamente **ipotesi nulla** ed è simbolizzata con  $H_0$ . Per esempio se l'affermazione è che la media del campione è uguale alla media della popolazione, l'ipotesi è matematizzata con  $H_0 : \bar{x} = \mu_0$



L'inferenza statistica consiste nel paragonare il valore del campione  $\bar{x}$  di dimensione  $n$  con la media della popolazione  $\mu_0$  usando la distribuzione campionaria<sup>4</sup>  $\bar{x} \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{n})$  traendo una delle due conclusioni seguenti rispetto all'affermazione iniziale:

- 1) Si rifiuta l'ipotesi nulla circa la media della popolazione poiché vi è evidenza sufficiente per dubitare della sua validità, e si conclude che  $\mu_0$  non è la media della popolazione da cui è stato estratto il campione in esame.
- 2) Non si rifiuta l'ipotesi nulla circa la media della popolazione poiché non vi è evidenza sufficiente per dubitare della sua validità. Si conclude che per quanto raccolto dai dati campionari  $\mu_0$  può essere considerata la media della popolazione da cui è stato estratto il campione in esame.

Il ragionamento è il seguente: si supponga che l'affermazione sia vera e la media della popolazione sia  $\mu_0$ . Qual è la probabilità che un campionamento casuale estratto da questa popolazione fornisca una media campionaria che si scosta dalla media vera di un ammontare pari o maggiore a quello della  $\bar{x}$  osservata?

Se questa probabilità è sufficientemente piccola vi è ragione di credere che la media campionaria osservata non sia un risultato plausibile. Quindi la premessa di base deve essere errata. Di conseguenza  $\mu_0$  non è la media della popolazione, cioè si è portati alla prima delle due conclusioni. Se d'altro lato questa probabilità non è sufficientemente piccola allora la media campionaria osservata è un risultato plausibile e non si è contraddetta la premessa di base che  $\mu_0$  sia effettivamente la media della popolazione da cui è stato estratto il campione. Si giunge così alla seconda delle due conclusioni.

È necessario essere precisi e definire che cosa si intende per **probabilità sufficientemente piccola**. Questa definizione è realizzata scegliendo arbitrariamente un valore di probabilità e considerando questo sufficientemente piccolo; in molte situazioni si scelgono i valori sia del 5% sia dell'1%.<sup>5</sup> In particolare una soglia del 5% viene definita come **differenza significativa** mentre una dell'1% come **differenza molto significativa**.

Una scelta del 5% indica che una probabilità 0.05 (1 su 20) fornisce evidenza sufficiente per contraddire l'affermazione che  $\mu_0$  è la media della popolazione. Una scelta dell'1% (1 su 100) è naturalmente un valore più **conservativo**. Si deve ora attivare questa strategia. L'analisi consiste nel determinare, sotto l'assunto che il campione provenga da una popolazione la cui media è  $\mu_0$ , la probabilità di osservare uno scostamento dalla media  $\mu_0$  pari o maggiore a quello della media campionaria. A questo punto si confronta questa probabilità con la probabilità prefissata **sufficientemente piccola**. Sostituendo il termine **frequenza relativa** al termine **probabilità** si può dire che il problema consiste nel determinare la frequenza relativa con cui le medie di un campione di grandezza  $n$  si scostano da  $\mu_0$ , in misura uguale o maggiore della osservata media campionaria  $\bar{x}$ , quando  $\mu_0$  è la media della popolazione. Questo è di fatto, un quesito concernente le aree delle code della distribuzione di campionamento delle medie; aree quindi sotto la curva di Gauss. Quindi si calcola innanzitutto il valore  $z_{stat}$  del campione:

$$z_{stat} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \quad (8)$$

e quindi si usano le tavole degli integrali della distribuzione normale (se  $n$  è sufficientemente grande, in caso contrario si usano le tabelle della distribuzione *t di Student*) e si confronta il risultato con la prefissata soglia di probabilità **sufficientemente piccola**. L'inferenza non comporta la generazione di una distribuzione di campionamento delle medie! Infatti si usa ciò che è noto di questa distribuzione di campionamento (grazie alle relazioni con i dati sulla popolazione) per trovare la relazione tra l'osservazione di una singola media campionaria e la media della popolazione.

Inoltre la determinazione della probabilità dello scarto di una media campionaria pari o maggiore di quella osservata richiede che si specifichi se l'interesse sia rivolto alla **scarto in una sola direzione o in ambedue le direzioni** da  $\mu_0$ . Qualora si volesse eseguire un test a due code la probabilità sufficientemente piccola andrà divisa in parti uguali nelle due code, ciascuna del 2.5%. In sintesi un test di significatività sulla media richiede i seguenti passaggi logici:

<sup>4</sup>Come formalismo matematico si usa indicare i parametri della distribuzione normale specificando la varianza e non la DS.

<sup>5</sup>La soglia del 5% ha ragioni prettamente storiche. Fu il livello arbitrariamente scelto da William Gosset (vero nome di Student) quando iniziò ad usare i suoi test per controlli di qualità presso la birreria Guinness che lo aveva assunto. Gosset per divergenze insanabili non era riuscito a laurearsi in matematica e tutti i suoi studi li ha pubblicati con lo pseudonimo di *studente*.

- 1) La determinazione dell'ipotesi nulla  $H_0$ . Per esempio che la media campionaria sia uguale alla media della popolazione  $H_0 : \bar{x} = \mu_0$ . Va anche specificato l'ES della distribuzione campionaria, recuperando il valore di  $\sigma$  dalla popolazione.
- 2) La determinazione di un livello di probabilità  $P < \alpha$  sufficientemente piccolo da fornire una ragionevole evidenza contro  $H_0$ ; per esempio  $\alpha < 5\%$
- 3) La determinazione di un'ipotesi alternativa  $H_A$  soprattutto relativa a qual è lo scostamento di interesse sperimentale. Per esempio per un test a due code  $H_A : \bar{x} \neq \mu_0$
- 4) Il calcolo di una grandezza statistica, per esempio  $z_{stat}$
- 5) Il confronto tra il valore della grandezza statistica  $z_{stat}$  con quello teorico  $z_{crit}$  permette di prendere una decisione di rifiuto o di non rifiuto dell'ipotesi nulla  $H_0$ . In alternativa è possibile il calcolo esatto della probabilità  $P_{stat}$  da confrontare con il valore prefissato  $\alpha$  oppure la determinazione degli intervalli di confidenza (vedere più avanti).

## 4.2 Esempi di determinazione della significatività

**Esempio 4.1.** La popolazione è rappresentata dai livelli di acido urico serico di pazienti maschi affetti da diabete mellito. Il livello di acido urico serico nella popolazione è  $5.4[\frac{mg}{100ml}]$  di siero con  $\sigma = 1.0[\frac{mg}{100ml}]$ . Una probabilità del 5% è sufficientemente piccola. Interessano gli scostamenti in una sola direzione, relativamente a valori maggiori della media della popolazione, pertanto  $H_A : \bar{x} > \mu_0$ .

In un campione di 25 diabetici il livello medio di acido urico serico è  $5.9[\frac{mg}{100ml}]$ . Si vuole determinare se la media della popolazione è da rifiutare o se il campione è compatibile con tale valore. Calcolando la grandezza standard per  $5.9[\frac{mg}{100ml}]$  si ottiene:

$$z_{stat} = \frac{5.9 - 5.4}{\frac{1.0}{\sqrt{25}}} = 2.5$$

e dalla tabella 15 delle aree della curva di Gauss si legge che ciò comporta  $P_{stat} = 0.6\%$ , che è significativamente meno della soglia  $\alpha = 5\%$ . In tal caso si è portati a rifiutare il valore di  $5.4[\frac{mg}{100ml}]$  come valore medio della popolazione dell'acido urico serico per i diabetici. Il campione offre un'evidenza ragionevole per dire che i livelli di acido urico serico dei diabetici hanno una media della popolazione **superiore** a  $5.4[\frac{mg}{100ml}]$ . Si può anche verificare che per una probabilità del 5% il valore critico che si può trovare sulla tabella degli integrali di Gauss è  $z_{crit} = 1.645$ . Siccome il valore assoluto calcolato  $z_{stat} = 2.5 > 1.645 = z_{crit}$  è maggiore di quello critico, si può respingere  $H_0$ .

**Esempio 4.2.** La popolazione è rappresentata dai tempi di sopravvivenza di pazienti affetti da tumore che sono stati trattati con un nuovo farmaco. Il tempo medio di sopravvivenza nella popolazione è 38.3 mesi con  $\sigma = 43.3$  mesi. Una probabilità del 5% è sufficientemente piccola. Interessano scostamenti in ambedue le direzioni dalla media della popolazione. Un campione di 100 pazienti affetti da tumore e trattati con un farmaco ha una sopravvivenza media di 46.9 mesi. Calcolando il valore statistico  $z$  si ottiene che:

$$z_{stat} = \frac{46.9 - 38.3}{\frac{43.3}{\sqrt{100}}} = 1.99$$

e per tale valore l'area nelle code della curva di Gauss è pari al  $P_{stat} = 4.7\%$  (appena al di sotto della soglia  $\alpha = 5\%$ ). In effetti il valore critico al 5% in due code corrisponde quello ad una coda del 2.5% e che corrisponde a  $z_{crit} = 1.96$ . Siccome  $1.99 > 1.96$  il valore  $z_{stat}$  del test è superiore a quello  $z_{crit}$ . Conclusione: si rifiuta l'ipotesi nulla ( $H_0 : \bar{x} = \mu_0$ ) che il tempo medio di sopravvivenza dei pazienti affetti da tumore e trattati con il farmaco è 38.3 mesi. Vi è evidenza sufficiente che il tempo medio di sopravvivenza nella popolazione di pazienti affetti da tumore e trattati con il farmaco sia maggiore di 38.3 mesi.

**Esercizio 4.1.** Eseguire un test con i dati dell'esempio 4.2 con l'ipotesi che il farmaco uniamente allunghi la sopravvivenza.  $P_{stat} = 0.02351$

### 4.3 Due tipi di errori $\alpha$ e $\beta$

Il livello di significatività è già stato definito come il rischio di rifiutare erroneamente l'ipotesi nulla  $H_0$  quando in realtà è vera. Questo errore è chiamato l'errore  $\alpha$  o errore di I tipo. Tale terminologia implica che vi sia un errore  $\beta$ , o errore di II tipo. Si definisce come errore  $\beta$  la probabilità di non rifiutare erroneamente l'ipotesi nulla  $H_0$  quando questa in realtà è falsa, in quanto è vera invece un'ipotesi alternativa  $H_A$ . Per esempio, con i livelli di acido urico serico dei diabetici dell'esempio 3.1, si supponga che la media vera della popolazione sia realmente quella della ipotesi alternativa  $H_A : \mu_A = 6.0 \left[ \frac{mg}{100ml} \right]$ . In tale caso si sarebbe portati a preferire che un test di significatività della media che sotto l'ipotesi nulla è pari a  $H_0 : \mu_0 = 5.4 \left[ \frac{mg}{100ml} \right]$  producesse una conclusione **statisticamente significativa**. È possibile che si raggiunga la conclusione **statisticamente non significativa**. L'errore  $\beta$  è la probabilità di ottenere questa conclusione errata. Da notare che di per sé è possibile definire l'errore  $\beta$  solo in presenza di una precisa definizione dell'ipotesi alternativa con un valore  $\mu_A$ .

**Esempio 4.3.** Qual è l'errore  $\beta$  per lo studio dei livelli di acido urico serico dei diabetici? Si ricordi che con un campione di 25 unità è stato fatto un test di significatività ad una coda al livello del 5%. Il calcolo dell'errore di II tipo è illustrato nella figura 3 sottostante. Innanzitutto si determina quale valore medio per un campione di dimensione 25 non sarebbe in grado di rifiutare l'ipotesi nulla  $H_0$  quando questa è vera. Ciò implica di trovare il valore soglia corrispondente al 5% della coda di destra della distribuzione di campionamento delle medie assumendo che l'ipotesi nulla sia vera. A questo quesito si risponde con l'elaborazione sulla distribuzione di campionamento delle medie per mezzo della quale si ottenne un valore pari a  $5.73 \left[ \frac{mg}{100ml} \right]$ .

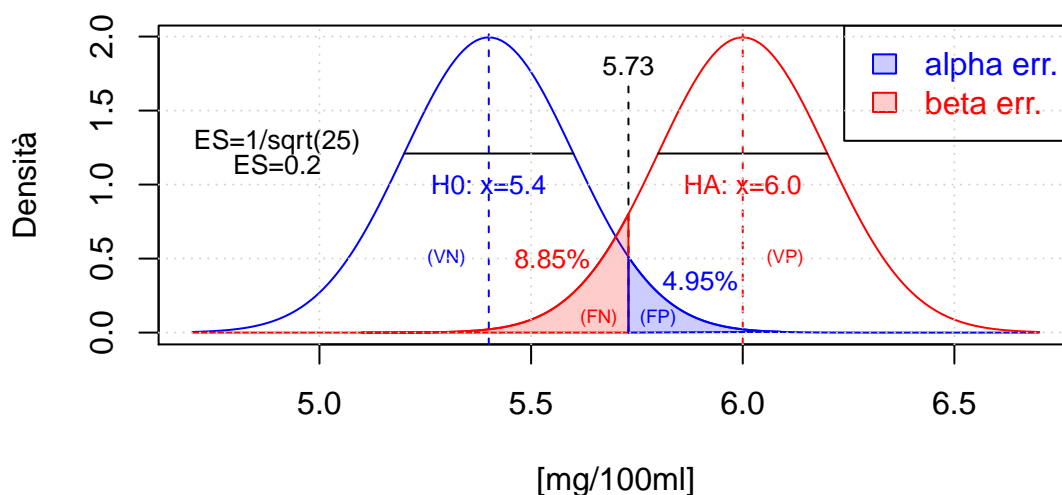


Figura 3: Errore alpha ed errore beta

Di conseguenza qualsiasi media di campione di grandezza 25 che è maggiore di  $5.73 \left[ \frac{mg}{100ml} \right]$  fornirà un'evidenza per rifiutare l'ipotesi nulla che la media sia  $5.4 \left[ \frac{mg}{100ml} \right]$ ; una qualsiasi media inferiore a tale valore non fornirà una base per rifiutare detta ipotesi nulla (linea tratteggiata nella figura 3). L'errore  $\beta$  o errore di II tipo può essere definito come la probabilità di non rifiutare l'ipotesi nulla quando è valida la media della popolazione sotto l'ipotesi alternativa  $H_A$ . In questo caso l'alternativa è  $\mu_A = 6.0 \left[ \frac{mg}{100ml} \right]$ . Ancora una volta l'interesse si concentra sulla distribuzione di campionamento delle medie di campioni di grandezza 25, ma ora con una media pari a  $6.0 \left[ \frac{mg}{100ml} \right]$  e precisamente sulla curva posta sul lato destro della tabella. Dato che una media campionaria inferiore a  $5.73 \left[ \frac{mg}{100ml} \right]$  non è in grado di fornire un'evidenza per il rifiuto dell'ipotesi nulla, la domanda è ora: quale parte della curva alternativa centrata su  $H_A$ , posta sul lato destro cade al di sotto di  $5.73 \left[ \frac{mg}{100ml} \right]$ ? Il rapporto  $z_{stat}$  è:

$$z_{stat} = \frac{5.73 - 6.0}{\frac{1}{\sqrt{25}}} = \frac{-0.27}{0.2} = -1.35$$

La tabella 15 della distribuzione normale ad una coda ci informa che l'area richiesta per questo valore di  $z_{stat}$  è pari a una probabilità  $P_{z_{stat}} = 0.089$ . Quindi l'errore II è 8.9% quando la media della popolazione è realmente  $6.0 \left[ \frac{mg}{100ml} \right]$ .

Con  è agevole calcolare direttamente la probabilità degli errori  $\alpha$  e  $\beta$ :

```
(soglia.c<-qnorm(0.95,mean=5.4,sd=0.2)) # Calcola il quantile normale per la coda di dx
## [1] 5.728971
(alpha<-1-pnorm(soglia.c,mean=5.4,sd=0.2)) # Verifica, ovviamente è il 5%
## [1] 0.05
(beta<-pnorm(soglia.c,mean=6,sd=0.2)) # Calcola l'errore beta
## [1] 0.08768546
```

La conseguenza della riduzione del livello di significatività che porta ad un più piccolo errore  $\alpha$  o di I tipo può essere facilmente interpretato. Per lasciare meno del 5% nella coda di destra della curva posta sul lato di sinistra della Figura 3, il punto corrispondente al valore soglia deve spostarsi alla destra di  $5.73[\frac{mg}{100ml}]$ . Ciò naturalmente aumenta l'area della coda di sinistra della curva posta sul lato di destra. Ne consegue che un **decremento nell'errore di I tipo provoca un incremento nell'errore di II tipo** o errore  $\beta$  e viceversa. L'errore di II tipo o errore  $\beta$  può essere diminuito solo a rischio di un incremento dell'errore di I tipo o errore  $\alpha$ .

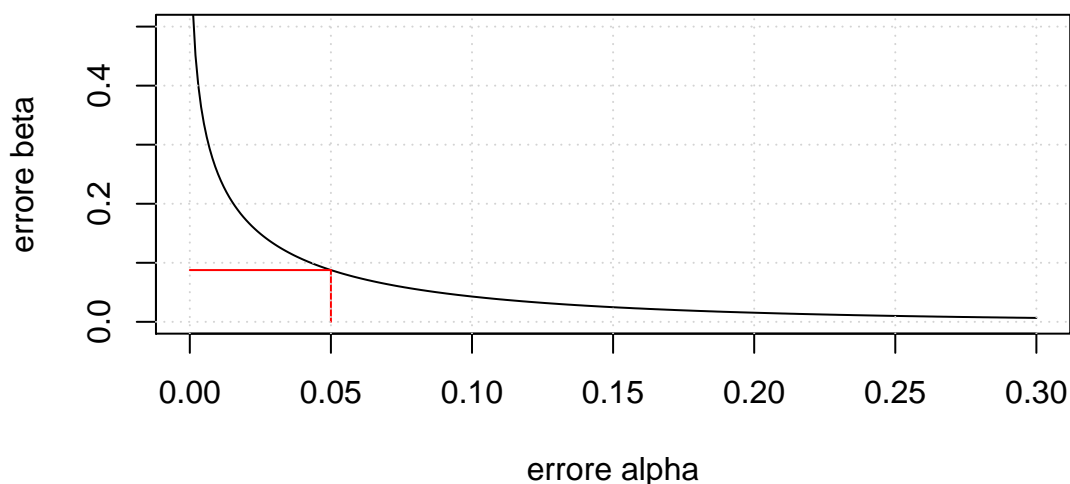


Figura 4: Legame tra errore alpha e errore beta

Nella figura 4 è rappresentato un grafico che mostra la variazione dell'errore  $\beta$  in funzione dell'errore  $\alpha$  relativo all'esempio 3.1. Si può notare che l'andamento non è lineare e che quindi scegliere un errore  $\alpha$  molto basso ha come conseguenza quella di alzare notevolmente l'errore  $\beta$ . La soglia storica del 5% nasce soprattutto da questa **esigenza di compromesso**.

Ancora più spesso come mostrato nella Figura 5 si rappresenta la **potenza** (o complemento dell'errore di II tipo)  $1 - \beta$  in funzione di  $\alpha$ ; questo tipo di grafico si chiama **ROC<sup>6</sup> (Receiver Operating Characteristic)** e viene usato soprattutto per determinare le soglie (cut off) di positività/negatività ad un dato test. In base alla soglia infatti variano le proporzioni di **falsi negativi** (errore  $\beta$ ) e **falsi positivi** (errore  $\alpha$ ). Si immagini come *Gedankenexperiment* di determinare la positività al diabete usando i dati sull'acido urico. Spostando il livello di  $5.73[\frac{mg}{100ml}]$  si possono aumentare/diminuire i falsi negativi e i falsi positivi indicati nella figura 3.

*Nota* . Per fissare un cut off ottimale solitamente si sceglie un punto il più vicino possibile al punto (0; 1). Incidono poi anche considerazioni sul rischio associato ad un falso positivo e a un falso negativo. L'interpretazione di tale rischio è chiaramente fortemente dipendente dalla situazione particolare del caso. Un falso negativo nella detezione di messaggi spam non ha lo stesso peso di un falso negativo per una malattia infettiva contagiosa.

<sup>6</sup><https://www.medcalc.org/manual/roc-curves.php> e [https://en.wikipedia.org/wiki/Receiver\\_operating\\_characteristic](https://en.wikipedia.org/wiki/Receiver_operating_characteristic)

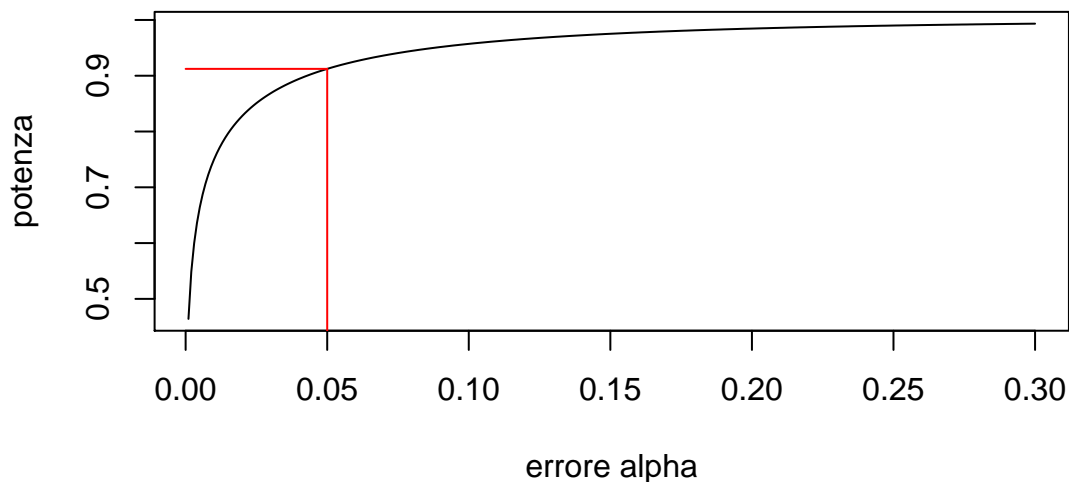
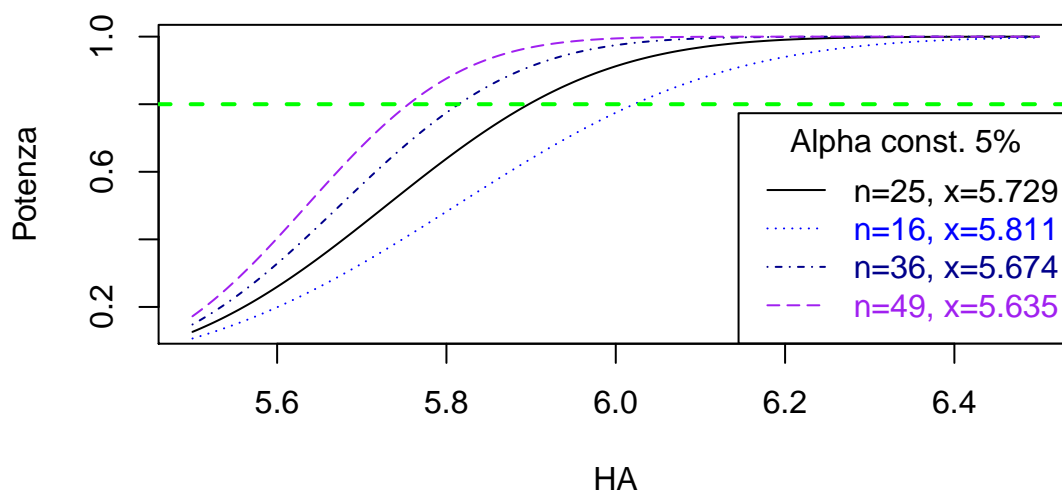


Figura 5: Curva ROC

Tabella 2: Riassunto sugli errori  $\alpha$  e  $\beta$ 

Conclusione del test di significatività	La media della popolazione è $\mu_0$ . Ipotesi nulla vera.	La media della popolazione non è $\mu_0$ ma $\mu_A$ . Ipotesi nulla non vera.
Statisticamente non significativo Non si rifiuta l'ipotesi nulla $H_0$	Conclusione corretta	Errore di II tipo o errore $\beta$
Statisticamente significativo si rifiuta l'ipotesi nulla $H_0$	Errore di I tipo o errore $\alpha$	Conclusione corretta

Si noti che l'errore di II tipo corrisponde ad una particolare media alternativa della popolazione che è denotata da  $\mu_A$ . Si potrebbe determinare la potenza per un insieme di valori di  $\mu_A$ . Per esempio, con i livelli di acido urico serico dei diabetici si potrebbe determinare per  $\mu_A = 5.5[\frac{mg}{100ml}]$ ,  $\mu_A = 5.6[\frac{mg}{100ml}]$ , e così via. In realtà si può costruire un grafico con i valori di  $\mu_A$  sull'ascissa e i corrispondenti valori di  $\beta$  sull'ordinata, o ancora meglio la potenza  $1 - \beta$ . Questo grafico fornisce quella che è nota come la curva operativa caratteristica (CO) del test. La curva operativa caratteristica per lo studio dei livelli di acido urico serico dei diabetici è mostrata nelle Figure 6 e 7.

Figura 6: Potenza in funzione di  $\mu_A$  con  $\alpha$  costante

Si noti l'effetto della dimensione del campione nella Figura 6. Siccome  $n$  influenza l'errore standard di entrambe le curve la potenza varia in modo molto marcato in funzione di  $n$ . Le quattro curve mostrano la potenza di poter discriminare una certa ipotesi alternativa  $H_A$  data da un certo valore  $\mu_A$  mantenendo costante  $\alpha = 5\%$ . Per esempio dalla lettura grafica, per una dimensione di campione  $n = 16$  l'80% della potenza si trova a circa  $6[\frac{mg}{100ml}]$ ; ciò significa che con un campione  $n = 16$  con media si ha l'80% di probabilità di confermare l'ipotesi  $H_A : \mu_A = 6[\frac{mg}{100ml}]$  rifiutando l'ipotesi nulla  $H_0 : \mu_A = 5.4[\frac{mg}{100ml}]$  al 5% (ipotesi che vedrebbe la sua soglia critica a  $x_{crit} = 5.811[\frac{mg}{100ml}]$ )

Se invece si vuole conoscere la potenza misurabile da una ipotesi  $H_A$  fissa, ignorando il test svolto sull'errore  $\alpha$  le curve un po' più vicine come mostrato dalla figura 7.

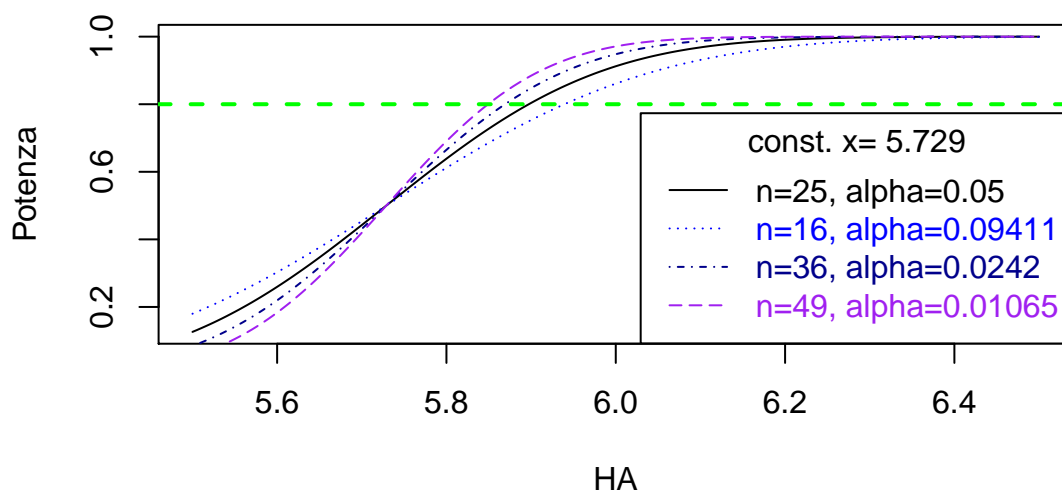


Figura 7: Potenza in funzione di  $\mu_A$  con la differenza di  $\mu$  fissa

In entrambi i grafici è mostrato il livello di potenza dell'80%: questo è il livello di riferimento abituale usato per determinare **la dimensione del campione** se si ha un'idea di quello che può essere  $H_A$ .

Relativamente ai due tipi di errore, un altro punto degno di nota è che la grandezza dell'errore di II tipo, per un prefissato errore di I tipo, dipende dal grado di sovrapposizione delle curve. Minore è la sovrapposizione, migliore è la situazione relativamente all'errore di II tipo. Ma allora che cosa può essere fatto per diminuire la sovrapposizione? A parte l'ovvio slittamento dell'intera curva posta sul lato di destra ancora più a destra, si può ridurre la sovrapposizione diminuendo la dispersione delle due curve. Ambedue le curve sono distribuzioni di campionamento delle medie con  $DS = \frac{\sigma}{\sqrt{n}}$ . Benché non si possa far nulla nei riguardi di  $\sigma$  (la variabilità da soggetto a soggetto propria della popolazione esaminata), si può aumentare  $n$ . Questo riduce  $\frac{\sigma}{\sqrt{n}}$  e di conseguenza la sovrapposizione. Pertanto per un prefissato errore di I tipo l'errore di II tipo diminuisce con l'aumento della dimensione del campione. A tal proposito si veda il paragrafo riguardante la determinazione della numerosità dei campioni.

#### 4.4 Sensibilità, specificità e valori predittivi

La discussione a proposito dell'errore  $\alpha$  e  $\beta$  ci porta a riflettere sulla definizione della sensibilità, della specificità e dei valori predittivi. Nella figura 3 sono marcati i quattro valori fondamentali: veri negativi (VN), veri positivi (VP), falsi negativi (FN) e falsi positivi (FP). Qualsiasi esame clinico è comunque basato su una soglia di detezione; c'è sempre un cut off, come quello della figura posizionato a  $5.7354[\frac{mg}{100ml}]$ . Considerando la curva blu come il totale dei sani (accertati come gold standard) e la curva rossa dei malati (anch'essi accertati come gold standard) e tenendo a mente che le aree totali di queste due curve sono normate a 1 si ha:

**Definizione 4.1.** Qui le definizioni più importanti:

- la sensibilità è il rapporto tra i veri positivi e il totale dei malati (totale della curva rossa, somma di veri positivi e falsi negativi)  

$$Sens = \frac{VP}{VP+FN} = \frac{1-\beta}{(1-\beta)+\beta} = 1 - \beta$$
. In pratica in questo caso è la *potenza* del test;
- la specificità è il rapporto tra i veri negativi e il totale dei sani (totale della curva blu, somma di

veri negativi e falsi positivi)

$Spec = \frac{VN}{VN+FP} = \frac{1-\alpha}{(1-\alpha)+\alpha} = 1 - \alpha$ . In pratica il complemento all'errore  $\alpha$ ;

- il valore predittivo positivo è il rapporto tra i veri positivi e il totale dei positivi (somma di veri positivi e falsi positivi)  $VPP = \frac{VP}{VP+FP} = \frac{1-\beta}{(1-\beta)+\alpha}$
- Il valore predittivo negativo è il rapporto tra i veri negativi e il totale dei negativi (somma di veri negativi e falsi negativi)  $VPN = \frac{VN}{VN+FN} = \frac{1-\alpha}{(1-\alpha)+\beta}$

Chiaramente  $\alpha$  e  $\beta$  potrebbero anche essere scambiate nella logica sperimentale se si concentrasse il test sulla determinazione dei malati anziché dei sani, cioè se si scambiasse il ruolo dell'ipotesi nulla e di quella alternativa. Ma in fin dei conti è unicamente un problema di nomenclatura. "L'uomo è misura delle cose"

Tutti questi valori si trovano confinato nell'intervallo  $I = [0; 1]$ . In generale avendo unicamente conteggi per le 4 possibili osservazioni tutti questi valori rientrano nella teoria sulle proporzioni e sui test sulle proporzioni. Si rimanda il lettore allo specifico paragrafo 11.1.

## 5 La significatività per i piccoli campioni e la curva di Student

### 5.1 Quando $\sigma$ della popolazione non è noto...

Nei casi precedenti  $\mu$  e  $\sigma$  della popolazione erano noti. Quando si ha a che fare con piccoli campioni questo nella maggior parte dei casi non è dato e si deve ricorrere agli stimatori. Per la media non sussistono grossi problemi visto che la media campionaria è uno stimatore corretto della media della popolazione. Per la stima della DS si deve invece ricorrere alla DS corretta campionaria  $\hat{s}$ . Per queste inferenze valgono tutte le regole precedentemente viste ma si dovrà utilizzare la **distribuzione di Student** al posto di quella di Gauss. Si noti come nella tabella 16 della curva di Student si trovano i valori associati ai **gradi di libertà** (*degree of freedom* in inglese). Sostanzialmente i gradi di libertà corrispondono al numero di osservazioni indipendenti che si hanno. Se si ha un campione di 30 pazienti di cui si conosce la glicemia si posso calcolare media  $\bar{x}$  e scarto corretto  $\hat{s}$ . Conoscendo la media e i primi 29 pazienti, il trentesimo paziente non è più libero di avere una glicemia qualsiasi. Quindi i gradi di libertà sostanzialmente sono dati dal numero di campioni meno il numero di vincoli (di solito 1) già noti (in questo caso la media).

Si noti anche che sono possibili infinite curve di Student ma con  $n$  crescente esse si approssimano asintoticamente alla curva normale. A seconda di vari autori tra loro discordanti, a partire da campioni sufficientemente grandi (di solito  $n > 100$  o  $n > 200$ , per taluni  $n > 30$ ) si può tornare ad utilizzare la tabella 15 di Gauss. Tuttavia con l'ampia disponibilità di calcolatori elettronici questa necessità è meno importante.

Inoltre si noti come siano riportati solo i valori relativi a determinate percentuali dell'area (0.95, 0.99, 0.999, ecc); questo perché col tempo si sono costituite delle consuetudini sulla significatività delle misure statistiche (campione significativo (95%), molto significativo (99%), ecc). A livello cartaceo sarebbe proibitivo riportare gli integrali come per la curva normale. Anche qui chiaramente il supporto informatico ha parecchio cambiato le consuetudini.

**Esempio 5.1.** Analizzando un campione di  $n = 10$  studenti fu trovato avere una media  $\bar{x} = 68.7$  e una DS corretta  $\hat{s} = 8.67$  battiti per minuto. Un manuale di valori clinici normali indica che la frequenza media di pulsazione per i maschi giovani è di  $\mu_0 = 72$  battiti per minuto. Questi dati sono compatibili con la norma riportata dal manuale? È questa media campionaria verosimilmente derivata da una popolazione la cui media è di 72 battiti per minuto? Si può eseguire un test di significatività, precisamente un test  $t$ .

L'ipotesi nulla  $H_0$  è che la media riportata dal libro ( $\mu_0 = 72$  battiti) sia corretta; si vuole vedere se la rilevazione sui dieci studenti concorda con questa ipotesi oppure no; il test è eseguito al livello di significatività del 5% e l'ipotesi alternativa è  $H_A : \mu_A \neq \mu_0$ , quindi il test è bilaterale con un'area del 2.5% per ogni coda. Si usa la stessa formula dei test precedenti ottenendo però un valore che va letto sulla distribuzione di Student.

$$t_{stat} = \frac{\bar{x} - \mu_0}{\frac{\hat{s}}{\sqrt{n}}} \quad (9)$$

Quindi sostituendo:

$$t_{stat} = \frac{68.7 - 72}{\frac{8.67}{\sqrt{10}}} = -1.20$$

Il valore assoluto di  $|t_{stat}| = 1.20$  è inferiore a quello tabulato per l'area di 0.975 e 9 gradi di libertà nella tabella 16 di Student ( $t_{crit,9g.l,0.975} \pm 2.26$ ). Con questi dati non si può rifiutare l'ipotesi nulla  $H_0$ .

## 6 Gli intervalli di confidenza e l'inferenza sulle medie

**Definizione 6.1.** Gli intervalli di confidenza (CI) (*confidence intervals* in inglese) sono quei limiti entro il quale si deve trovare il valore dell'ipotesi  $H_0$  affinché quest'ultima non sia rifiutata.

Nel caso del test  $t$  si calcolano con la seguente formula:

$$\bar{x} \pm t_{n-1,\alpha} \cdot \frac{\hat{s}}{\sqrt{n}} \quad (10)$$

**Esempio 6.1.** Prendiamo l'esempio precedente 5.1 e chiediamoci: entro quali limiti può oscillare  $\bar{x}$ . Considerando come accettabile il fatto che l'errore sia del 5% si hanno i seguenti limiti di confidenza:

$$68.7 \pm 2.2622 \cdot \frac{8.67}{\sqrt{10}}$$

E quindi l'intervallo di confidenza è  $I = [62.5, 74.9]$ . In altre parole vi è una certezza del 95% che la velocità di pulsazione media nella popolazione da cui il campione è estratto giace entro questi limiti. Siccome l'intervallo  $I$  comprende anche il valore  $\mu_0 = 72$  ciò significa che non c'è una differenza significativa e l'ipotesi nulla  $H_0$  non può essere respinta.

Si noti che aumentare la numerosità del campione ha un duplice effetto sulla riduzione dei limiti di confidenza. Non solo diminuisce il rapporto  $\frac{\hat{s}}{\sqrt{n}}$  ma anche in valore di  $t_{n-1,\alpha}$ . Si osservi a tal proposito la variazione dei valori in verticale nella rispettiva colonna della tabella @ref(tab:tabstudent. Ne consegue che analizzando campioni più numerosi si restringono le possibilità di errore dovute alla naturale fluttuazione statistica dei dati. Si noti inoltre che questo metodo di calcolo è spesso usato per il procedimento inverso, cioè per la determinazione di un valore medio della popolazione a partire dai dati di un campione.

## 7 Confronto tra medie

Scopo di questo paragrafo è di estendere gli argomenti trattati nel paragrafo precedente al confronto di 2 medie campionarie. Dato che la maggior parte della scienza è, per sua natura, comparativa, questa è una situazione molto diffusa, più comune di quella delle singole medie del paragrafo precedente.

In generale vi sono due campioni di osservazioni da due popolazioni sottostanti (spesso gruppi di soggetti sottoposti a trattamento e di soggetti di controllo) le cui medie sono denotate da  $\mu_T$  e  $\mu_C$  e le cui deviazioni standard sono denotate da  $\sigma_T$  e  $\sigma_C$ . Ricordando che un test di significatività comporta un'ipotesi nulla che specifica i valori per le quantità della popolazione, la rilevante ipotesi nulla e che le medie siano identiche, cioè  $H_0 : \mu_T = \mu_C$  oppure  $H_0 : \mu_T - \mu_C = 0$

La logica per il test di significatività è identica a quella presentata in precedenza. Assumendo che l'ipotesi nulla  $H_0$  sia vera (in altri termini che non vi sia differenza nelle medie delle due popolazioni) si calcola la probabilità di ottenere differenze tra le medie campionarie così grandi o ancor maggiori di quelle osservate. Se questa probabilità è *sufficientemente piccola*, vi è una ragionevole evidenza per dubitare della validità dell'ipotesi nulla; quindi si conclude che vi è una differenza tra le medie delle due popolazioni. In altri termini si rifiuta l'ipotesi nulla.

### 7.1 Medie non indipendenti

In molti casi i dati del gruppo di controllo sono collegati a quelli del gruppo sperimentale:



- Auto accoppiamento: gli stessi soggetti ricevono prima un trattamento con un farmaco e poi un trattamento con un altro farmaco o con un placebo<sup>7</sup>.
- Situazioni naturalmente appaiate: i soggetti di sperimentazione e di controllo estratti da una stessa nidia di cavie (topolini); gruppi di gemelli omozigoti, uno ammalato e uno sano; ragazzi ammalati di una data malattia che si trovano in una data classe e come controllo ragazzi sani della stessa classe, ecc.
- Appaiamento artificiale creato dallo sperimentatore: per esempio studiando l'efficacia sui i parti prematuri di una determinata procedura si ricercano soggetti simili con simili fattori di rischio (età, pressione sanguigna della madre, ecc)

In tutti questi casi si valuta solitamente la **differenza** tra il campione e il suo elemento di controllo, come illustrato dall'esempio seguente:

**Esempio 7.1.** Studio su pazienti ipertesi. A ciascuno di 11 pazienti ipertesi sono stati somministrati molti farmaci in momenti differenti. I risultati qui presentati sono quelli del placebo in confronto con l'idroclorotiazide. Ciascun paziente ha ricevuto per primo il placebo e successivamente il farmaco. La pressione sanguigna è stata registrata in modo identico in questi due momenti (placebo e farmaco). I dati riportati sono pressioni sistoliche. Sulla base di questi risultati sperimentali vi è una evidenza di una differenza nella pressione media sistolica legata a questi due trattamenti?

L'interesse è in un test a due code. Vi è chiaramente una situazione di auto appaiamento dato dall'identità del paziente così che l'interesse si concentra sulla differenza inter soggetto. In un contesto **ANOVA** si direbbe che l'appaiamento è dato dal **fattore di blocco** (*blocking factor* in inglese) costituito dall'identità del soggetto: ogni soggetto corrisponde ad un livello del fattore.

Tabella 3: Tabella dati autoappaiati

Paziente	Placebo	Idroclorotiazide	Differenza
1	211	181	30
2	210	172	38
3	196	210	-14
4	203	191	12
5	196	167	29
6	190	161	29
7	191	178	13
8	160	177	-17
9	173	149	24
10	170	119	51
11	163	156	7

L'ipotesi nulla in questo caso è che la media delle differenze della popolazione sia nulla. Cioè  $H_0 : \delta_0 = 0$ . In questo modo si procede con un normalissimo test  $t_{stat}$  sulle **differenze**. Con la funzione `summary()` di **R** si estraggono i valori necessari per il test.

```
summary(test.dip)
##      Paziente      Placebo      Idroclorotiazide      Differenza
## Min.   : 1.0   Min.   :160.0   Min.   :119.0   Min.   : -17.00
## 1st Qu.: 3.5   1st Qu.:171.5   1st Qu.:158.5   1st Qu.:  9.50
## Median : 6.0   Median :191.0   Median :172.0   Median : 24.00
## Mean   : 6.0   Mean   :187.5   Mean   :169.2   Mean   : 18.36
## 3rd Qu.: 8.5   3rd Qu.:199.5   3rd Qu.:179.5   3rd Qu.: 29.50
## Max.   :11.0   Max.   :211.0   Max.   :210.0   Max.   : 51.00
```

$$t_{stat} = \frac{\bar{d} - \delta_0}{\frac{\hat{s}}{\sqrt{n}}} = \frac{18.364 - 0}{\frac{20.834}{\sqrt{11}}} = 2.923$$

e i limiti di confidenza per le differenze saranno:

<sup>7</sup>In realtà gli esperimenti di questo tipo più moderni sono più complessi in quanto va anche determinato se c'è un effetto nell'ordine della somministrazione; gli schemi passano da alternanze binarie TC/CT ad alternanze ternarie TTC TCC ecc. Il programma non include la discussione di questi modelli.

$$\bar{d} \pm t_{10,0.05} \cdot \frac{\hat{s}}{\sqrt{n}} = 18.36 \pm 2.2281 \cdot \frac{20.834}{\sqrt{11}} = 18.36 \pm 14$$

Quindi l'intervallo è:  $I = [4.367; 32.36]$

### 7.1.1 Tre metodi equivalenti per l'inferenza

- Siccome  $t_{stat} = 2.9234$  è superiore al valore critico  $t_{crit,10,0.05} = 2.2281$  preso dalle tabelle, si rifiuta l'ipotesi nulla  $H_0 : \delta_0 = 0$ .
- Siccome l'intervallo di confidenza  $I = [32.36; 4.367]$  non contiene l'ipotesi nulla  $H_0 : \delta_0 = 0$ , questo è un secondo modo di svolgere l'inferenza rifiutando  $H_0$ .
- Siccome la probabilità di realizzare  $\bar{d} = 24$  con  $\delta_0 = 0$  e  $\hat{s} = 20.834$  è di  $P_{stat} = 0.01521$  e quindi  $P_{stat} \ll \alpha = 0.05$  anche questo è un modo per falsificare l'ipotesi nulla.

### 7.1.2 Test con R

In  è estremamente semplice eseguire un test  $t$ , basta usare la funzione `t.test()` come illustrato qui sotto

```
head(test.dip) # Controllare la struttura dati
##   Paziente Placebo Idroclorotiazide Differenza
## 1         1      211             181          30
## 2         2      210             172          38
## 3         3      196             210         -14
## 4         4      203             191          12
## 5         5      196             167          29
## 6         6      190             161          29
t.test(test.dip$Placebo, test.dip$Idroclorotiazide, # Dati
       paired = T, # Da specificare !!!
       alternative="two.sided", # In realtà per default è già a due code; si può omettere
       mu = 0, # Anche questo è un valore di default che può essere omesso.
       )
##
## Paired t-test
##
## data: test.dip$Placebo and test.dip$Idroclorotiazide
## t = 2.9234, df = 10, p-value = 0.01521
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.367181 32.360092
## sample estimates:
## mean of the differences
##                18.36364
```

## 7.2 Medie indipendenti

In molti casi è troppo *artificiale* cercare degli appaiamenti. Talvolta non è proprio possibile. Ammettiamo per esempio di avere un gruppo di pazienti che hanno avuto un tipo di trattamento e un gruppo di pazienti che hanno ricevuto un placebo e i due gruppi sono costituiti da persone diverse. Si avranno due medie di risultati ( $\mu_T$  e  $\mu_C$  per le popolazioni e  $\bar{x}_T$  e  $\bar{x}_C$  per il campione) e associate ad esse due deviazioni standard ( $\sigma_T$  e  $\sigma_C$  se sono note le DS della popolazione oppure più frequentemente  $\hat{s}_T$  e  $\hat{s}_C$  se sono note solamente le DS corrette dei due campioni).

Non si può procedere alla costruzione di una tabella delle differenze in quanto i dati non sono accoppiati e per di più non **sono necessariamente ugualmente numerosi**:  $n_T$  non è per forza uguale a  $n_C$ .

Per valutare se la differenza tra i due gruppi è significativa si può comunque trovare un metodo. Innanzitutto si può porre l'ipotesi  $H_0$ : “i gruppi di terapia e di controllo hanno lo stesso risultato” e quindi matematicamente la sottrazione delle due medie deve essere pari a zero:  $\mu_T - \mu_C = 0$  per la popolazione e  $\bar{x}_T - \bar{x}_C = 0$  per il campione.

Il problema più grosso è la valutazione dell'ES per il calcolo di  $t_{stat}$ ; questo perché abbiamo due DS distinte per campione e non un'unica DS per il *sistema*. Fortunatamente ci viene in aiuto un teorema sugli errori standard che esprime questa uguaglianza<sup>8</sup>:

$$ES(\bar{x}_T - \bar{x}_C) = \sqrt{[ES\bar{x}_T]^2 + [ES\bar{x}_C]^2} \quad (11)$$

<sup>8</sup>Deriva direttamente dal teorema di additività della varianza.

Quindi avendo a disposizione gli scarti quadratici delle popolazioni ciò diventa:

$$ES(\bar{x}_T - \bar{x}_C) = \sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}} \quad (12)$$

e il calcolo del valore  $t$ :

$$t_{stat} = \frac{\bar{x}_T - \bar{x}_C - 0}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}} \quad (13)$$

Tuttavia raramente sono noti gli scarti quadratici delle popolazioni; nella maggioranza dei casi saranno note unicamente gli scarti quadratici corretti dei campioni  $\hat{s}$ , che saranno usati come stima degli scarti quadratici delle popolazioni e quindi si dovrà anche passare all'utilizzo della curva di Student (come precedentemente trattato) e questo test allora sarà chiamato: test  $t$  per il confronto di due medie indipendenti.

Prima di procedere vi è ancora un'altra difficoltà che riguarda le DS delle popolazioni,  $\sigma_T$  e  $\sigma_C$ . Il test  $t$  per medie indipendenti, che prolifera nella letteratura medica, si basa sull'assunto ulteriore che le DS nelle popolazioni sottostanti sono **omoschedastiche**, cioè che  $\sigma_T \approx \sigma_C$ ; non devono essere uguali ma devono appartenere allo stesso potenziale pool di  $\sigma$ . Questo fatto deve essere verificato sempre prima dell'esecuzione del test  $t$  con ulteriori appropriati test presentati nel paragrafo 8.2 (ad esempio il test F di Fisher, il test di Bartlett, o anche il test di Levene). Quando le DS non sono uguali, vi sono difficoltà teoriche nello sviluppare un test  $t$  appropriato e spesso si finisce per ricorrere a metodi non parametrici come mostrato nel paragrafo 9. Ne consegue che questo testo assume, da questo momento in poi, che  $\sigma_T$  e  $\sigma_C$  siano omoschedastiche.

Ci si può chiedere se questo è un assunto ragionevole. Il fondamento logico sta nel fatto che, in molte situazioni, l'applicazione di un nuovo trattamento ad un gruppo, mentre l'altro gruppo è lasciato come controllo, tende al massimo ad influenzare il valore medio e a lasciare immutata la variabilità. L'esperienza ha posto questo in evidenza.

Con un  $\sigma$  comune, per ciascuna delle popolazioni dei trattati e dei controlli, si dimostra che l'errore standard della differenza tra le medie diviene:

$$ES(\bar{x}_T - \bar{x}_C) = \sqrt{\sigma^2 \cdot \left(\frac{1}{n_T} + \frac{1}{n_C}\right)} \quad (14)$$

Rimane ora da ottenere la stima di  $\sigma$ , basata sui risultati campionari. Si hanno due stime: la varianza campionaria nel gruppo trattato,  $\hat{s}_T$ , e quella nel gruppo di controllo,  $\hat{s}_C$ . Si potrebbe suggerire di eseguire una media semplice delle due. Si dimostra che ciò è inappropriato: dato che le grandezze campionarie possono differire sostanzialmente nei due gruppi, una varianza campionaria potrebbe essere una stima di  $\sigma$  molto più affidabile dell'altra. Pertanto è più appropriata una media ponderata di  $\hat{s}_T$  e  $\hat{s}_C$ , avente come pesi quantità che dipendono, in un certo qual modo, dall'affidabilità di ciascuna varianza campionaria. Matematicamente si può verificare che i pesi ottimali sono i g.l. (gradi di libertà) di ciascuna varianza campionaria; precisamente  $n_T - 1$  e  $n_C - 1$ . La stima risultante combinata di  $\sigma^2$  è chiamata la *stima pooled della varianza comune*:

$$\hat{s}_{pooled}^2 = \frac{(n_T - 1)\hat{s}_T^2 + (n_C - 1)\hat{s}_C^2}{(n_T - 1) + (n_C - 1)} \quad (15)$$

O altrimenti riscritta come:

$$\hat{s}_{pooled}^2 = \frac{\sum (x_T - \bar{x}_T)^2 + \sum (x_C - \bar{x}_C)^2}{n_T + n_C - 2} \quad (16)$$

<sup>9</sup>È più facile e intuitivo introdurre il concetto di omoschedasticità discutendo di regressione lineare. Questo concetto verrà ripreso più avanti.

Quindi usando la distribuzione di Student per il rapporto critico si ha:

$$t_{stat} = \frac{\bar{x}_T - \bar{x}_C - 0}{\sqrt{\hat{s}_{pooled}^2 \cdot \left(\frac{1}{n_T} + \frac{1}{n_C}\right)}} \quad (17)$$

e per i limiti di confidenza:

$$(\bar{x}_T - \bar{x}_C) \pm t_{crit, n_T+n_C-2, \alpha} \cdot \sqrt{\hat{s}_{pooled}^2 \cdot \left(\frac{1}{n_T} + \frac{1}{n_C}\right)} \quad (18)$$

**Esempio 7.2.** Si intende verificare la differenza tra due gruppi di Cladocero Daphnia magna (del quale sono stati misurate le dimensioni di 5 individui per il campione 1 e 7 individui per il campione 2, come riportato nella tabella 4 dopo 20 giorni dalla schiusa delle uova: si vuole investigare se hanno raggiunto dimensioni medie significativamente differenti.

Tabella 4: Tabella delle misure di Cladocero Daphnia in [mm]

misura	4.290	3.900	3.783	3.900	4.095	3.120	3.112	3.120	3.847	3.081	3.042	3.742
gruppo	A	A	A	A	A	B	B	B	B	B	B	B

Si estraggono i necessari parametri

Tabella 5: Tabella riassuntiva Cladocero Daphnia

	A	B
media	3.993600	3.294857
sd	0.200006	0.343768
varianza	0.040002	0.118176
n	5.000000	7.000000

Si calcola prima  $\hat{s}_{pooled}^2$  e in seguito il valore  $t_{stat}$

$$\hat{s}_{pooled}^2 = \frac{(5-1) \cdot 0.040002 + (7-1) \cdot 0.11818}{(5-1) + (7-1)} = 0.086907$$

$$t_{stat} = \frac{3.9936 - 3.2949 - 0}{\sqrt{0.086907 \cdot \left(\frac{1}{5} + \frac{1}{7}\right)}} = 4.0479$$

Il valore di  $t_{stat}$  (assoluto) è molto superiore a quello critico per 10 g.l. a due code  $t_{crit, n_T+n_C-2, \alpha} = 2.2281$  quindi l'ipotesi nulla  $H_0 : \bar{x}_C - \bar{x}_T = 0$  (che la differenza sia nulla) è respinta; la probabilità è pari a  $P_{H_0} = 0.002332$ , molto inferiore al prefissato  $\alpha = 0.05$

Questo test può essere eseguito comodamente con  come mostrato dal code chunk qui sotto riportato:

```
cd.a
## [1] 4.290 3.900 3.783 3.900 4.095
cd.b
## [1] 3.120 3.112 3.120 3.847 3.081 3.042 3.742
t.test(cd.a, cd.b, # Dati
       paired=F, # Non accoppiati
       var.equal = T # Importante, altrimenti viene eseguito il test nella variante di Welch
       )
##
## Two Sample t-test
##
## data: cd.a and cd.b
## t = 4.0479, df = 10, p-value = 0.002331
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
## 0.3141287 1.0833570
## sample estimates:
## mean of x mean of y
## 3.993600 3.294857
```

Nota .  $\mathbb{R}$  di default eseguirebbe la variante di Welch<sup>10</sup> del test  $t$ . Tale variante permette un calcolo un po' più attendibile (ma non troppo) se il vincolo di omoschedasticità è violato.

## 8 Verifica dei prerequisiti per i test $t$

I test  $t$  sono molto popolari e vengono usati in modo spesso **troppo superficiale**. Infatti per ogni test  $t$  è necessario verificare che i prerequisiti siano validi. In particolare vanno verificate:

- la normalità dei dati nei gruppi;
- l'omoschedasticità delle deviazioni standard;
- l'indipendenza dei valori tra loro; ha a che fare con l'ordine, le serie temporali e i fenomeni di autocorrelazione evidenziabili per esempio con i plot ACF e PACF (non trattata in questa scheda).

### 8.1 Verifica della normalità

Ci sono tanti test che permettono verificare se un gruppo di dati ha una distribuzione normale. Sono tutti da considerare “**old styled statistics**”; non andrebbero usati. Questo perché quando si costruisce l'esperimento statistico il desiderio è quello di **non falsificare** l'ipotesi nulla  $H_0 : X \sim N(\bar{x}, \sigma^2)$  che i dati siano normali. Il fatto di non respingere l'ipotesi nulla **non permette** di dire che i **dati sono veramente normali**, ma permette unicamente di dire che **non c'è evidenza che non lo siano**; e questa “**non evidenza**” dipende unicamente dal fatto che molto probabilmente il data set è troppo piccolo. In un data set di un milione di osservazioni anche la più piccola deviazione dalla normalità verrebbe registrata.

Nota . “**Absence of evidence is not evidence of absence**” è una delle massime da tenere sempre bene a mente. Solamente un test che **rifiuta un'ipotesi nulla**  $H_0$  può essere considerato **significativo**. **Un non rifiuto non equivale ad una accettazione**.

Quindi se incontrate test della normalità come *Komgorov-Smirnov*, *Lilliefors*, *Shapiro-Wilk*, *Cramér-von-Mises*, ecc. sappiate che sono superati. Sono implementati anche in  $\mathbb{R}$  ma non ha più senso presentarli. La prassi della statistica moderna prevede che quando si vuole “**non falsificare**” una ipotesi nulla si usino unicamente argomenti **grafici** e/o “**occhiometrici**”.

Per verificare la normalità di una distribuzione si usa il grafico chiamato **Q-Q plot**. Su un asse si riportano i quantili normali (il valore  $z$  dei singoli dati in base alla media e alla deviazione standard del campione) e sull'altro la grandezza assoluta come mostrato nella figura 8. La figura ha due plot pressoché identici: quello di sinistra ottenuto calcolando i parametri a mano e quello di destra ottenuto con le funzioni `qqnorm()` e `qqline()` del pacchetto base di  $\mathbb{R}$ .

```
# Q-Q plot tracciato a mano"
par(mfrow=c(1,2)) # Due grafici nella stessa figura
cd.b.s <- sort(cd.b)
i<-1:length(cd.b.s)
cd.r<-(i-0.5)/length(cd.b) # Quantile empirico. Si può calcolare automaticamente con ppoints()
cd.q<-qqnorm(cd.r)
plot(cd.q,cd.b.s,main="Q-Q plot manuale")
f.q<-lm(cd.b.s ~ cd.q)
abline(f.q,col="red")
## Q-Q plot automatico
cd.qq<-qqnorm(cd.b) # Funzione integrata
cd.qql<-qqline(cd.b) # La regressione qui è robusta
```

<sup>10</sup>[https://en.wikipedia.org/wiki/Welch%27s\\_t-test](https://en.wikipedia.org/wiki/Welch%27s_t-test)

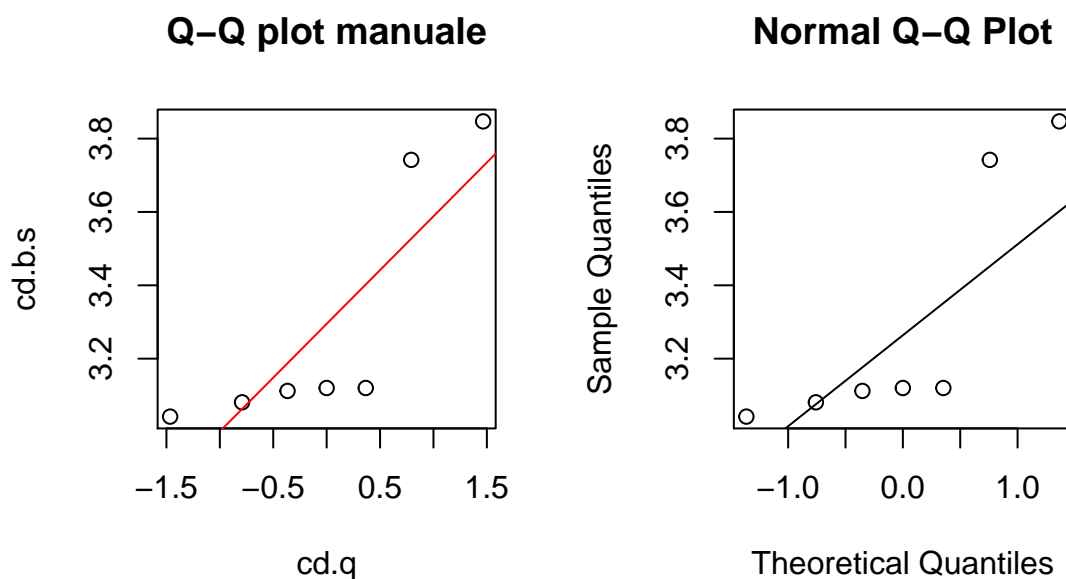


Figura 8: Q-Q plot manuale vs automatizzato con `qqnorm()` e `qqline()`

In alternativa nel pacchetto `car` è disponibile la funzione `qqPlot()` che aggiunge anche gli intervalli di confidenza della regressione. Qui sotto nella figura 9 i plot per tutti e due i gruppi del test  $t$  dell'esempio 7.2.

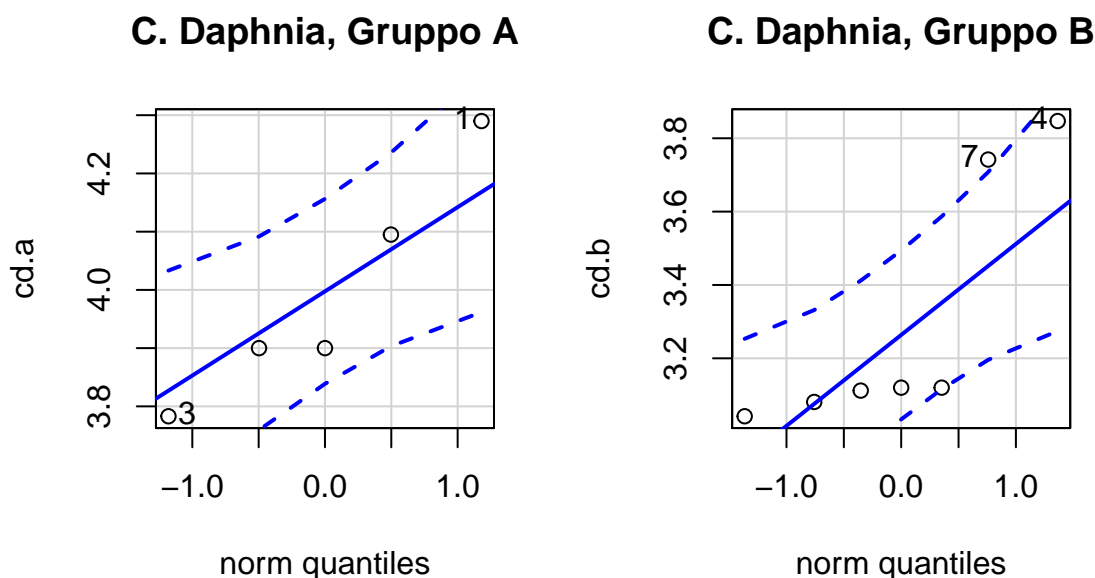


Figura 9: Q-Q plot tracciati con la funzione `qqPlot()` di `car`

I dati sono veramente pochi per verificare in modo attendibile che i punti siano posizionati su una retta. Tuttavia per il gruppo B si può quasi pensare che la normalità non sia garantita. Questa violazione è già sufficiente per rifiutare i risultati del test. In questo specifico caso sarebbe meglio ricorrere ad un metodo non parametrico (vedi paragrafo 9).

## 8.2 Test per l'omoschedasticità

Anche questi test dovrebbero essere **deprecati** a favore di analisi “occhiometriche”. Purtroppo però per validare i test  $t$  non sono disponibili metodi grafici<sup>11</sup>. Pertanto vengono qui presentati il test  $F$  per l'omogeneità delle varianze (da non confondere con l'ANOVA!) e il test di *Levene*.

<sup>11</sup>Se invece si costruisse un test ANOVA, perfettamente equivalente al test  $t$ , si potrebbe procedere graficamente con l'analisi del Tukey-Anscombe plot già usato nell'ambito della regressione.

### 8.2.1 Test F

Per la verifica degli assunti del test  $t$ , la omoschedasticità tra due gruppi (A e B) è verificata con un test bilaterale, dove l'ipotesi nulla  $H_0 : \sigma_A = \sigma_B$  e l'ipotesi alternativa  $H_A : \sigma_A \neq \sigma_B$ . In realtà per precisione, trattandosi di un test  $F$  l'ipotesi nulla sarebbe da indicare come  $H_0 : \frac{\sigma_A}{\sigma_B} = 1$  e quella alternativa  $H_A : \frac{\sigma_A}{\sigma_B} \neq 1$

Il test  $F$ <sup>12</sup> bilaterale, il primo ad essere proposto e tuttora il più diffuso, è fondato sul rapporto tra la varianza campionaria  $\hat{s}_{max}^2$  maggiore e la varianza campionaria minore  $\hat{s}_{min}^2$  (uno dei due gruppi ha la varianza più grande dell'altra), secondo la formula:

$$F_{(n_{max}-1), (n_{min}-1), \alpha} = \frac{\hat{s}_{max}^2}{\hat{s}_{min}^2} \quad (19)$$


Fondato sul fatto che sia vera l'ipotesi nulla, il valore dovrebbe essere uguale a 1 ma che chiaramente essendo esiguo il campione potrebbe casualmente (o significativamente) essere diverso da 1.


L'esecuzione in  è estremamente semplice:

```
var.test(cd.a, cd.b)
##
## F test to compare two variances
##
## data: cd.a and cd.b
## F = 0.3385, num df = 4, denom df = 6, p-value = 0.3138
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.0543582 3.1132646
## sample estimates:
## ratio of variances
##      0.3384973
```

Come si può vedere in questo esempio non si può osservare una violazione della normalità con il test  $F$ .


### 8.2.2 Test di Levene

L'ipotesi nulla è identica a quella del test  $F$ . Il test di Levene ha varie alternative<sup>13</sup>. Come punto di riferimento si può usare la media, la media trimmed oppure la mediana del gruppo. Si consiglia questa ultima alternativa, che risulta essere anche la versione di default in  (`leveneTest()` è nella libreria `car`).

Si calcolano le devianze assolute dalla mediana, separate nei due gruppi. Si procede poi ad un test  $t$  che compara i due gruppi di devianze assolute. Nell'ultima implementazione in  il test  $t$  finale è stato sostituito del perfettamente equivalente test ANOVA ad una via, sempre sulle devianze assolute. Per eseguire il calcolo con la funzione integrata i dati vanno longitudinalizzati. Una colonna per i valori e una colonna per il gruppo di riferimento.

```
head(cd.df)
## misura gruppo
## 1 4.290 A
## 2 3.900 A
## 3 3.783 A
## 4 3.900 A
## 5 4.095 A
## 6 3.120 B
leveneTest(misura ~ gruppo, data=cd.df) # Tipica sintassi dei sistemi lineari, y in funzione di x, "-" = "in funzione di"
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 1      0.2 0.6642
##      10
```

Anche in questo caso non viene respinta l'ipotesi nulla dell'uguaglianza delle varianze.

**Esercizio 8.1.** Calcolare il valore del test di Levene, con l'aiuto di , senza usare la funzione del pacchetto `car`, calcolando le devianze assolute a mano ed eseguendo il relativo test  $t$  sui due gruppi di devianze.

<sup>12</sup>[https://en.wikipedia.org/wiki/F-test\\_of\\_equality\\_of\\_variances](https://en.wikipedia.org/wiki/F-test_of_equality_of_variances)

<sup>13</sup>[https://en.wikipedia.org/wiki/Levene%27s\\_test](https://en.wikipedia.org/wiki/Levene%27s_test)

## 9 Test equivalenti *non parametrici*

Quando i requisiti di un test  $t$  sono violati è buona norma ricorrere a test **non parametrici**; questi ultimi hanno **requisiti più laschi**. Per quasi ogni test classico (parametrico), esiste un test analogo non parametrico. Qui vengono riportati unicamente i due test usati più di frequente. Il Wilcoxon per dati appaiati e il Wilcoxon-Mann-Whitney per dati non appaiati. Consiglio per entrambi una lettura dettagliata dei capitoli sui test non parametrici della monografia del Prof. Soliani (cap 8 e 9)<sup>14</sup>.

Questi test confrontano le mediane, e non le medie dei gruppi! Quindi l'ipotesi nulla è sulla differenza delle mediane  $H_0 : \delta_{mediana} = 0$  e  $H_A : \delta_{mediana} \neq 0$

### 9.1 Il test Wilcoxon per dati appaiati.

Si calcolano le differenze, e successivamente i ranghi delle differenze assolute, e si assegna il segno al rango in base al segno della differenza. Riprendendo i dato sull'isoclorotiazide:

Tabella 6: Tabella dati autoappaiati con il rango dotato di segno

Paziente	Placebo	Idroclorotiazide	Differenza	S.Rango
1	211	181	30	9.0
2	210	172	38	10.0
3	196	210	-14	-4.0
4	203	191	12	2.0
5	196	167	29	7.5
6	190	161	29	7.5
7	191	178	13	3.0
8	160	177	-17	-5.0
9	173	149	24	6.0
10	170	119	51	11.0
11	163	156	7	1.0

Si sommano quindi i ranghi con lo stesso segno, applicando poi il valore assoluto al risultato. Si ottiene la somma dei ranghi positivi pari a 57 e quella dei ranghi negativi pari a 9. Il minore di questi due valori è chiamato  $T$ ; in questo caso  $T = 9$

La somma dei ranghi è valore noto (con deviazioni in presenza di **valori ripetuti** altrimenti chiamati **ties**)

$$S = \frac{n+1}{2} \cdot n \quad (20)$$

Secondo l'ipotesi nulla  $H_0$ , la differenza tra le due serie di osservazioni appaiate dovrebbe essere uguale a zero. Di conseguenza, nella colonna delle differenze la somma dei ranghi con segno positivo e la somma dei ranghi con segno negativo dovrebbero essere uguali. Perciò il totale minore dovrebbe tendere ad un valore medio atteso  $\mu_T$  determinato da  $n$ , il numero di differenze o ranghi, secondo la relazione

$$\mu_T = \frac{n \cdot (n+1)}{4} \quad (21)$$

in questo caso si ha

$$\mu_T = \frac{11 \cdot (11+1)}{4} = 33$$

Nel caso in questione va verificato se la differenza tra il rango calcolato 9 e il rango atteso teorico 33.

Ci sono tabelle di valori critici per  $T$  possibili. In questo caso il valore critico esatto calcolato da  $\mathbb{R}$  è pari a  $T_{crit} = 11$ . Essendo  $T = 9$  inferiore si palesa una differenza significativa. (Per essere significativa il

<sup>14</sup><http://www.dsa.unipr.it/soliani/soliani.html>



valore statistico  $T$  deve essere inferiore a quello critico). Con  $\mathbb{R}$  è facile anche calcolare la probabilità associata con la funzione `psignrank()` e per due code si ottiene  $P = 0.032227$ .

Una tabella di  $T$  critici è riportata nell'appendice (tabella 18).

Tutta questa analisi può essere svolta agevolmente con  $\mathbb{R}$  con un unico comando, come mostrato qui sotto nel code chunk:

```
wilcox.test(test.dip$Iidroclorotiazide,test.dip$Placebo,paired = T,
            exact=F,correct=F,conf.int = T)
##
## Wilcoxon signed rank test
##
## data: test.dip$Iidroclorotiazide and test.dip$Placebo
## V = 9, p-value = 0.03277
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
## -31.999967 -5.000044
## sample estimates:
## (pseudo)median
## -20.28401
```

Inoltre con i confronti tra campioni più grandi, esiste l'approssimazione normale in cui la media attesa è quella già indicata, mentre la deviazione standard  $\sigma_T$  è data da:

$$\sigma_T = \sqrt{\frac{n \cdot (n+1) \cdot (2n+1)}{24}} \quad (22)$$

e si usa la distribuzione di Gauss convertendo  $T$  e  $\mu_T$  con la usuale formula

$$z_{stat} = \frac{T - \mu_T}{\sigma_T} \quad (23)$$

per cui con l'esempio precedente si ha

$$\sigma_T = \sqrt{\frac{11 \cdot 12 \cdot 23}{24}} = 11.274$$

e

$$z_{stat} = \frac{9 - 33}{11.274} = -2.128$$

il cui valore assoluto è maggiore di 1.96. Pertanto la differenza è significativa.

Anche se molti ignorano la cosa anche i test non-parametrici hanno dei requisiti<sup>15</sup>. Per il test di Wilcoxon si assume che la distribuzione delle differenze abbia una forma **simmetrica**. Si può anche dimostrare che ciò è legato al fatto che considerando i due gruppi separatamente, questi devono avere la stessa distribuzione (forma) e differire quindi unicamente per il parametro di location (la mediana). Per verificare questi assunti si può tracciare un **istogramma delle differenze**, ma spesso i dati sono troppo pochi, oppure un **ecdf per ciascun gruppo** (spesso più efficace), oppure anche un boxplot con i singoli punti esplicitati sul grafico.

## 9.2 Il test U di Wilcoxon-Mann-Whitney

Questo test è simile al Wilcoxon signed rank test. Va usato per **campioni non appaiati**. La logica è simile a quella precedente ma usa le **precedenze** anziché i ranghi. Mettendo in ordine i dati, tenendo memoria del gruppo di provenienza, si conta quanti elementi di un gruppo precedono l'altro gruppo. La somma delle precedenze (solitamente quella più piccola) dà il valore  $U$  (in  $\mathbb{R}$  denominato  $W$ ) che va poi confrontato con le tabelle apposite oppure va trasformato con l'approssimazione normale. Usando i dati dell'esempio 7.2 sulla Cladocero Daphnia si ha:

<sup>15</sup>Una interessante discussione su questi requisiti la trovate su stackexchange all'URL <https://stats.stackexchange.com/questions/348057/wilcoxon-signed-rank-symmetry-assumption>

Tabella 7: Precedenze sui dati della Cladocero Daphnia

	11	10	7	6	8	12	3	9	2	4	5	1
misura	3.042	3.081	3.112	3.120	3.120	3.742	3.783	3.847	3.900	3.900	4.095	4.290
gruppo	B	B	B	B	B	B	A	B	A	A	A	A
precA	NA	NA	NA	NA	NA	NA	6	NA	7	7	7	7
precB	0	0	0	0	0	0	NA	1	NA	NA	NA	NA

Chiaramente le precedenze sono per  $B = 1$ , mentre per  $A = 6 + 7 + 7 + 7 + 7 = 34$ . Quindi il valore della statistica è  $U = 1$ . Se l'ipotesi nulla è che non ci siano differenze tra le mediane, si ha il valore teorico:

$$\mu_U = \frac{n_1 \cdot n_2}{2} \quad (24)$$

Per piccoli campioni sono disponibili valori critici di  $U$  calcolati in base alla numerosità dei due campioni (tabelle 19, 20,21). Per i campioni grandi si ricorre all'approssimazione normale, convertendo:


$$z_{stat} = \frac{U - \mu_U}{\sigma_U} \quad (25)$$


e usando come deviazione standard  $\sigma_U$

$$\sigma_U = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}} \quad (26)$$

Qui con  si ottiene il seguente output

```
## [1] 4.290 3.900 3.783 3.900 4.095
## [1] 3.120 3.112 3.120 3.847 3.081 3.042 3.742
##
## Wilcoxon rank sum test
##
## data: cd.a and cd.b
## W = 34, p-value = 0.007166
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  0.1580703 1.0530352
## sample estimates:
## difference in location
##                0.7800098
```

La funzione `pwilcox()` fornisce la probabilità. In questo caso per due code  $2 \cdot \text{pwilcox}(34, 7, 5) = 0.0050505$ . I valori non coincidono in quanto il software applica una correzione a causa della presenza di ties. L'intervallo di confidenza non parametrico è basato su una procedura complessa. Rimando all'help di  per i dettagli.<sup>16</sup>

*Nota* . In  viene usata la  $W$  al posto della  $U$  per il test con gruppi indipendenti. Si tratta dello stesso test; solo la denominazione è alternativa.

Anche per questo test il requisito è che i due gruppi abbiano la stessa distribuzione; pertanto l'`ecdf` è il grafico di elezione per questa verifica.

## 10 Stima della numerosità di un campione

Questa è una parte molto vasta della statistica. Nella maggior parte dei casi è necessario creare simulazioni Monte Carlo. Qui si presentano unicamente due esempi facili senza simulazione.

<sup>16</sup>Nel pacchetto `coin` vi è una funzione alternativa `wilcox_test` con metodi di calcolo Monte Carlo per la determinazione di P-values e intervalli di confidenza; è consigliato soprattutto in presenza di ties.

## 10.1 Con i limiti di confidenza per una ipotesi

Visto che il numero di osservazioni in un campione incide sui limiti di confidenza si vuole con questo calcolo determinare il numero minimo di osservazioni necessarie per ottenere una determinata ampiezza dei limiti di confidenza.

**Esempio 10.1.** Riferendosi sempre all'esempio sui battiti cardiaci tra studenti immaginiamo che quello eseguito sia unicamente uno studio pilota per una ricerca di maggiori dimensioni. Si supponga che si desideri determinare con la confidenza del 95% un intervallo di  $\pm 2$  battiti al minuto entro cui si ipotizza che si trovi la media della popolazione sottostante. Su quanti soggetti si dovrà eseguire la rilevazione? Dalla tabella 16 di Student i limiti di confidenza al 95% di una media con infiniti gradi di libertà è pari a 1.96, quindi i limiti di confidenza saranno dati dalla seguente relazione:

$$\pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} = \pm 2$$

(Si noti come in questo caso è stato scritto  $\sigma$  al posto di  $\hat{s}$ ; in effetti per campioni di dimensioni crescenti questa relazione è lecita).

Quindi risolvendo l'equazione per  $n$  si ottiene che (prendendo per ipotesi il fatto che la deviazione standard corretta del campione possa simulare la popolazione:

$$n = \left(1.96 \cdot \frac{\sigma}{2}\right)^2 = 72.2$$

Quindi come prima ipotesi ci vogliono almeno 73 rilevazioni per avere un intervallo così ristretto di valori di pulsazioni al minuto. Naturalmente questa è solo una stima iniziale in quanto con 72 gradi di libertà il rapporto critico è leggermente diverso da 1.96 e inoltre non è detto che la deviazione standard corretta sia molto aderente a quella della popolazione.

## 10.2 Con ipotesi nulla $H_0$ e ipotesi alternativa $H_A$ (errori $\alpha$ e $\beta$ )

Siano dati un'ipotesi nulla  $H_0$  e la sua media associata  $\mu_0$  e un rischio di tipo  $\alpha$  (significatività al...) riferito all'errore di tipo I e contemporaneamente un'ipotesi alternativa  $H_A$  con la media  $\mu_A$  e il rischio  $\beta$  associato all'errore di tipo II quale sarà la quantità minima di campioni necessari? I rapporti critici per  $\alpha$  e  $\beta$  sono i seguenti:

$$z_\alpha = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$$z_\beta = \frac{\bar{x} - \mu_A}{\frac{\sigma}{\sqrt{n}}}$$

in cui  $z_\beta$  si riferisce alla coda sinistra della distribuzione pari a  $\beta$ . Dalle cui relazioni si ottiene:

$$n = \left[ \frac{(z_\alpha - z_\beta) \cdot \sigma}{\mu_A - \mu_0} \right]^2 \quad (27)$$

**Esempio 10.2.** Come esempio per la determinazione della dimensione del campione si considerino i precedenti dati di pressione sanguigna sistolica nella valutazione di un farmaco antipertensivo. Questi dati illustrano il test  $t$  per campioni appaiati, ma la limitazione alle differenze di ciascun paio ridusse la situazione a quella del test  $t$  per una sola media e precisamente la differenza media delle paia. Si supponga che un nuovo farmaco sia ora a disposizione del ricercatore e si voglia saggiare questo farmaco contro un placebo in maniera simile a quanto fatto con l'idroclorotiazide. Qual è la dimensione del campione di pazienti ipertesi ora necessaria? Il ricercatore specifica che egli desidera eseguire un test a due code, ad un livello di significatività del 5% dell'ipotesi nulla; quest'ultima asserisce che non vi è differenza tra le medie della popolazione. Ciò comporta  $\mu_0 = 0$  e  $z_\alpha = 1.96$  (il valore soglia che lascia un'area del 5% nelle due code della distribuzione normale standardizzata). Il ricercatore specifica inoltre che se la vera differenza fosse pari a 10 mm di mercurio egli desidererebbe correre un rischio solo del 5% di non essere in grado di rifiutare l'ipotesi nulla. Ciò comporta  $\mu_A = 10$  e  $z_\beta = 1.65$  (il valore soglia che lascia un'area del 5% nella coda di sinistra della distribuzione normale standardizzata). Per completare il calcolo il

ricercatore ha bisogno ora di conoscere una stima di  $\sigma$ . I dati relativi al trattamento con idroclorotiazide davano un  $\hat{s} = 13.1$  e non disponendo di altri dati utilizza quindi questa stima:

$$n = \left[ \frac{(1.96 + 1.65) \cdot 20.834}{10 - 0} \right]^2 = 46.749$$

Saranno quindi necessari almeno 47 pazienti. Si noti come questa è per vari motivi un'approssimazione grossolana.

## 11 Inferenza con le proporzioni (distribuzione binomiale)

Quando si valuta quale percentuale di una popolazione  $n$  possiede un determinato livello  $S$  di un certo fattore, si è visto nel capitolo di statistica di base che questa non è altro che la frequenza  $f = \frac{S}{n}$  e che la frequenza non è null'altro che una probabilità  $p$  il cui valore è limitato nell'intervallo  $[0; 1]$ . Per questo motivo la distribuzione di riferimento è quella **binomiale**. La distribuzione binomiale è definita unicamente dalla probabilità  $p$  e dal numero di osservazioni  $n$ . In particolare vale:

$$E(f) = p \quad (28)$$

Cioè il valore atteso della frequenza è pari alla probabilità  $p$ . Inoltre per la varianza vale:

$$V(f) = \frac{p \cdot (1 - p)}{n} \quad (29)$$

**Non c'è un secondo parametro di dispersione** che deve essere definito insieme al parametro di centralità (media), come avviene per le distribuzioni normali. C'è unicamente il parametro di centralità  $p$  che in base alla numerosità del campione  $n$  definisce anche automaticamente la **varianza**  $\sigma^2$ .

Il calcolo delle probabilità può essere svolto in due modi distinti. Si può calcolare la **probabilità esatta** sommando le singole probabilità (la distribuzione binomiale è discreta, non servono integrali ma è una procedura tediosa da fare a mano) oppure si può usare l'approssimazione normale alla distribuzione binomiale. In tal caso la media è  $\mu = n \cdot p$  e la varianza è come indicato sopra. In questo modo si ottiene una standardizzazione per le proporzioni: con una ipotesi nulla  $H_0 : p = \pi$  si ha come indicato qui sotto<sup>17</sup>:

$$z_{stat} = \frac{p - \pi - \frac{1}{2n}}{\sqrt{\frac{\pi \cdot (1 - \pi)}{n}}} \quad (30)$$

oppure per il numero  $x$  di successi:

$$z_{stat} = \frac{x - n \cdot \pi - \frac{1}{2}}{\sqrt{n\pi \cdot (1 - \pi)}} \quad (31)$$

Il fattore  $2n$  è la *correzione di Yates* e serve a correggere la distorsione data dall'utilizzo di una curva continua (normale) per un insieme di dati discreti. Per campioni molto numerosi il fattore di correzione diventa trascurabilmente piccolo e può anche essere omesso dal calcolo.

<sup>17</sup>L'approssimazione della distribuzione normale va usata solamente quando  $n \cdot \pi \geq 5$  e anche per la varianza  $n(1 - \pi) \geq 5$ . Negli altri casi andrebbe calcolata la proporzione esatta.

Tabella 8: Tabella riassuntiva delle varianti di test binomiale

Risultato	Risultato in campione di $n$ unità	Media popola- zione	Errore standard	Coda destra	Coda sinistra
Numero di successi spesso unicamente ipotizzato	$x$	$n\pi$	$\sqrt{n\pi \cdot (1-\pi)}$	$z = \frac{p-\pi-\frac{1}{2}}{\sqrt{n \cdot \pi \cdot (1-\pi)}}$	$z = \frac{p-\pi+\frac{1}{2}}{\sqrt{n \cdot \pi \cdot (1-\pi)}}$
Distribuzione di frequenza dei campioni (più campioni ciascuno di ampiezza $n$ )	$p = \frac{x}{n}$	$\pi$	$\sqrt{\frac{\pi \cdot (1-\pi)}{n}}$	$z = \frac{p-\pi-\frac{1}{2n}}{\sqrt{\frac{\pi \cdot (1-\pi)}{n}}}$	$z = \frac{p-\pi+\frac{1}{2n}}{\sqrt{\frac{\pi \cdot (1-\pi)}{n}}}$
Un campione di $n$ elementi	$100 \cdot p$	$100 \cdot \pi$	$100 \cdot \sqrt{\frac{n\pi \cdot (1-\pi)}{n}}$	$z = \frac{100p-100\pi-\frac{100}{2n}}{\sqrt{\frac{\pi \cdot (1-\pi)}{n}}}$	$z = \frac{100p-100\pi+\frac{100}{2n}}{\sqrt{\frac{\pi \cdot (1-\pi)}{n}}}$


**Esempio 11.1.** In una sperimentazione clinica per confrontare due analgesici denotati A e B, ciascuno di dieci pazienti sofferenti di emicrania cronica, ha ricevuto gli analgesici A e B in due diversi momenti. A caso, metà dei pazienti ha ricevuto prima A e in seguito B; l'altra metà prima B e in seguito A. A ciascun paziente si è chiesto di indicare quale farmaco egli preferisse, cioè quale trovasse più efficace nel diminuire i sintomi della sua emicrania. I pazienti sono stati costretti a scegliere, ad essi non è stato consentito indicare che li preferivano entrambi o che non preferivano nessuno dei due. Un semplice elenco dei risultati ha rivelato che otto pazienti hanno preferito A e due B. Ponendo l'ipotesi nulla  $H_0 : x = n\pi$  con  $\pi = 0.5$  si ha, usando la formula (31):

$$z_c = \frac{x - np - \frac{1}{2}}{\sqrt{n\pi(1-\pi)}} = \frac{8 - 5 - 0.5}{\sqrt{10 \cdot \frac{1}{2} \cdot \frac{1}{2}}} = 1.58$$

Dato che si tratta di un test a due code (non si sa quale dei due farmaci potrebbe essere preferito) il valore è inferiore ad un test con  $P_{stat} > 0.95$  e la proporzione non è significativa. In effetti l'area sottesa alla curva di Gauss corrisponde a 0.114 (11.4%).

**Esempio 11.2.** Il problema precedente può anche essere risolto in modo esatto col calcolo combinatorio, invece di ricorrere all'approssimazione normale. In tal caso si calcolano con la formula per la probabilità delle distribuzioni binomiali, le probabilità degli estremi (0,1,2,8,9,10 preferenze) ottenendo che la somma delle probabilità di questi casi è 0.109. Come esercizio si provi ad eseguire il calcolo per esteso, ricordando che la probabilità per una distribuzione binomiale è:


$$P = C_{n,k} \cdot p^k \cdot (1-p)^{n-k} \quad (32)$$

in cui  $p$  è la probabilità per l'evento favorevole,  $k$  è il numero di volte che l'evento favorevole si realizza e  $n$  è il numero totale di osservazioni. Con  è estremamente semplice eseguire il test in quanto è presente la funzione `binom.test()` come mostrato qui sotto nel code chunk:

```
binom.test(x=8,n=10,p=0.5,alternative="two.sided",conf.level = 0.95)
##
## Exact binomial test
##
## data: 8 and 10
## number of successes = 8, number of trials = 10, p-value = 0.1094
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4439045 0.9747893
## sample estimates:
## probability of success
##                0.8
```

In questo caso anche col calcolo esatto si ottiene il medesimo risultato. Si noti anche che il sistema fornisce gli intervalli di confidenza di  $\pi$ . Tali limiti, per l'approssimazione normale si calcolano con la seguente formula:

$$\pi = p \pm 1.96 \cdot \sqrt{\frac{p \cdot (1 - p)}{n}} \quad (33)$$

**Esercizio 11.1.** Calcolare con  la probabilità esatta dell'esempio precedente, considerando l'ipotesi alternativa  $H_A : \pi \geq 0.5$ , (farmaco B come placebo).  $I = [0.4930987; 1]$ , non signif.

## 11.1 Sensibilità, specificità, & co (reloaded)

Nel paragrafo 4.4 abbiamo trattato concettualmente sensibilità specificità e valori predittivi. Con gli strumenti appena introdotti sulle proporzioni è elementare poter non solo calcolare le proporzioni ma anche i relativi intervalli di confidenza. Esiste nel pacchetto *epyR* la funzione *epi.test()* che è in grado di determinare tutti i parametri di interesse medico con un solo comando.

**Esempio 11.3.** Prendendo l'esempio proposto dal pacchetto *epyR* si ha un certo nuovo test diagnostico provato su 1586 pazienti. Tra 744 pazienti che sono noti per avere un certa malattia, 670 sono risultati positivi. Sugli 842 pazienti non ammalati, 640 sono risultati negativi. Si ha la seguente tabella:

Tabella 9: Dati per un nuovo metodo diagnoastico

	Diagnosi+	Diagnosi-
Test+	670	202
Test-	74	640

Con un solo comando si calcolano tutti i parametri e i relativi intervalli di confidenza, come mostrato nel code chunk qui sotto.

```
epi.tests(dat.scott)
##           Outcome +   Outcome -   Total
## Test +           670         202     872
## Test -            74         640     714
## Total            744         842    1586
##
## Point estimates and 95 % CIs:
## -----
## Apparent prevalence           0.55 (0.52, 0.57)
## True prevalence               0.47 (0.44, 0.49)
## Sensitivity                   0.90 (0.88, 0.92)
## Specificity                   0.76 (0.73, 0.79)
## Positive predictive value     0.77 (0.74, 0.80)
## Negative predictive value     0.90 (0.87, 0.92)
## Positive likelihood ratio     3.75 (3.32, 4.24)
## Negative likelihood ratio     0.13 (0.11, 0.16)
## -----
```

Per le definizioni della prevalenza<sup>18</sup> apparente e/o vera e positive/negative likelihood ratios<sup>19</sup> potete documentarvi in rete, se avete interesse.

Prendendo i dati della tabella 9 si può provare a calcolare a mano per esempio la sensibilità.

$$Sens = \frac{VP}{VP + FN} = \frac{670}{670 + 74} = 0.9005$$

E per gli intervalli di confidenza, usando la formula (33) si ha:

$$p = 0.9005 \pm 1.96 \cdot \sqrt{\frac{0.9005 \cdot (1 - 0.9005)}{670 + 74}} = 0.9005 \pm 0.02151$$

Si può perfino ottenere gli intervalli usando la funzione *binom.test()* ottenendo il seguente output:

<sup>18</sup>[https://influentialpoints.com/Training/estimating\\_true\\_prevalence.htm](https://influentialpoints.com/Training/estimating_true_prevalence.htm)

<sup>19</sup><https://www.statisticshowto.com/likelihood-ratio/>

```
binom.test(670,670+74)
##
## Exact binomial test
##
## data: 670 and 670 + 74
## number of successes = 670, number of trials = 744, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.8767462 0.9210923
## sample estimates:
## probability of success
##           0.9005376
```

Da notare che in questo caso gli intervalli di confidenza non sono simmetrici in quanto la funzione `binom.test()` usa un algoritmo diverso che si adatta meglio ai dati<sup>20</sup>, in particolare quando si trovano vicino alle probabilità 0 oppure 1

## 11.2 Proporzioni indipendenti (test binomiale)

Si consideri per esempio l'analisi di un principio attivo in cui si ha un gruppo di pazienti trattati con il farmaco (T) e un gruppo di controllo (C). Si ha la seguente tabella di grandezze:

Tabella 10: Tabella riassuntiva per proporzioni di trattamento e controllo

Campione	Gruppo trattamento	Gruppo di controllo	Totale
Proporzioni di successi	$\pi_T$	$\pi_C$	
Dimensione del campione	$n_T$	$n_C$	$n_T + n_C$
Numero di successi	$x_T$	$x_C$	$x_T + x_C$
Proporzioni di successi	$p_T = \frac{x_T}{n_T}$	$p_C = \frac{x_C}{n_C}$	$p = \frac{x_T + x_C}{n_T + n_C}$

In questa situazione si deve stimare l'errore standard come media ponderata degli errori standard dei due campioni a confronto. Operando vari passaggi algebrici qui omessi si ottiene la seguente formula che permette il calcolo del rapporto critico considerando l'ipotesi nulla  $H_0 : \pi_T = \pi_C$

$$z = \frac{p_T - p_C}{\sqrt{pq \cdot \left(\frac{1}{n_T} + \frac{1}{n_C}\right)}} \quad (34)$$

Si noti che in questa formula la correzione di Yates non è presa in considerazione. Per poterlo fare si deve usare la formula con le occorrenze come qui indicato:

$$z = \frac{\frac{x_T - \frac{1}{2}}{n_T} - \frac{x_C - \frac{1}{2}}{n_C}}{\sqrt{pq \cdot \left(\frac{1}{n_T} + \frac{1}{n_C}\right)}} \quad (35)$$

Con  $p = \frac{x_T + x_C}{n_T + n_C}$  e  $q = (1 - p)$

**Esempio 11.4.** I dati nella tabella 11 forniscono esempio ideale per il confronto di proporzioni indipendenti. Questi dati si riferiscono ad una sperimentazione clinica del propranololo su pazienti affetti da infarto miocardico acuto. I due gruppi di pazienti sono quelli trattati con propranololo e un gruppo di controllo che non riceve il farmaco. La risposta dicotomica consisteva nell'essere ciascun paziente ancora vivo al ventottesimo giorno dopo la sua ammissione allo studio, o nell'essere egli venuto a mancare in un certo momento compreso entro questo periodo di 28 giorni. Sono riprodotti qui sotto i dati relativi, insieme con i tassi di sopravvivenza, in ciascuno dei due gruppi e nel totale.

Considerando un test a due code con significatività del 5% si ha il seguente calcolo:

<sup>20</sup>Per default è usato l'algoritmo di Clopper e Pearson, si veda l'help del comando per i dettagli.

Tabella 11: Tabella proporzioni pazienti trattati con propanololo

Risultato	Gr. Propanololo	Gr. Controllo	Totale ris.
Sopravvissuti	38	29	67
Non sopravvissuti	7	17	24
Total	45	46	91

$$z = \frac{\frac{38-0.5}{45} - \frac{29-0.5}{49}}{\sqrt{0.7363 \cdot 0.2637 \cdot \left(\frac{1}{45} + \frac{1}{46}\right)}} = 2.08$$

Si può quindi affermare che i pazienti trattati con propanololo hanno una sopravvivenza significativamente migliore di quelli del gruppo di controllo. Si consiglia di dare un'occhiata alla funzione `prop.test()` di R, il cui funzionamento è molto simile a quello della funzione `chisq.test()` presentata nel prossimo paragrafo. Entrambe usano la distribuzione  $\chi^2$  invece di usare l'approssimazione normale. Qui sotto un code chunk con l'esecuzione del test.

```
mat.prop<-matrix(c(Propanololo,Controllo),nrow=2,ncol=2,byrow = F)
prop.test(mat.prop)
##
## 2-sample test for equality of proportions with continuity correction
##
## data:  mat.prop
## X-squared = 4.3198, df = 1, p-value = 0.03767
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.03007647 0.52091855
## sample estimates:
##   prop 1   prop 2
## 0.5671642 0.2916667
```

### 11.3 Proporzioni indipendenti (test $\chi^2$ )

Si chiama anche test dell'indipendenza in certi pacchetti software. Sostanzialmente è un test sulle proporzioni, con il vantaggio che possono esserci più di due gruppi e ci possono anche essere più di due risultati possibili. La formula che si applica è la seguente:

$$\chi_{(n-1)(m-1)}^2 = \frac{\sum (|O - E| - \frac{1}{2})^2}{E} = 4.3195 \quad (36)$$



in cui  $O$  sono i valori osservati e  $E$  i valori attesi calcolati in base ai totali marginali. Per esempio per la prima casella si ha  $E_{1,1} = \frac{45 \cdot 67}{91} = 33.1319$ . Qui sotto la tabella dei valori attesi per l'esempio in questione.

Tabella 12: Valori attesi dei pazienti trattati con propanololo

	Propanololo	Controllo
Sopravvissuti	33.132	33.868
Non sopravvissuti	11.868	12.132

Il risultato di  $\chi_{1g,l}^2 = 4.3195$  risulta significativo, con una probabilità  $P = 0.03768$

*Nota* . Il calcolo dei gradi di libertà è semplicemente  $g \cdot l = (n - 1)(m - 1)$  in cui  $n$  e  $m$  sono il numero di righe e il numero di colonne. Per una tabella  $2 \times 2$  si ha un solo grado di libertà. Il confronto va fatto usando una tabella di valori critici della distribuzione  $\chi_{g,l}^2$ , o in alternativa tramite R. A tal proposito si guardi la funzione `pchisq()`.

**Esercizio 11.2.** Eseguire il calcolo dell'esempio 11.4 qui sopra tramite  usando `chisq.test()` con e senza la correzione di Yates. Va passata ad  la matrice dei conteggi dei casi.



Per evitare di usare l'approssimazione normale, si può ricorrere al test esatto di Fisher<sup>21</sup> che usa le probabilità esatte estratte dalla corrispondente distribuzione multinomiale. Lascio all'interesse dello studente l'approfondimento di questa tecnica. In **R** è disponibile la funzione `fisher.test()`.

## 11.4 Proporzioni dipendenti con il McNemar test

Nel caso in cui ci possa essere un appaiamento dei dati delle proporzioni va usato il test di McNemar al posto dei test precedentemente indicati. Il test con una tabella 2 x 2 di McNemar utilizza unicamente gli elementi fuori dalla diagonale. Con la tabella seguente

Tabella 13: Tabella 2 x 2 per proporzioni dipendenti

	Test 2 positivo	Test 2 negativo	Totale righe
Test 1 positivo	$a$	$b$	$a + b$
Test 1 negativo	$c$	$d$	$c + d$
Totale col	$a + c$	$b + d$	$n = a + b + c + d$

Si considerano unicamente  $c$  e  $d$ . Per il test asintotico che usa la distribuzione  $\chi^2_{1g.l.}$  si ha:

$$\chi^2_{stat} = \frac{(b - c)^2}{b + c} \quad (37)$$

Con i dati sulle risposte all'approvazione del presidente USA con due richieste fatte ad un mese di distanza a 1000 persone si ha un appaiamento dato dal soggetto. La tabella 14 rappresenta i dati:

Tabella 14: Approvazione del presidente USA

	Approvo 2	Disapprovo 2
Approvo 1	794	150
Disapprovo 1	86	570

Per cui il test McNemar con il  $\chi^2$  ha il seguente aspetto:

$$\chi^2_{stat} = \frac{(150 - 86)^2}{150 + 86} = 17.356$$

Il valore di  $\chi^2_{stat}$  è molto maggiore del rapporto critico  $\chi^2_{1g.l.} = 3.8414588$  e quindi nel mese intercorso c'è stata una modifica significativa dell'opinione sull'operato del presidente. La procedura può essere automatizzata la funzione `mcnemar.test()` di **R**, come mostrato nel code chunk qui sotto.

<sup>21</sup>[https://en.wikipedia.org/wiki/Fisher%27s\\_exact\\_test](https://en.wikipedia.org/wiki/Fisher%27s_exact_test)

```
##  
## McNemar's Chi-squared test with continuity correction  
##  
## data: Performance  
## McNemar's chi-squared = 16.818, df = 1, p-value = 4.115e-05
```

L'alternativa esatta è invece quella di applicare un test binomiale con uno dei due conteggi fuori dalla diagonale considerando come totale la loro somma  $n = b + c$  e supponendo come ipotesi nulla  $H_0 : \pi_0 = 0.5$ . In pratica si presume  $b \sim Bin(n = b + c, p = 0.5)$ . Il test esatto con `binom.test()` dà il seguente risultato:

```
##  
## Exact binomial test  
##  
## data: 150 and 236  
## number of successes = 150, number of trials = 236, p-value = 3.716e-05  
## alternative hypothesis: true probability of success is not equal to 0.5  
## 95 percent confidence interval:  
## 0.5706732 0.6970596  
## sample estimates:  
## probability of success  
## 0.6355932
```

Si vede che i risultati sono simili ma non esattamente uguali. Questo fatto non sorprende.

## 12 Tabelle

### 12.1 Integrali della curva di Gauss

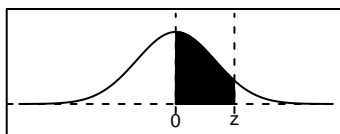


Tabella 15: Tabella degli integrali di Gauss

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41309	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.49900
3.1	0.49903	0.49906	0.49910	0.49913	0.49916	0.49918	0.49921	0.49924	0.49926	0.49929
3.2	0.49931	0.49934	0.49936	0.49938	0.49940	0.49942	0.49944	0.49946	0.49948	0.49950
3.3	0.49952	0.49953	0.49955	0.49957	0.49958	0.49960	0.49961	0.49962	0.49964	0.49965
3.4	0.49966	0.49968	0.49969	0.49970	0.49971	0.49972	0.49973	0.49974	0.49975	0.49976
3.5	0.49977	0.49978	0.49978	0.49979	0.49980	0.49981	0.49981	0.49982	0.49983	0.49983
3.6	0.49984	0.49985	0.49985	0.49986	0.49986	0.49987	0.49987	0.49988	0.49988	0.49989
3.7	0.49989	0.49990	0.49990	0.49990	0.49991	0.49991	0.49992	0.49992	0.49992	0.49992
3.8	0.49993	0.49993	0.49993	0.49994	0.49994	0.49994	0.49994	0.49995	0.49995	0.49995
3.9	0.49995	0.49995	0.49996	0.49996	0.49996	0.49996	0.49996	0.49996	0.49997	0.49997

## 12.2 Valori critici di Student

Tabella 16: Tabella dei valori critici di Student

d.f	X0.90	X0.95	X0.975	X0.99	X0.995
1	3.0777	6.3137	12.7062	31.8205	63.6567
2	1.8856	2.9200	4.3026	6.9646	9.9248
3	1.6377	2.3534	3.1824	4.5407	5.8409
4	1.5332	2.1319	2.7765	3.7470	4.6041
5	1.4759	2.0151	2.5706	3.3649	4.0321
6	1.4398	1.9432	2.4469	3.1427	3.7074
7	1.4149	1.8946	2.3646	2.9979	3.4995
8	1.3968	1.8596	2.3060	2.8965	3.3554
9	1.3830	1.8331	2.2622	2.8214	3.2498
10	1.3722	1.8125	2.2281	2.7638	3.1693
11	1.3634	1.7959	2.2010	2.7181	3.1058
12	1.3562	1.7823	2.1788	2.6810	3.0545
13	1.3502	1.7709	2.1604	2.6503	3.0123
14	1.3450	1.7613	2.1448	2.6245	2.9768
15	1.3406	1.7530	2.1315	2.6025	2.9467
16	1.3368	1.7459	2.1199	2.5835	2.9208
17	1.3334	1.7396	2.1098	2.5669	2.8982
18	1.3304	1.7341	2.1009	2.5524	2.8784
19	1.3277	1.7291	2.0930	2.5395	2.8609
20	1.3253	1.7247	2.0860	2.5280	2.8453
21	1.3232	1.7207	2.0796	2.5177	2.8314
22	1.3212	1.7171	2.0739	2.5083	2.8188
23	1.3195	1.7139	2.0687	2.4999	2.8073
24	1.3178	1.7109	2.0639	2.4922	2.7969
25	1.3163	1.7081	2.0595	2.4851	2.7874
26	1.3150	1.7056	2.0555	2.4786	2.7787
27	1.3137	1.7033	2.0518	2.4727	2.7707
28	1.3125	1.7011	2.0484	2.4671	2.7633
29	1.3114	1.6991	2.0452	2.4620	2.7564
30	1.3104	1.6973	2.0423	2.4573	2.7500
31	1.3095	1.6955	2.0395	2.4528	2.7440
32	1.3086	1.6939	2.0369	2.4487	2.7385
33	1.3077	1.6924	2.0345	2.4448	2.7333
34	1.3070	1.6909	2.0322	2.4411	2.7284
35	1.3062	1.6896	2.0301	2.4377	2.7238
36	1.3055	1.6883	2.0281	2.4345	2.7195
37	1.3049	1.6871	2.0262	2.4314	2.7154
38	1.3042	1.6860	2.0244	2.4286	2.7116
39	1.3036	1.6849	2.0227	2.4258	2.7079
40	1.3031	1.6839	2.0211	2.4233	2.7045
45	1.3007	1.6794	2.0141	2.4121	2.6896
50	1.2987	1.6759	2.0086	2.4033	2.6778
55	1.2971	1.6730	2.0040	2.3961	2.6682
60	1.2958	1.6706	2.0003	2.3901	2.6603
70	1.2938	1.6669	1.9944	2.3808	2.6479
80	1.2922	1.6641	1.9901	2.3739	2.6387
90	1.2910	1.6620	1.9867	2.3685	2.6316
100	1.2901	1.6602	1.9840	2.3642	2.6259
110	1.2893	1.6588	1.9818	2.3607	2.6213
120	1.2887	1.6577	1.9799	2.3578	2.6174
140	1.2876	1.6558	1.9770	2.3533	2.6114
160	1.2869	1.6544	1.9749	2.3499	2.6069
200	1.2858	1.6525	1.9719	2.3451	2.6006
250	1.2850	1.6510	1.9695	2.3414	2.5956
300	1.2844	1.6500	1.9679	2.3388	2.5923
500	1.2833	1.6479	1.9647	2.3338	2.5857

12.3 Valori critici per il  $\chi^2$ Tabella 17: Tabella dei valori critici per  $\chi^2$ 

d.f	X0.90	X0.95	X0.975	X0.99	X0.995
1	2.7055	3.8415	5.0239	6.6349	7.8794
2	4.6052	5.9915	7.3778	9.2103	10.5966
3	6.2514	7.8147	9.3484	11.3449	12.8382
4	7.7794	9.4877	11.1433	13.2767	14.8603
5	9.2364	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5476
7	12.0170	14.0671	16.0128	18.4753	20.2777
8	13.3616	15.5073	17.5345	20.0902	21.9550
9	14.6837	16.9190	19.0228	21.6660	23.5893
10	15.9872	18.3070	20.4832	23.2093	25.1882
11	17.2750	19.6751	21.9200	24.7250	26.7569
12	18.5494	21.0261	23.3367	26.2170	28.2995
13	19.8119	22.3620	24.7356	27.6883	29.8195
14	21.0641	23.6848	26.1190	29.1412	31.3194
15	22.3071	24.9958	27.4884	30.5779	32.8013
16	23.5418	26.2962	28.8453	31.9999	34.2672
17	24.7690	27.5871	30.1910	33.4087	35.7185
18	25.9894	28.8693	31.5264	34.8053	37.1564
19	27.2036	30.1435	32.8523	36.1909	38.5823
20	28.4120	31.4104	34.1696	37.5662	39.9969
21	29.6151	32.6706	35.4789	38.9322	41.4011
22	30.8133	33.9244	36.7807	40.2894	42.7957
23	32.0069	35.1725	38.0756	41.6384	44.1813
24	33.1962	36.4150	39.3641	42.9798	45.5585
25	34.3816	37.6525	40.6465	44.3141	46.9279
26	35.5632	38.8851	41.9232	45.6417	48.2899
27	36.7412	40.1133	43.1945	46.9629	49.6449
28	37.9159	41.3371	44.4608	48.2782	50.9934
29	39.0875	42.5570	45.7223	49.5879	52.3356
30	40.2560	43.7730	46.9792	50.8922	53.6720
31	41.4217	44.9853	48.2319	52.1914	55.0027
32	42.5847	46.1943	49.4804	53.4858	56.3281
33	43.7452	47.3999	50.7251	54.7755	57.6484
34	44.9032	48.6024	51.9660	56.0609	58.9639
35	46.0588	49.8019	53.2034	57.3421	60.2748
36	47.2122	50.9985	54.4373	58.6192	61.5812
37	48.3634	52.1923	55.6680	59.8925	62.8833
38	49.5126	53.3835	56.8955	61.1621	64.1814
39	50.6598	54.5722	58.1201	62.4281	65.4756
40	51.8051	55.7585	59.3417	63.6907	66.7660
45	57.5053	61.6562	65.4102	69.9568	73.1661
50	63.1671	67.5048	71.4202	76.1539	79.4900
55	68.7962	73.3115	77.3805	82.2921	85.7489
60	74.3970	79.0819	83.2977	88.3794	91.9517
70	85.5270	90.5312	95.0232	100.4252	104.2149
80	96.5782	101.8795	106.6286	112.3288	116.3211
90	107.5650	113.1453	118.1359	124.1163	128.2989
100	118.4980	124.3421	129.5612	135.8067	140.1695
110	129.3851	135.4802	140.9166	147.4143	151.9485
120	140.2326	146.5674	152.2114	158.9502	163.6482
140	161.8270	168.6130	174.6478	181.8403	186.8468
160	183.3106	190.5165	196.9151	204.5301	209.8239
200	226.0211	233.9943	241.0579	249.4451	255.2642
250	279.0504	287.8815	295.6886	304.9396	311.3462
300	331.7885	341.3951	349.8745	359.9064	366.8444
500	540.9303	553.1268	563.8515	576.4928	585.2066

## 12.4 Valori critici $T$ di Wilcoxon (paired oppure sign rank test)

Tabella 18: Tabella dei valori critici di Wilcoxon per il signed rank test con uno o due campioni

dim.sample	q.10	q.05	q.025	q.01	q.005	q.0025	q.001	q.0005
4	1	0	0	0	0	0	0	0
5	3	1	0	0	0	0	0	0
6	4	3	1	0	0	0	0	0
7	6	4	3	1	0	0	0	0
8	9	6	4	2	1	0	0	0
9	11	9	6	4	2	1	0	0
10	15	11	9	6	4	2	1	0
11	18	14	11	8	6	4	2	1
12	22	18	14	10	8	6	3	2
13	27	22	18	13	10	8	5	3
14	32	26	22	16	13	10	7	5
15	37	31	26	20	16	13	9	7
16	43	36	30	24	20	16	12	9
17	49	42	35	28	24	20	15	12
18	56	48	41	33	28	24	19	15
19	63	54	47	38	33	28	22	19
20	70	61	53	44	38	33	27	22
21	78	68	59	50	43	38	31	26
22	87	76	66	56	49	43	36	31
23	95	84	74	63	55	49	41	36
24	105	92	82	70	62	55	46	41
25	114	101	90	77	69	61	52	46
26	125	111	99	85	76	68	59	52
27	135	120	108	93	84	75	65	58
28	146	131	117	102	92	83	72	65
29	158	141	127	111	101	91	80	72
30	170	152	138	121	110	99	87	79
31	182	164	148	131	119	108	95	87
32	195	176	160	141	129	117	104	95
33	208	188	171	152	139	127	113	103
34	222	201	183	163	149	137	122	112
35	236	214	196	174	160	147	132	121
36	251	228	209	186	172	158	142	131
37	266	242	222	199	183	169	152	141
38	282	257	236	212	195	181	163	151
39	298	272	250	225	208	193	174	162
40	314	287	265	239	221	205	186	173
45	403	372	344	313	292	273	250	234
50	504	467	435	398	374	351	324	305
55	616	574	537	494	466	440	408	386
60	740	691	649	601	568	538	502	477
65	876	821	773	719	682	648	607	578
70	1023	961	908	847	806	768	722	690
75	1182	1113	1054	987	941	899	848	812
80	1352	1277	1212	1137	1087	1040	984	944
85	1534	1452	1381	1299	1243	1192	1131	1087
90	1728	1639	1561	1472	1411	1356	1288	1241
95	1934	1837	1753	1656	1590	1530	1456	1405
100	2152	2046	1956	1851	1780	1715	1635	1579
120	3140	3002	2882	2744	2651	2565	2460	2386
140	4318	4144	3994	3820	3702	3593	3460	3366
160	5687	5474	5290	5078	4933	4800	4637	4523
180	7247	6993	6774	6520	6347	6188	5994	5857
200	8999	8702	8445	8147	7945	7758	7530	7369
250	14220	13805	13445	13029	12746	12484	12165	11939
300	20647	20101	19629	19081	18709	18364	17943	17646

## 12.5 Valori critici U di Wilcoxon

Tabella 19: Valori critici di U: sopra la diagonale  $p < 0.05$ , sotto la diagonale  $p < 0.025$ 

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3		0	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9
4	1		2	3	4	5	5	6	7	8	9	10	11	12	12	13	14	15
5	2	3		4	6	7	8	9	10	12	13	14	15	16	18	19	20	21
6	3	4	6		7	9	11	12	14	15	17	18	20	22	23	25	26	28
7	3	5	7	9		11	13	15	17	19	21	23	25	27	29	31	33	35
8	4	6	9	11	14		16	18	20	23	25	27	30	32	35	37	39	42
9	4	7	10	13	16	19		21	24	27	29	32	35	38	40	43	46	49
10	5	8	12	15	18	21	25		27	30	34	37	40	43	46	49	53	56
11	6	9	13	17	20	24	28	32		34	38	41	45	48	52	56	59	63
12	6	10	14	18	22	27	31	35	39		42	46	50	54	58	62	66	70
13	7	11	16	20	25	29	34	38	43	48		51	55	60	64	68	73	77
14	8	12	17	22	27	32	37	42	47	52	57		60	65	70	75	79	84
15	8	13	19	24	29	34	40	45	51	56	62	67		71	76	81	86	91
16	9	15	20	26	31	37	43	49	55	61	66	72	78		82	87	93	99
17	10	16	21	27	34	40	46	52	58	65	71	78	84	90		94	100	106
18	10	17	23	29	36	42	49	56	62	69	76	83	89	96	103		107	113
19	11	18	24	31	38	45	52	59	66	73	81	88	95	102	110	117		120
20	12	19	26	33	40	48	55	63	70	78	85	93	101	108	116	124	131	

Tabella 20: Valori critici di U: sopra la diagonale  $p < 0.01$ , sotto la diagonale  $p < 0.005$ 

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3		0	0	0	0	0	1	1	1	2	2	2	3	3	3	3	4	4
4	0		0	1	1	2	2	3	3	4	4	5	6	6	7	7	8	9
5	0	1		2	2	3	4	5	6	7	8	8	9	10	11	12	13	14
6	0	2	3		4	5	6	7	8	10	11	12	13	14	16	17	18	19
7	1	2	4	5		7	8	10	11	13	14	16	17	19	20	22	23	25
8	1	3	5	7	8		10	12	14	16	18	19	21	23	25	27	29	31
9	2	4	6	8	10	12		14	17	19	21	23	25	28	30	32	34	37
10	2	4	7	9	12	14	17		19	22	25	27	30	32	35	38	40	43
11	2	5	8	10	13	16	19	23		25	28	31	34	37	40	43	46	49
12	3	6	9	12	15	18	22	25	29		32	35	38	42	45	48	52	55
13	3	6	10	13	17	21	24	28	32	36		39	43	46	50	54	58	61
14	3	7	11	14	18	23	27	31	35	39	44		47	51	55	59	64	68
15	4	8	12	16	20	25	29	34	38	43	48	52		56	61	65	70	74
16	4	8	13	17	22	27	32	37	42	47	52	57	62		66	71	75	80
17	5	9	14	19	24	29	34	39	45	50	56	61	67	72		76	82	87
18	5	10	15	20	25	31	37	42	48	54	60	66	71	77	83		88	93
19	5	10	16	21	27	33	39	45	51	57	64	70	76	83	89	95		100
20	6	11	17	23	29	35	41	48	54	61	68	74	81	88	94	101	108	

Tabella 21: Valori critici di U per due gruppi con la stessa dimensione n

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$p < 0.05$	0	2	5	8	12	16	22	28	35	43	52	62	73	84	97	110	124	139
$p < 0.025$	0	1	3	6	9	14	18	24	31	38	46	56	65	76	88	100	114	128
$p < 0.01$	0	0	2	4	7	10	15	20	26	32	40	48	57	67	78	89	102	115
$p < 0.005$	0	0	1	3	5	8	12	17	22	28	35	43	52	61	71	82	94	106