

Università di Torino

QUADERNI  
DIDATTICI  
del  
Dipartimento di Matematica

M.GARETTO

**STATISTICA**  
**Lezioni ed esercizi**

Corso di Laurea in Biotecnologie  
A.A. 2002/2003

Quaderno # 13 – Novembre 2002



## 8. Test di ipotesi

### 8.1 Introduzione

Come è già stato messo in evidenza, uno degli scopi più importanti di un'analisi statistica è quello di utilizzare dei dati provenienti da un campione per fare inferenze sulla popolazione da cui è stato tratto il campione.

Nel Cap. 7 si è visto ad esempio come, utilizzando la media campionaria, si può stimare il valore del corrispondente parametro della popolazione.

Ci sono altri problemi in cui invece dobbiamo decidere se un'affermazione riguardante un parametro di una popolazione è vera o falsa.

Ad esempio il responsabile della produzione in un'azienda può ipotizzare che le confezioni prodotte abbiano un peso medio di 250g; un medico può ipotizzare che un certo farmaco sia efficace nel 90% dei casi in cui viene usato. Con la verifica delle ipotesi si può determinare se tali congetture sono compatibili con i dati disponibili dal campione.

#### Definizioni 1

Un'ipotesi formulata in termini di parametri di una popolazione, come media e varianza, è detta **ipotesi statistica**.

Il procedimento che consente di rifiutare o accettare un'ipotesi statistica utilizzando i dati di un campione, viene chiamato **test di ipotesi**.

### 8.2 Ipotesi statistiche

Per illustrare i concetti generali riguardanti la verifica delle ipotesi, consideriamo i seguenti esempi.

#### Esempio 1

Si vuole sottoporre a test l'affermazione di un produttore di vernici secondo cui il tempo medio di asciugatura di una nuova vernice è  $\mu = 20$  minuti.

A questo scopo si prende un campione di 35 lattine di vernice, si effettuano 35 prove di verniciatura con la vernice delle diverse confezioni e si calcola il tempo medio di asciugatura, con l'intenzione di rifiutare l'affermazione del produttore se la media osservata supera il valore di 20 minuti, o di accettarla in caso contrario.

#### Esempio 2

Si vuole verificare se le lattine di caffè confezionate automaticamente da una ditta contengono in media il peso dichiarato  $\mu = 250$  g.

A tale scopo si estrae un campione di 30 lattine, se ne pesa il contenuto e si calcola il peso medio, per stabilire se e quanto differisca da 250g, tenendo presente che il produttore avrebbe un danno sia vendendo lattine con un peso superiore, perché guadagnerebbe meno, sia con un peso inferiore, perché perderebbe i clienti.

La verifica delle ipotesi statistiche inizia con la definizione del problema in termini di ipotesi sul parametro di interesse.

Per prima cosa si stabilisce l'ipotesi da sottoporre a test, detta **ipotesi nulla**, indicata con  $H_0$ , ossia l'ipotesi che si ritiene vera fino a prova contraria.

Oltre all'ipotesi nulla occorre specificare anche un'adeguata **ipotesi alternativa**, indicata con  $H_1$ , ossia un'affermazione che contraddice l'ipotesi nulla. Nell'esempio 1 l'ipotesi nulla è

$$H_0: \quad \mu \leq 20 \text{ minuti}$$

e l'ipotesi alternativa è

$$H_1: \quad \mu > 20 \text{ minuti.}$$

Nell'esempio 2 l'ipotesi nulla è

$$H_0: \mu = 250 \text{ g}$$

e l'ipotesi alternativa è

$$H_1: \mu \neq 250 \text{ g.}$$

Il termine "ipotesi nulla" può essere chiarito dai seguenti esempi.

Se si vuole stabilire se un metodo di insegnamento di una lingua straniera è più efficiente di un altro, si ipotizza che i due metodi siano ugualmente efficienti; se si vuole verificare se un farmaco è più efficace di un altro, si ipotizza che siano ugualmente efficaci.

Questo in altre parole significa ipotizzare che non ci sia nessuna differenza fra i due metodi o fra i due farmaci: per questo motivo l'ipotesi si dice "nulla".

In generale attualmente il termine "ipotesi nulla" viene usato per ogni ipotesi stabilita per decidere se può essere rifiutata o no.

Gli esempi seguenti illustrano la scelta dell'ipotesi nulla e dell'ipotesi alternativa in varie situazioni, nelle quali il parametro sottoposto a test è la media.

### **Esempio 3**

Si supponga di voler dimostrare che il tempo medio richiesto per svolgere una certa operazione è minore di 30 minuti. In tal caso si sceglie l'ipotesi nulla

$$H_0: \mu \geq 30 \text{ minuti}$$

e l'ipotesi alternativa

$$H_1: \mu < 30 \text{ minuti.}$$

### **Esempio 4**

Il contenuto dichiarato dal produttore delle bottiglie di acqua minerale di una certa marca è 920ml.

Un'associazione di consumatori sostiene che in realtà le bottiglie contengono in media una quantità inferiore di acqua.

L'ipotesi nulla è che il produttore non imbrogli, ossia

$$H_0: \mu \geq 920 \text{ ml}$$

e l'ipotesi alternativa è

$$H_1: \mu < 920 \text{ ml.}$$

### **Esempio 5**

Un ingegnere suggerisce alcune modifiche che si potrebbero apportare a una linea produttiva per aumentare il numero di pezzi prodotti giornalmente.

Per decidere se applicare queste modifiche occorre che i dati sperimentali indichino con forte evidenza che la macchina modificata è più produttiva di quella originaria.

Se  $\mu_0$  è il numero medio di pezzi prodotti prima della modifica, si sceglie l'ipotesi nulla

$$H_0: \mu \leq \mu_0$$

e l'ipotesi alternativa è

$$H_1: \mu > \mu_0.$$

### **Osservazione**

E' importante sottolineare che con la verifica delle ipotesi, e in generale con l'inferenza statistica, non si arriva alla dimostrazione di un'ipotesi; si ha solo un'indicazione del fatto che l'ipotesi sia o meno avvalorata dai dati disponibili: quando non si rifiuta un'ipotesi nulla, non si dice che essa è vera, ma che può essere vera; in altre parole se non rifiutiamo l'ipotesi nulla, possiamo solo concludere che il campione non fornisce prove sufficienti a garantirne il rifiuto, ma ciò non implica alcuna dimostrazione.

### 8.3 Tipi di errore e livello di significatività

Dopo aver formulato le ipotesi, occorre specificare quale risultato del campione porterà al rifiuto dell'ipotesi nulla.

Ricordiamo che una statistica campionaria, media o varianza, è uno stimatore corretto del corrispondente parametro della popolazione. Poiché il valore della statistica è calcolato da un campione, anche se l'ipotesi nulla è vera, è però molto probabile che la statistica differisca dal valore vero del parametro di una certa quantità, per effetto del caso; ciò nonostante, se l'ipotesi nulla è vera, ci aspettiamo che la statistica campionaria sia vicina al parametro della popolazione.

Se ciò accade non ci sono prove sufficienti per rifiutare l'ipotesi nulla. Se nell'esempio 1, la media campionaria fosse ad esempio di 20.50 minuti, potremmo ragionevolmente concludere che l'ipotesi nulla è vera, ossia l'affermazione del produttore è vera, perché il valore campionario è "abbastanza vicino" al valore  $\mu = 20$  minuti.

Analogamente, nel caso dell'esempio 2, se la media campionaria fosse di 245 g o di 255 g, potremmo ragionevolmente decidere di accettare l'ipotesi nulla che il peso medio sia  $\mu = 250$  g, perché la differenza dal peso dichiarato è piccola; se invece la differenza dal peso medio fosse "troppo grande" potremmo decidere di rifiutare l'ipotesi.

Il processo decisionale non può certo essere basato sui termini "abbastanza vicino" o "troppo grande" usati negli esempi.

In generale, utilizzando le proprietà della distribuzione di campionamento della statistica soggetta a test, si può identificare un intervallo di valori di quella statistica che verosimilmente non si presentano se l'ipotesi nulla è vera.

La distribuzione di campionamento della statistica test è, di solito, una distribuzione nota, come la normale o la distribuzione  $t$ , e ricorriamo a queste distribuzioni per sottoporre a verifica un'ipotesi nulla. La distribuzione di campionamento della statistica test è divisa in due regioni, una regione di rifiuto e una regione di accettazione, delimitate da uno o più valori, detti **valori critici**.

#### Definizioni 2

La **regione di rifiuto** corrisponde all'insieme dei valori di una statistica campionaria che conducono al rifiuto dell'ipotesi nulla.

L'insieme dei valori che portano invece all'accettazione dell'ipotesi nulla si chiama **regione di accettazione**.

Se la statistica test, in base ai dati del campione, assume un valore che cade nella regione di rifiuto, l'ipotesi nulla deve essere rifiutata; se al contrario il valore cade nella regione di accettazione, l'ipotesi nulla non può essere rifiutata.

La regione di rifiuto può essere vista come l'insieme dei valori della statistica test che non è probabile che si verifichino quando l'ipotesi nulla è vera, mentre è probabile che si verifichino quando l'ipotesi nulla è falsa. Pertanto, se il campione porta a un valore della statistica test che cade nella regione di rifiuto, rifiutiamo l'ipotesi nulla perché non è probabile che sia vera.

Quando si usa una statistica campionaria per prendere una decisione sul parametro della popolazione si corre sempre il rischio di giungere a una conclusione sbagliata. In effetti nella verifica di ipotesi si individuano due **tipi di errore**.

Per illustrare questo problema riprendiamo in esame l'esempio 1. Supponiamo di aver scelto la regione di accettazione, stabilendo di accettare l'ipotesi nulla se la media del campione non supera i 20.50 minuti.

C'è una prima possibilità che la media del campione superi i 20.50 minuti stabiliti, mentre la media effettiva della popolazione è  $\mu = 20$  minuti; c'è anche una seconda possibilità che la media del campione possa essere minore o uguale a 20.50 minuti, ma la media effettiva non sia  $\mu = 20$  minuti, ma sia ad esempio  $\mu = 21$  minuti.

La situazione appena descritta in questo esempio è tipica dei test di ipotesi: anche se si fa il test in modo corretto, si possono commettere questi due tipi di errore, che possono portare a conseguenze dannose.

**Definizioni 3**

Se l'ipotesi  $H_0$  è vera, ma viene erroneamente rifiutata, si commette un **errore del I tipo**; la probabilità di commettere tale errore è indicata con  $\alpha$ .

Se l'ipotesi  $H_0$  è falsa, ma erroneamente viene accettata, si commette un **errore del II tipo**; la probabilità di commettere questo tipo di errore è indicata con  $\beta$ .

Questi risultati possono essere riassunti nel seguente schema.

	$H_0$ vera	$H_0$ falsa
Rifiutiamo $H_0$	Errore del I tipo Probabilità = $\alpha$	Decisione corretta
Accettiamo $H_0$	Decisione corretta	Errore del II tipo Probabilità = $\beta$

Un'analogia che può chiarire le idee precedenti è quella del processo a un imputato. In tribunale una persona sottoposta a processo viene ritenuta innocente fino a prova contraria. L'ipotesi nulla  $H_0$  è quindi "l'imputato è innocente"; l'ipotesi alternativa  $H_1$  è "l'imputato è colpevole".

L'errore del I tipo è condannare un innocente, l'errore del II tipo è assolvere un colpevole.

Riassumiamo questi concetti con lo schema seguente.

	Imputato innocente	Imputato colpevole
Imputato condannato	Errore del I tipo	Decisione corretta
Imputato assolto	Decisione corretta	Errore del II tipo

Scegliere come ipotesi nulla  $H_0$  "l'imputato è innocente" significa ritenere che condannare un innocente sia un errore più grave che assolvere un colpevole.

In generale l'errore di I tipo è quello considerato più grave: questo significa che l'ipotesi nulla  $H_0$  va formulata in modo che quello che si ritiene sia l'errore più grave coincida con l'errore di I tipo.

Servendoci ancora degli esempi 1 e 2, calcoliamo la probabilità  $\alpha$  di commettere un errore del I tipo; usiamo a tale scopo le proprietà della distribuzione della media campionaria.

**Esempio 1 – parte 2**

Assumiamo che sia noto dall'esperienza che lo scarto quadratico medio del tempo di asciugatura della vernice è  $\sigma = 2$  minuti e studiamo la probabilità di commettere un errore del I tipo, ossia la probabilità  $\alpha$  che la media del campione superi 20.5 minuti, anche se la media effettiva della popolazione è  $\mu = 20$  minuti<sup>1</sup>.

Come è noto dal Cap. 6, la distribuzione della media campionaria per grandi campioni ( $n \geq 30$ ) è approssimativamente normale, quindi la probabilità suddetta è data dall'area della regione rappresentata nella figura 1

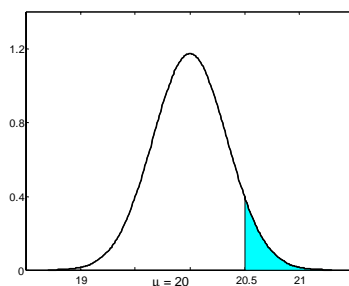


Figura 1

<sup>1</sup> Si ricordi che l'ipotesi nulla e l'ipotesi alternativa in questo esempio sono

$$\begin{aligned} H_0: & \quad \mu = 20 \text{ minuti} \\ H_1: & \quad \mu < 20 \text{ minuti.} \end{aligned}$$

La regione a destra del valore 20.5 è la regione di rifiuto, quella a sinistra è la regione di accettazione: se il valore della media campionaria cade a destra di 20.5 l'ipotesi nulla viene rifiutata, altrimenti non viene rifiutata.

Se la popolazione da cui proviene il campione è sufficientemente grande da poterla considerare infinita<sup>2</sup>, applicando il teorema 1, Cap. 6, pag. 171, si calcola la deviazione standard della distribuzione della media campionaria

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{35}} = 0.34$$

Standardizzando il valore  $\bar{x} = 20.5$  si ha

$$Z = \frac{20.5 - 20}{0.34} = 1.47.$$

Utilizzando le tavole della distribuzione normale, si trova che l'area della regione a destra di 20.5 è  $P(Z > 1.47) = 1 - P(Z < 1.47) = 1 - 0.9292 = 0.0708$

quindi la probabilità di rifiutare erroneamente l'ipotesi nulla è  $\alpha = 0.0708$

### Esempio 2 – parte 2

Assumiamo che lo scarto quadratico medio della popolazione sia  $\sigma = 15\text{g}$  e studiamo la probabilità  $\alpha$  che la media del campione non sia compresa fra 245g e 255g, anche se la media effettiva della popolazione è  $\mu = 250\text{g}$ <sup>3</sup>.

La probabilità che si vuole calcolare è data dalla somma delle due aree rappresentate nella figura 2

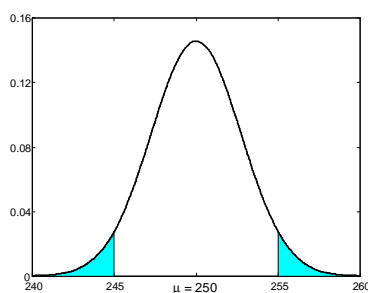


Figura 2

La regione di rifiuto in questo caso è costituita dai valori a sinistra di 245g e dai valori a destra di 255g; se il valore della media campionaria cade nell'intervallo (245, 255), che è la regione di accettazione, l'ipotesi nulla viene accettata, altrimenti viene rifiutata.

Seguendo il procedimento già descritto nell'esempio precedente si trova

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{30}} \cong 2.74.$$

La regione di accettazione è un intervallo simmetrico rispetto a  $\mu = 250$ ; standardizzando il valore  $\bar{x} = 255$  si ha

$$Z = \frac{255 - 250}{2.74} = 1.82.$$

Utilizzando le tavole della distribuzione normale, si trova che l'area della regione colorata è

$$P(|Z| > 1.82) = 2 \cdot [1 - P(Z < 1.82)] = 2 \cdot (1 - 0.9656) = 0.0688$$

quindi la probabilità di rifiutare erroneamente l'ipotesi nulla è  $\alpha = 0.0688$ .

<sup>2</sup> Si ricordi quanto detto nel Cap. 6, pag. 171, a proposito della correzione per popolazioni finite e si veda lo schema riassuntivo a pag. 172, punto 2 b.

<sup>3</sup> Si ricordi che l'ipotesi nulla e l'ipotesi alternativa in questo esempio sono

$$\begin{aligned} H_0: & \quad \mu = 250 \text{ g} \\ H_1: & \quad \mu \neq 250 \text{ g}. \end{aligned}$$

Quando la regione di rifiuto è costituita da un intervallo (come accade nell'esempio 1, parte 2), il **test** si dice **unilaterale** o anche **test a una coda**; quando invece la regione di rifiuto è costituita da due intervalli, ossia da due code della distribuzione (come accade nell'esempio 2, parte 2), il **test** si dice **bilaterale** o anche **test a due code**.

I valori che dividono la regione di accettazione da quella di rifiuto sono detti **valori critici**; tali valori sono uno o due, a seconda che il test sia a una coda o a due code.

#### **Definizione 4**

La probabilità  $\alpha$  di commettere un errore del I tipo, ossia di rifiutare un'ipotesi nulla vera, è detta **livello di significatività**.

Negli esempi 1 e 2 (parte 2) si è mostrato come calcolare la probabilità  $\alpha$  di commettere un errore del I tipo, per regioni di rifiuto scelte arbitrariamente.

Il metodo usato più frequentemente nelle applicazioni consiste invece nello specificare un valore per il livello di significatività  $\alpha$  e poi identificare la regione di rifiuto che soddisfa tale valore.

Poiché l'errore di I tipo è quello considerato più grave, si scelgono per  $\alpha$  valori piccoli; i valori più usati sono  $\alpha = 0.01$  e  $\alpha = 0.05$ .

In corrispondenza al livello di significatività  $\alpha$ , il valore  $(1 - \alpha) \cdot 100\%$  coincide con il **grado di fiducia** già introdotto a proposito degli intervalli di confidenza.

Se si sceglie ad esempio un livello di significatività  $\alpha = 0.05$ , ossia del 5%, ci sarà una probabilità del 5% di rifiutare un'ipotesi che avrebbe dovuto essere accettata; in altre parole siamo fiduciosi al 95% di aver preso la decisione giusta.

#### **Definizione 5**

La probabilità di commettere un errore del II tipo, indicata con  $\beta$ , viene anche chiamata **rischio del consumatore**.

Si può controllare il rischio connesso a un errore del I tipo scegliendo un valore di  $\alpha$  piccolo, ad esempio  $\alpha = 0.01$ : questo deve essere fatto se si ritiene che le conseguenze di un errore del I tipo siano gravi. Tuttavia al diminuire di  $\alpha$ , aumenta  $\beta$ , ossia ad una riduzione dell'errore del I tipo si accompagna un aumento dell'errore del II tipo. Quindi nei casi in cui è molto importante evitare, per quanto possibile, un errore del II tipo, è meglio scegliere come valore di  $\alpha$  un valore non troppo piccolo, ad esempio  $\alpha = 0.05$ .

Un modo per controllare e ridurre l'errore del II tipo consiste nell'aumentare la dimensione del campione. Un'elevata dimensione del campione consente di solito di individuare anche piccole differenze tra la statistica campionaria e il parametro della popolazione. Si tenga presente però che aumentare di molto l'ampiezza del campione potrebbe essere troppo costoso.

Per un fissato valore di  $\alpha$  l'aumento dell'ampiezza del campione riduce il rischio del consumatore  $\beta$ , quindi aumenta la probabilità  $1 - \beta$  di rifiutare l'ipotesi nulla quando è falsa, e quindi dovrebbe essere rifiutata. La probabilità  $1 - \beta$  si chiama anche **potenza del test**.

La scelta dei valori di  $\alpha$  e  $\beta$  dipende dai costi che ciascun errore comporta (vedere esempio 8).

Riassumiamo nello schema seguente i passi in cui si articola un test di ipotesi.

#### **Schema riassuntivo – Test di ipotesi**

- 1 – Si scelgono l'ipotesi nulla e l'ipotesi alternativa.
- 2 – Si sceglie il livello di significatività  $\alpha$  a cui si vuole eseguire il test.
- 3 – In funzione del valore  $\alpha$  scelto, si determina la regione di rifiuto.
- 4 – Si calcola dai dati del campione il valore della statistica test e si vede se appartiene o no alla regione di rifiuto.
- 5 – Si prende la decisione: rifiutare o non rifiutare l'ipotesi nulla al livello di significatività stabilito.

E' opportuno sottolineare che, quando l'ipotesi nulla non è rifiutata, non si dovrebbe dire che tale ipotesi viene accettata, bensì che l'ipotesi nulla non viene rifiutata: questo perché è possibile che di commetta un errore del II tipo; poiché spesso la probabilità di commettere un errore del II tipo è abbastanza elevata, non ci si dovrebbe impegnare troppo dicendo che si accetta l'ipotesi nulla. Tuttavia, anche se impropriamente, spesso si usa il termine "si accetta l'ipotesi nulla".

#### 8.4 Test di ipotesi sulla media (varianza nota)

Descriviamo il procedimento per eseguire un test di ipotesi sulla media di una popolazione avente varianza  $\sigma^2$  nota.

Il test si basa sulla statistica

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

dove  $n$  è l'ampiezza del campione e  $\mu_0$  è il valore della media assunto nell'ipotesi nulla

$$H_0: \mu = \mu_0.$$

Il test qui illustrato è essenzialmente un **test per grandi campioni** ( $n \geq 30$ ); in tal caso la distribuzione della media campionaria può essere approssimata dalla distribuzione normale e la variabile aleatoria  $Z$  ha approssimativamente la distribuzione normale standardizzata.

Nel caso particolare in cui il campione è estratto da una popolazione con distribuzione normale, la variabile  $Z$  ha distribuzione normale standardizzata, qualunque sia l'ampiezza del campione (vedere esempi 13 e 14).

Sia, come al solito,  $z_\alpha$  il valore di  $Z$  per cui l'area a destra di  $z_\alpha$  al di sotto della curva normale standardizzata è uguale a  $\alpha$ .

Nelle figure seguenti si illustrano le regioni di rifiuto per un dato livello di significatività  $\alpha$ , a seconda dell'ipotesi alternativa stabilita.

L'ipotesi nulla è

$$H_0: \mu = \mu_0.$$

Nei primi due casi si fa un test a una coda, nel terzo caso un test a due code.

##### 1° caso – Test a una coda (figura 3)

Ipotesi alternativa  $H_1: \mu > \mu_0.$

Regione di rifiuto<sup>4</sup>  $Z > z_\alpha$

Regione di accettazione  $Z < z_\alpha$

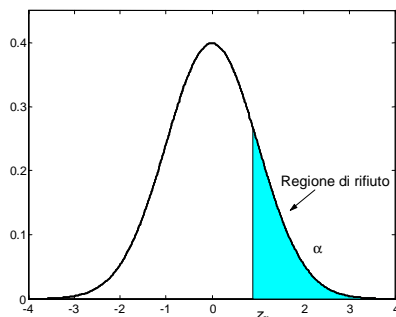


Figura 3

<sup>4</sup> L'utilizzo o meno del segno di uguale nelle regioni di rifiuto e di accettazione, in questo e nei casi seguenti, è assolutamente ininfluente, dal momento che la distribuzione normale è una distribuzione continua.

**Esempio 10**

Un campione di 36 osservazioni avente media  $\bar{x} = 86.2$  proviene da una distribuzione avente varianza  $\sigma^2 = 100$ . In passato la media della distribuzione era  $\mu = 83.0$ , ma si ipotizza che recentemente la media possa essere cambiata.

Usando il livello di significatività del 5%, sottoporre a test l'ipotesi nulla

$$H_0: \mu = 83.0$$

a – supporre di non sapere, nel caso che la media sia cambiata, se è aumentata o diminuita;

b – supporre di sapere che, nel caso che la media sia cambiata, essa può solo essere aumentata.

a – Nel primo caso l'ipotesi alternativa è

$$H_1: \mu \neq 83.0$$

Si effettua un test a due code; per il livello di significatività  $\alpha = 0.05$  la regione di rifiuto è costituita dai valori  $Z < -1.96$  e  $Z > 1.96$ .

Il valore della statistica test è

$$Z = \frac{86.2 - 83.0}{\frac{10}{\sqrt{36}}} = 1.92.$$

Il valore  $Z = 1.92$  non appartiene alla regione di rifiuto, perciò si decide di non rifiutare l'ipotesi nulla. In altre parole non c'è un'evidenza significativa, al livello del 5%, che la media sia cambiata.

b – Nel secondo caso l'ipotesi alternativa è

$$H_1: \mu > 83.0.$$

Si effettua un test a una coda; per il livello di significatività  $\alpha = 0.05$  il valore critico è  $z_\alpha = 1.645$ ; la regione di rifiuto è costituita dai valori  $Z > 1.645$ .

Il valore  $Z = 1.92$  appartiene alla regione di rifiuto, perciò si decide di rifiutare l'ipotesi nulla. In altre parole si ha un'evidenza significativa, al livello del 5%, che la media è aumentata.

Si noti che le decisioni prese sono diverse nei due casi, e ciò dipende dal fatto che la stessa ipotesi nulla viene testata contro alternative diverse.

Il test descritto in questo paragrafo richiede che sia noto il valore  $\sigma$  dello scarto quadratico medio; se  $\sigma$  non è conosciuto, ma il campione è grande, si può sostituire  $\sigma$  con il valore  $s$  dello scarto quadratico medio del campione.

**Esempio 11**

Una ditta produttrice di pneumatici afferma che la durata media di un certo tipo di pneumatici per auto è di almeno 50000km.

Per sottoporre a test questa affermazione un campione di 40 pneumatici viene sottoposto a prove su strada e si misura una durata media  $\bar{x} = 48900$  km, con uno scarto quadratico medio  $s = 2500$  km.

Sottoporre a test l'affermazione, con un livello di significatività  $\alpha = 0.01$ .

L'ipotesi nulla è

$$H_0: \mu \geq 50000$$

e l'ipotesi alternativa è

$$H_1: \mu < 50000.$$

Si effettua un test ad una coda e, dato che la probabilità di un errore del I tipo è massima quando  $\mu = 50000$ , si procede come se l'ipotesi nulla fosse

$$H_0: \mu = 50000$$

Per il livello di significatività  $\alpha = 0.01$  il valore critico è  $z_\alpha = -2.326$  e la regione di rifiuto è costituita dai valori  $Z < -2.326$ .

Lo scarto quadratico medio della popolazione non è noto e viene sostituito con lo scarto quadratico medio del campione.

Il valore della statistica test è

$$Z = \frac{48900 - 50000}{\frac{2500}{\sqrt{40}}} = -2.78.$$

Il valore  $Z = -2.78$  appartiene alla regione di rifiuto, perciò l'ipotesi nulla deve essere rifiutata al livello di significatività  $\alpha = 0.01$ , e concludiamo che l'affermazione del produttore non può essere accettata.

### **Esempio 12**

In un dato anno il voto medio all'esame di maturità classica è stato di 73/100.

In una commissione che ha esaminato 70 candidati, si è registrato un voto medio di 76.2/100 con uno scarto quadratico medio  $s = 14$ .

Verificare l'ipotesi che non ci sia differenza significativa tra la media generale e la media del campione, al livello di significatività del 5%.

L'ipotesi nulla è

$$H_0: \quad \mu = 73$$

e l'ipotesi alternativa è

$$H_1: \quad \mu \neq 73.$$

Il test è a due code e al livello di significatività del 5% la regione di rifiuto è costituita dai valori  $Z < -1.96$  e  $Z > 1.96$ .

Lo scarto quadratico medio della popolazione non è noto e viene sostituito con lo scarto quadratico medio del campione  $s = 14$ .

Il valore della statistica test è

$$Z = \frac{76.2 - 73}{\frac{14}{\sqrt{70}}} = 1.91.$$

Il valore  $Z = 1.91$  non appartiene alla regione di rifiuto, perciò al livello di significatività del 5% l'ipotesi nulla non deve essere rifiutata; concludiamo quindi che la differenza tra il risultato generale e il risultato della particolare commissione è dovuta a fluttuazioni casuali, ossia con una probabilità del 95% la differenza non è imputabile né ai candidati, né alla commissione d'esame.

E' evidente che nella scelta delle ipotesi non si vuole indagare su una maggiore o minore severità della commissione.

Se la popolazione da cui proviene il campione è normale, questo test può essere applicato anche nel caso di piccoli campioni.

### **Esempio 13**

Supponiamo che i punteggi di un test sul quoziente di intelligenza di una certa popolazione di adulti si distribuiscano normalmente con uno scarto quadratico medio  $\sigma = 15$ .

Un campione di 25 adulti estratti da questa popolazione ha un punteggio medio di 105.

Sottoporre a test l'ipotesi che il punteggio medio sia 100, con un livello di significatività del 5%.

Poiché la popolazione da cui proviene il campione ha distribuzione normale con scarto quadratico medio noto  $\sigma = 15$ , quanto detto per i grandi campioni è valido anche per un piccolo campione.

L'ipotesi nulla è

$$H_0: \quad \mu = 100$$

e l'ipotesi alternativa è

$$H_1: \quad \mu \neq 100.$$

Si effettua un test a due code; per il livello di significatività  $\alpha = 0.05$  la regione di rifiuto è costituita dai valori  $Z < -1.96$  e  $Z > 1.96$ .

Il valore della statistica test è

$$Z = \frac{105 - 100}{\frac{15}{\sqrt{25}}} = 1.67.$$

Il valore  $Z = 1.67$  non appartiene alla regione di rifiuto, perciò si decide di non rifiutare l'ipotesi nulla.

#### **Esempio 14**

Da una popolazione normale avente scarto quadratico medio  $\sigma = 2$ , si estrae un campione di ampiezza  $n = 10$ . Il valor medio del campione sia  $\bar{x} = 18.58$ .

Sottoporre a test l'ipotesi nulla

$$H_0: \mu = 20$$

scegliendo come ipotesi alternativa

$$H_1: \mu \neq 20$$

ai livelli di significatività dell'1% e del 5%.

Poiché la popolazione da cui proviene il campione è normale, si può effettuare il test per grandi campioni anche se l'ampiezza del campione è  $n = 10$ .

a – Livello di significatività  $\alpha = 0.01$ .

Si effettua un test a due code; per il livello di significatività  $\alpha = 0.01$  la regione di rifiuto è costituita dai valori  $Z < -2.576$  e  $Z > 2.576$ .

Il valore della statistica test è

$$Z = \frac{18.58 - 20}{\frac{2}{\sqrt{10}}} = -2.245.$$

Il valore  $Z = -2.245$  non appartiene alla regione di rifiuto, perciò si decide di non rifiutare l'ipotesi nulla.

b – Livello di significatività  $\alpha = 0.05$ .

Si effettua un test a due code; per il livello di significatività  $\alpha = 0.05$  la regione di rifiuto è costituita dai valori  $Z < -1.96$  e  $Z > 1.96$ .

Il valore  $Z = -2.245$  appartiene alla regione di rifiuto, perciò si decide di rifiutare l'ipotesi nulla.

Nel caso trattato in questo esempio si possono dunque trarre le seguenti conclusioni:

a – I dati campionari non consentono di rifiutare l'ipotesi nulla al livello di significatività dell'1%.

b – I dati campionari consentono di rifiutare l'ipotesi nulla al livello di significatività del 5%.

Come si vede, la decisione che si prende non dipende solo dai dati campionari, ma anche dal livello di significatività fissato. In questo caso, la differenza fra la media del campione  $\bar{x} = 18.58$  e il valore ipotizzato  $\mu = 20$  per il parametro della popolazione viene ritenuta statisticamente significativa al livello del 5%, ma non al livello dell'1%.

Queste conclusioni ci portano alle seguenti considerazioni.

Ogni test di ipotesi porta al confronto di due numeri, il valore della statistica  $Z$ , che può essere calcolato in base ai dati campionari, e il valore critico (o i due valori critici nel test a due code), che invece dipende dal livello di significatività fissato. Nell'esempio sono stati confrontati il valore della statistica  $Z = -2.245$  e i valori critici  $z_{\frac{\alpha}{2}} = -1.96$  e  $z_{\frac{\alpha}{2}} = 1.96$ . Se tra i valori suddetti vale

una certa disuguaglianza, si rifiuta l'ipotesi, altrimenti non si rifiuta.

Poiché, come abbiamo visto nell'esempio precedente, un livello  $\alpha$  diverso può condurre a una decisione diversa (rifiutare/non rifiutare) risulta interessante determinare qual è il valore  $\alpha$  che fa da spartiacque fra le due diverse conclusioni.

Nell'esempio precedente ci poniamo la seguente domanda: fissati i dati del campione, e quindi il valore di  $Z$ , qual è il più piccolo livello di significatività  $\alpha$  per cui si rifiuta l'ipotesi nulla?

Nel caso dell'esempio la regione di rifiuto è costituita dai valori  $Z$  tali che

$$|Z| > z_{\frac{\alpha}{2}}$$

Il più piccolo valore di  $\alpha$  per cui si rifiuta l'ipotesi si trova risolvendo l'equazione

$$2.245 = z_{\frac{\alpha}{2}}$$

Da qui segue<sup>1</sup>

$$\frac{\alpha}{2} = P(Z > 2.245) = 1 - P(Z < 2.245) = 1 - 0.9876 = 0.0124$$

$$\alpha = 0.0248$$

Questo significa che, con i dati campionari disponibili, il livello di significatività che fa da spartiacque tra la decisione di rifiutare l'ipotesi nulla e quella di non rifiutarla è il livello del 2.48%: questo livello è quindi il più piccolo livello a cui i dati disponibili permettono di rifiutare l'ipotesi nulla.

#### Definizione 6

In un test di ipotesi, dopo aver effettuato il campionamento e calcolato il valore della statistica necessaria per eseguire il test, si dice **P-value** il più piccolo valore del livello di significatività a cui i dati campionari consentono di rifiutare l'ipotesi nulla.

Un  $P$ -value quasi uguale a zero significa che siamo praticamente certi di non sbagliare rifiutando l'ipotesi nulla; un  $P$ -value dell'ordine dei soliti livelli di significatività è imbarazzante, in quanto indica che la decisione se rifiutare o no l'ipotesi nulla è critica, dipende in modo cruciale dalla scelta del livello di significatività; un  $P$ -value molto maggiore indica invece che a qualsiasi livello ragionevole di significatività, non rifiutiamo l'ipotesi nulla; in questo caso si può anche dire che il test ci porta ad accettare l'ipotesi.

Il  $P$ -value può essere difficile da calcolare con precisione usando le tavole, ma viene di solito fornito dai più diffusi software statistici.

Per i test basati sulla distribuzione normale, come nel caso dell'esempio 12, il  $P$ -value è relativamente facile da calcolare. Se  $Z_0$  è il valore della statistica test, calcolato in base ai dati campionari, allora il  $P$ -value può essere ottenuto in base alle seguenti formule

$P\text{-value} = \begin{cases} 1 - P(Z < Z_0) \\ P(Z < Z_0) \\ 2[1 - P(Z <  Z_0 )] \end{cases}$	per il test a una coda con	$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$
	per il test a una coda con	$H_0 : \mu = \mu_0$	$H_1 : \mu < \mu_0$
	per il test a due code con	$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$

#### Esempio 15

Riprendiamo in esame i risultati ottenuti nell'esempio 8.

Le conclusioni tratte nel caso a sono piuttosto critiche e questo viene evidenziato dal  $P$ -value; si ha infatti

$$Z_0 = -2.085$$

$$P\text{-value} = 2[1 - P(Z < 2.085)] = 2(1 - 0.9815) = 0.0371$$

Il livello minimo che consente di rifiutare l'ipotesi nulla è del 3.71%.

Nel caso b invece le conclusioni non sono critiche; si ha infatti

$$Z_0 = -1.318$$

$$P\text{-value} = 2[1 - P(Z < 1.318)] = 2(1 - 0.9062) = 0.1875$$

In questo caso a ogni ragionevole livello di significatività possiamo accettare l'ipotesi nulla<sup>2</sup>.

<sup>1</sup> Il valore della probabilità  $P(Z < 2.245)$  è stato calcolato con un software statistico.

<sup>2</sup> In entrambi i casi i calcoli delle probabilità sono stati effettuati con un software statistico.

### 8.5 Test di ipotesi sulla media (varianza incognita)

Esaminiamo ora il caso in cui il campione usato per effettuare il test proviene da una popolazione di cui non è nota la varianza  $\sigma^2$ .

Come già osservato nel paragrafo precedente, se  $\sigma$  non è noto, ma il campione è grande, si può sostituire  $\sigma$  con il valore  $s$  dello scarto quadratico medio del campione.

Se invece il campione è piccolo, e la popolazione da cui proviene il campione ha distribuzione normale, o almeno approssimativamente normale, si può usare il teorema 3, Cap. 6; sulla base di tale teorema la statistica

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

è una variabile aleatoria avente la distribuzione  $t$  con grado di libertà  $v = n - 1$ .

I criteri per i test a una e a due code basati sull'uso di questa distribuzione sono analoghi a quelli già descritti nel paragrafo precedente, con  $z_\alpha$  e  $\frac{z_\alpha}{2}$  sostituiti da  $t_\alpha$  e  $\frac{t_\alpha}{2}$ ; questi valori critici per

un dato livello di significatività  $\alpha$  dipendono dal grado di libertà e devono essere letti di volta in volta sulle tavole della distribuzione  $t$ .

Nella tabella 2 riassumiamo i valori comunemente usati per il livello di significatività  $\alpha$  e i corrispondenti valori critici  $t_\alpha$  e  $\frac{t_\alpha}{2}$  per i test a una e a due code. L'ipotesi nulla è

$$H_0: \mu = \mu_0.$$

Test	Ipot. altern. $H_1$	Liv. signif. $\alpha$	Valori critici	Reg. rifiuto
una coda	$\mu > \mu_0$	0.01	$t_\alpha = t_{0.01}$	$T > t_{0.01}$
		0.05	$t_\alpha = t_{0.05}$	$T > t_{0.05}$
una coda	$\mu < \mu_0$	0.01	$t_\alpha = -t_{0.01}$	$T < -t_{0.01}$
		0.05	$t_\alpha = -t_{0.05}$	$T < -t_{0.05}$
due code	$\mu \neq \mu_0$	0.01	$\frac{t_\alpha}{2} = t_{0.005}$ $\frac{t_\alpha}{2} = -t_{0.005}$	$T > t_{0.005}$ $T < -t_{0.005}$
		0.05	$\frac{t_\alpha}{2} = t_{0.025}$ $\frac{t_\alpha}{2} = -t_{0.025}$	$T > t_{0.025}$ $T < -t_{0.025}$

Tabella 2

#### Esempio 16

Le bottiglie di vino poste in vendita contengono usualmente 750ml di vino.

Si effettua un controllo su un campione di 6 bottiglie e si misurano i seguenti valori in ml

747.0	751.5	752.0	747.5	747.0	749.0
-------	-------	-------	-------	-------	-------

Stabilire se questi dati confermano con un livello di significatività del 5% l'affermazione che le bottiglie hanno un contenuto in media pari a quanto dichiarato.

Se il test è effettuato per tutelare l'interesse del consumatore, l'ipotesi nulla è

$$H_0: \mu = 750$$

e l'ipotesi alternativa è

$$H_1: \mu < 750.$$

Calcolando la media e la varianza del campione si ottengono i seguenti valori

$$\bar{x} = \frac{747.0 + 751.5 + 752.0 + 747.5 + 747.0 + 749.0}{6} = 749.0$$

$$s^2 = \frac{1}{5} \cdot (747.0^2 + 751.5^2 + 752.0^2 + 747.5^2 + 747.0^2 + 749.0^2 - 6 \cdot 749^2) = 5.1$$

Il valore della statistica test è

$$T = \frac{749.0 - 750}{\frac{\sqrt{5.1}}{\sqrt{6}}} = -1.08.$$

Il test è a una coda, e per il livello di significatività del 5% e il grado di libertà  $v = 5$  il valore critico è

$$t_\alpha = -t_{0.05} = -2.015$$

La regione di rifiuto è data dai valori  $T < -2.015$ .

Il valore  $T = -1.08$  appartiene alla regione di accettazione, perciò non c'è un'evidenza significativa, al livello del 5%, che le bottiglie contengano meno di 750ml di vino.

### **Esempio 17**

Una prova del carico di rottura di 6 cavi d'acciaio costruiti da una ditta ha mostrato un carico di rottura medio  $\bar{x} = 7750$  kg e uno scarto quadratico medio  $s = 145$ kg, mentre il costruttore afferma che il carico di rottura medio è di 8000kg.

E' possibile sostenere l'affermazione del costruttore ad un livello di significatività del 5% e dell'1%?

L'ipotesi nulla e l'ipotesi alternativa sono

$$\begin{aligned} H_0: & \mu = 8000 \\ H_1: & \mu < 8000. \end{aligned}$$

Il valore della statistica test è

$$T = \frac{7750 - 8000}{\frac{145}{\sqrt{6}}} = -4.22.$$

Il test è a una coda, e per il livello di significatività del 5% e il grado di libertà  $v = n - 1 = 5$ , il valore critico è

$$t_\alpha = -t_{0.05} = -2.015$$

La regione di rifiuto è data dai valori  $T < -2.015$ . Il valore  $T = -4.22$  appartiene alla regione di rifiuto, perciò rifiutiamo l'ipotesi nulla al livello di significatività del 5%.

Per il livello di significatività dell'1% e il grado di libertà  $v = n - 1 = 5$ , il valore critico è

$$t_\alpha = -t_{0.01} = -3.365$$

La regione di rifiuto è data dai valori  $T < -3.365$ . Il valore  $T = -4.22$  appartiene alla regione di rifiuto, perciò anche al livello di significatività dell'1% rifiutiamo l'ipotesi nulla.

In conclusione non possiamo sostenere che l'affermazione del costruttore sia giustificata per nessuno dei due livelli di significatività.

### **Esempio 18**

Si estrae un campione di 8 confezioni di detersivo in polvere da una grossa produzione.

I pesi in g delle 8 confezioni sono

1998.5	2000.4	1999.9	2005.8
2011.5	2007.6	2001.3	2002.4

# Esercizi di Statistica, con soluzioni e non solo...

*G. Marchetti*

*2016 ver. 1.9*

## Indice

<b>1</b>	<b>Introduzione</b>	<b>1</b>
<b>2</b>	<b>Indici</b>	<b>3</b>
<b>3</b>	<b>Indici di associazione</b>	<b>6</b>
<b>4</b>	<b>Probabilità</b>	<b>7</b>
<b>5</b>	<b>Variabili casuali discrete</b>	<b>14</b>
<b>6</b>	<b>Variabili casuali doppie</b>	<b>21</b>
<b>7</b>	<b>Variabili casuali continue</b>	<b>24</b>
<b>8</b>	<b>Stima e stimatori</b>	<b>32</b>
<b>9</b>	<b>Test delle ipotesi</b>	<b>42</b>

## 1 Introduzione

- Introduzione alla statistica
- Fenomeni collettivi
- Distinzione tra unità e variabili
- Classificazione delle variabili qualitative/quantitative, discrete/continue
- Distribuzione di frequenza
- Diagramma stem-and-leaf
- Istogramma
- Distribuzioni doppie di frequenza

### 1.1

Considera 20 famiglie. Per ciascuna rileva il numero di componenti. Ecco i dati:

1 3 2 5 4 2 2 3 3 2 3 4 4 3 2 5 4 3 3 1

Costruisci la distribuzione di frequenza.

#### Soluzione

Componenti	1	2	3	4	5
Famiglie	2	5	7	4	2

## 1.2

Nell'esempio precedente dire

- qual è l'unità
- qual è la variabile
- La variabile è quantitativa o qualitativa?

### Soluzione

L'unità è la famiglia, la variabile il numero di componenti, quantitativa discreta.

## 1.3

Per 20 giorni hai registrato i minuti di ritardo del treno per arrivare a Firenze.

Eccoli:

28 5 4 12 17 12 14 5 4 4 11 8 4 26 17 6 0 19 8 38

Fai un grafico stem-and-leaf prendendo come stelo la cifra delle decine. Costruisci una distribuzione di frequenza con classi (in minuti)

0-9 10-19 20-29 30-39

### Soluzione

```
0 | 0444455688
1 | 1224779
2 | 68
3 | 8
```

## 1.4

Nell'esempio precedente dire

- qual è l'unità
- qual è la variabile
- La variabile è quantitativa o qualitativa?

### Soluzione

L'unità è il viaggio (o il treno), la variabile è il ritardo in min. quantitativa, continua.

## 1.5

In un'ora una libreria fa 20 scontrini per i seguenti importi in Euro

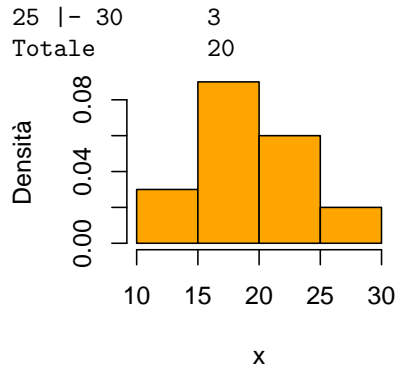
10 13 13 18 18 18 19 19 20 20 20 20 22 22 23 24 24 25 26 27

Fate un istogramma con classi

10 |- 15 , 15 |- 20, 20 |- 25, 25 |- 30

### Soluzione

Importo	n. clienti
10  - 15	3
15  - 20	5
20  - 25	9



## 1.6

Ecco la distribuzione del salario mensile (in Euro) di un campione di 1000 lavoratori:

Classi di reddito	Frequenze	Altezze dei rettangoli
0   - 500	100	0.2
500   - 1000	200	0.4
1000   - 2000	500	0.5
2000   - 4000	600	0.3
4000   - 8000	400	0.1

Disegnare l'istogramma e verificare che le aree dei rettangoli sono uguali alle frequenze. Come sono state calcolate le altezze dei rettangoli?

### Soluzione

Le altezze sono = frequenze/ampiezza di classe.

## 2 Indici

- Moda
- Media
- Mediana
- Deviazione standard e varianza
- Quartili e quantili
- Box-plot
- Introduzione alla disuguaglianza di Chebyshev

### 2.1

Le temperature nella località X alle 12 sono state

12.1 14.5 9.7 8.1 13.0 12.5 10.5

Calcola la media e la mediana.

### Soluzione

Media =  $80.4/7 = 11.48$  gradi centigradi.

Mediana = 12.1 gradi centigradi.

8.1 9.7 10.5 (12.1) 12.5 13.0 14.5

## 2.2

La distribuzione di 520 studenti per numero di esami superati è

Esami	Studenti
0	50
1	100
2	160
3	120
4	80
5	10

Calcolare le frequenze relative, la moda e il numero medio di esami superati. Calcolare la mediana.

### Soluzione

Esami	Studenti	Freq. relative	Cumulate	Prodotto
0	50	0.10	0.10	0
1	100	0.19	0.29	100
2	160	0.31	0.60	320
3	120	0.23	0.83	360
4	80	0.15	0.98	320
5	10	0.02	1.00	50
Totale	520	1.00		1150

La moda è 2 esami superati. La media è  $1150/520 = 2.21$  esami. La mediana si trova notando che le due unità centrali sono la 260 e la 261-ma nella successione ordinata e dalla distribuzione si vede che stanno entrambe nella classe di 2 esami.

Quindi la mediana è 2 esami superati.

## 2.3

La popolazione delle prime 10 città americane in milioni è la seguente.

New York (New York)	9.21
Los Angeles (California)	4.05
Chicago (Illinois)	2.83
Houston (Texas)	2.01
Phoenix (Arizona)	1.55
Filadelfia	1.45
Dallas (Texas)	1.31
San Diego (California)	1.30
San Antonio (Texas)	1.24
San Jose (California)	0.94

Calcolare la popolazione media e la popolazione mediana. Quale indice è migliore?

### Soluzione

La popolazione totale è 25.80 milioni. Quindi la media è 2.58 milioni. La mediana è la semisomma tra 1.45 e 1.55 cioè 1.5 milioni di abitanti.

È meglio la mediana perché non risente troppo dei valori anomali (come New York).

## 2.4

Ecco il voto di laurea di 5 studenti di Lettere

110 109 108 110 110

Ecco il voto di laurea di 5 studenti di Economia

90 98 110 105 102

C'è maggiore variabilità di voto a Economia o a Lettere? Giustificare calcolando le varianze del voto e le deviazioni standard.

### Soluzione

- A Lettere il voto medio di laurea è 109.4.
- A Economia il voto medio è 101.
- La varianza del voto a Lettere è 0.8
- La varianza del voto a Economia è 57. (Ho usato il denominatore  $n = 4$ .)

Le deviazioni standard sono perciò

- Lettere punti 0.8944
- Economia punti 7.5498.

Evidentemente la variabilità è minore a Lettere. Volendo usare il *coefficiente di variazione* per ottenere una misura relativa di variabilità si ottiene

- Lettere CV = 0.0082
- Economia CV = 0.0748

Quindi il coefficiente di variazione è minore a Lettere.

## 2.5

Considera la distribuzione di 1000 studenti secondo il voto di laurea

Voto	Frequenza
98	10
99	40
100	250
101	400
102	250
103	40
104	10

- Mostrate che la media la moda e la mediana sono uguali a 101. Verificate che la deviazione standard è di 1 punto.
- Calcolate la frequenza relativa di studenti che hanno preso un voto compreso tra  $101 - 2 = 99$  e  $101 + 2 = 103$ . Secondo la disuguaglianza di Chebyshev questa frequenza relativa quanto dovrebbe essere?

### Soluzione

La distribuzione è simmetrica e quindi media moda e mediana sono uguali. La moda e la mediana sono evidentemente uguali a 101 punti. La media è  $101000/1000 = 101$ .

Per calcolare la varianza compiliamo la tabella seguente

Voto	Freq.	Voto <sup>2</sup>	Voto <sup>2</sup> * freq.
98	10	9604	96040
99	40	9801	392040
100	250	10000	2500000
101	400	10201	4080400
102	250	10404	2601000
103	40	10609	424360
104	10	10816	108160
Totale	1000		10202000

Quindi la media dei voti al quadrato è

$$MQ = 10202000/1000 = 10202$$

e la varianza è uguale alla MQ meno la media al quadrato :

$$MQ - (media^2) = 10202 - 101^2 = 1.$$

Quindi anche la deviazione standard è 1.

La proporzione di studenti che hanno preso un voto tra 99 e 103 è  $980/1000 = 98\%$ .

Come vedremo la disuguaglianza di Chebychev asserisce che per forza questa proporzione deve essere maggiore di  $1 - 1/4 = 75\%$ . E infatti così avviene.

### 3 Indici di associazione

- Covarianza
- Coefficiente di correlazione
- Retta dei minimi quadrati

#### 3.1

Su 4 famiglie di 2 componenti misuriamo il reddito di Febbraio  $X$  e le relative spese per l'alimentazione  $Y$ .

X: 1500 1700 1400 1600  
Y: 200 350 150 300

Calcolare la covarianza e il coefficiente di correlazione e interpretarli.

#### Soluzione

$media(X) = 1550$ ,  $media(Y) = 250$ ,  $var(X) = 12500$ ,  $var(Y) = 6250$ .

$(X - media(X))(Y - media(Y)) : 2500, 15000, 15000, 2500$

$cov(X, Y) = 35000/4 = 8750$ .

$cor(X, Y) = 8750/\sqrt{(12500)(6250)} = 0.9899$ .

#### 3.2

Provate a calcolare la covarianza con la formula equivalente

$media(XY) - media(X) media(Y)$

#### Soluzione

Poiché  $\text{media}(XY) = (1500 \times 200 + 1700 \times 350 + 1400 \times 150 + 1600 \times 300)/4$  abbiamo  
 $\text{cov}(X, Y) = \text{media}(XY) - \text{media}(X) \text{media}(Y) = 396250 - (1550)(250) = 8750$ .

### 3.3

Rappresentate la relazione tra spesa e reddito con la retta dei minimi quadrati e verificate che questa è:  
 $\text{Spesa} = -835 + 0.7 \text{ Reddito}$ . Provate a interpretare il risultato.

#### Soluzione

Coefficiente angolare:

$$\text{cor}(X, Y) \sqrt{\text{var}(Y)/\text{var}(X)} = \text{cov}(X, Y)/\text{var}(X) = 8750/12500 = 0.7.$$

La retta deve passare per il punto  $(\text{media}(X), \text{media}(Y)) = (1550, 250)$  e quindi ha equazione

$$y = 250 + 0.7(x - 1550) = -835 + 0.7x.$$

La pendenza ha l'interpretazione: per ogni Euro in più di reddito la spesa aumenta di 70 centesimi.

### 3.4

Vero o Falso? Se  $X$  e  $Y$  sono due variabili con  $\text{var}(X) = 3.25$ ,  $\text{var}(Y) = 5.8$ ,  $\text{cov}(X, Y) = 14.703$  allora il coefficiente di correlazione è 0.78. Giustificare.

#### Soluzione

Falso. Perché  $0.78 = \text{cov}(X, Y)/(\text{var}(X)\text{var}(Y))$  mentre il coefficiente di correlazione è

$$\text{cor}(X, Y) = \text{cov}(X, Y)/\sqrt{\text{var}(X)\text{var}(Y)}.$$

## 4 Probabilità

1. Esperimenti aleatori
2. Eventi elementari, eventi
3. Probabilità (classica e frequentista)
4. Combinazioni
5. Probabilità condizionata
6. Indipendenza
7. Tavole di probabilità
8. Formula di Bayes

- Un esperimento aleatorio è un esperimento il cui esito è incerto.
- Un evento è un insieme di esiti specificati prima di fare l'esperimento.
- Un evento si verifica se si verifica uno dei suoi componenti elementari.
- La probabilità è una misura  $P(A)$  dell'incertezza associata al verificarsi di evento  $A$ .

Gli eventi si comportano e si combinano fra loro come insiemi.

Si distinguono:

- L'evento certo  $S$  che si verifica sempre
- L'evento impossibile  $\emptyset$  che non si verifica mai

## 7.12

Una variabile  $X$  ha una distribuzione con media 250 e deviazione standard 20. Dare indicazioni sulla probabilità:

- $P(210 < X < 290)$
- $P(220 < X < 280)$

Determinare le stesse probabilità sapendo che  $X$  è normale  $N(250, \sigma = 20)$ .

### Soluzione

Se la distribuzione di  $X$  è ignota si usa la disuguaglianza di Chebyshev. Entrambi gli intervalli hanno come punto centrale la media della distribuzione.

Per esempio, nel primo caso  $(210, 290)$  ha punto centrale  $(210 + 290)/2 = 250$ .

Si trova il raggio dell'intorno di 250 che è  $(290 - 210)/2 = 40$ . Quindi l'intervallo è  $250 \pm 40$ .

Quindi si esprime il raggio come un multiplo della deviazione standard cioè  $40 = k \cdot 20$  da cui si ottiene  $k = 2$  (il raggio 40 è il doppio della deviazione standard).

Infine si usa la disuguaglianza:

$$P(210 < X < 290) \geq 1 - 1/(k^2) = 1 - 1/4 = 0.75.$$

Nel secondo caso il raggio è  $(280 - 220)/2 = 60/2 = 30$  e poiché  $30 = k \cdot 20$  implica  $k = 1.5$ . Quindi abbiamo

$$P(220 < X < 280) \geq 1 - 1/(k^2) = 1 - 1/(1.5^2) = 0.555$$

Sapendo che  $X$  è normale le probabilità si possono ottenere usando le tavole.

- $P(210 < X < 290) = P(-2 < Z < 2) = 2(0.9772) - 1 = 0.9544$
- $P(220 < X < 280) = P(-1.5 < Z < 1.5) = 2(0.9332) - 1 = 0.8664$

## 8 Stima e stimatori

1. Campioni casuali e distribuzioni campionarie
2. Stima
3. Stimatori corretti
4. Errore standard
5. Teorema centrale del limite
6. Intervalli di confidenza per la media (varianza nota)
7. Intervalli di confidenza per la media (varianza incognita)
8. Distribuzione t e tavole dei quantili
9. Intervalli di confidenza asintotici per una proporzione
10. Ampiezza dell'intervallo di confidenza
11. Scelta della dimensione campionaria
  - Un campione casuale da una popolazione con una distribuzione  $X$  è un' $n$ -upla di osservazioni  $(x_1, \dots, x_n)$  che si può equiparare alle realizzazioni di  $n$  variabili aleatorie  $X_1, \dots, X_n$  che siano
    - indipendenti
    - identicamente distribuite come  $X$ .

- I campioni casuali sono importanti per stimare una caratteristica della popolazione senza fare un censimento.
- La caratteristica della popolazione  $X$  si chiama *parametro*. Per esempio  $\mu$  o  $\sigma^2$  sono parametri di una popolazione continua. Invece una proporzione di successi  $p$  è un parametro di una popolazione binaria (composta di 0 e 1).
- Il parametro si stima con i dati campionari  $x_1, \dots, x_n$ . Le stime più comuni sono
  - $\bar{x} = \sum_{i=1}^n x_i/n$  la media campionaria stima  $\mu$ .
  - $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$  stima la varianza  $\sigma^2$ .
  - $\hat{p}$  = proporzione di successi nel campione, stima  $p$ .
- Uno *stimatore* è la stima nel campionamento ripetuto. Si descrive con una variabile aleatoria che esprime il variare della stima nell'*universo dei campioni*. Gli stimatori corrispondenti alle stime precedenti sono
  - Lo stimatore di  $\mu$ :  $\bar{X} = \sum_{i=1}^n X_i/n$
  - Lo stimatore di  $\sigma^2$ :  $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ .
  - Lo stimatore di  $p$ :  $\hat{P} = \sum_{i=1}^n X_i/n$ .

- Uno stimatore è una variabile aleatoria con una distribuzione detta *distribuzione campionaria*.
- Dato un campione casuale iid  $(X_1, \dots, X_n)$  da una popolazione  $X$  continua con una media  $\mu$  si dice *stimatore media campionaria*  $\bar{X} = (X_1 + \dots + X_n)/n$ .
- Lo stimatore  $\bar{X}$  è rappresentato dalla sua *distribuzione campionaria* cioè dalla distribuzione ottenuta calcolando la media in ogni campione dell'*universo dei campioni*.
- Lo stimatore va distinto dalla *stima* della media cioè il semplice dato  $\bar{x} = (x_1 + \dots + x_n)/n$  indicato con la lettera minuscola.
- Lo stimatore *media campionaria*  $\bar{X} = (X_1 + \dots + X_n)/n$  è *corretto* per  $\mu$  cioè

$$E(\bar{X}) = \mu$$

qualunque sia la popolazione e qualunque sia  $\mu$ .

- Dati due stimatori corretti di  $\mu$  si dice che  $T_1$  è più *efficiente* di  $T_2$  se

$$\text{var}(T_1) \leq \text{var}(T_2)$$

Per esempio la media campionaria  $\bar{X}$  è più efficiente della mediana campionaria in campioni provenienti da una distribuzione binomiale.

- Lo stimatore  $\bar{X}$  ha varianza  $\sigma^2/n$  e la sua deviazione standard, chiamata *errore standard* nel contesto della stima, è

$$ES = \frac{\sigma}{\sqrt{n}}.$$

- Se la popolazione è normale lo stimatore  $\bar{X}$  ha distribuzione normale  $N(\mu, \sigma^2/n)$  esattamente, qualunque sia la dimensione del campione.
- Se la popolazione non è normale, ma la dimensione del campione è sufficientemente grande ( $> 100$ ) lo stimatore  $\bar{X}$  ha distribuzione approssimata da  $N(\mu, \sigma^2/n)$  (teorema centrale del limite).
- Data una popolazione dicotomica, cioè di 0 e 1, con una proporzione di 1 (i successi) pari a  $p$ , e un campione casuale iid  $(X_1, \dots, X_n)$  si dice *stimatore proporzione campionaria*  $\hat{P} = (X_1 + \dots + X_n)/n = \# \text{successi}/n$ .
- Lo stimatore  $\hat{P}$  è rappresentato dalla sua distribuzione campionaria che è esattamente Binomiale divisa per  $n$ .

- La proporzione campionaria  $\hat{P}$  è uno stimatore corretto di  $p$ , cioè

$$E(\hat{P}) = p$$

qualunque sia  $p$  e per ogni numerosità  $n$ .

- La varianza della proporzione  $\hat{P}$  è  $p(1-p)/n$ .
- L'errore standard di  $\hat{P}$  è

$$ES(\hat{P}) = \sqrt{p(1-p)/n}.$$

- la distribuzione campionaria di una proporzione  $\hat{P}$  è approssimativamente normale se  $np(1-p) > 9$

$$\hat{P} \approx N(p, p(1-p)/n).$$

- Gli errori standard dipendono dai parametri incogniti e quindi vengono stimati opportunamente.

$$\begin{aligned} - \hat{ES}(\bar{X}) &= s/\sqrt{n} \\ - \hat{ES}(\hat{P}) &= \sqrt{\hat{p}(1-\hat{p})/n} \end{aligned}$$

Questo fornisce una misura calcolabile dell'errore di campionamento che si commette nella stima del parametro.

- La distribuzione di  $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}$  è normale standard.
- La distribuzione di  $\frac{\bar{X}-\mu}{\sqrt{s^2/n}}$  dove

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

è *t di Student* con  $n-1$  gradi di libertà. I suoi quantili sono tabulati sul libro. Quando  $n > 100$  la *t di Student* si può approssimare con una normale standard.

- Un *intervallo di confidenza* per  $\mu$  al livello  $c$  (tipicamente  $c = 0.95$  o  $c = 0.99$ ) è un intervallo di stima con estremi  $A$  e  $B$  tali nell'universo dei campioni una proporzione  $c$  di campioni produce intervalli che contengono  $\mu$ :

$$P(A < \mu < B) = c.$$

- Un intervallo di confidenza di livello  $1-\alpha$  per la media di una distribuzione normale *con varianza nota* è

$$\bar{X} \pm z \cdot \sigma/\sqrt{n}$$

dove  $z$  è tale che  $P(Z < z) = 1 - \alpha/2$  ossia  $P(Z > z) = \alpha/2$  e  $Z$  è la normale standard.

- Un intervallo di confidenza di livello  $1-\alpha$  per la media di una distribuzione normale è

$$\bar{X} \pm t \cdot \sqrt{s^2/n}$$

dove  $s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$ . Il valore  $t$  è un quantile della *t di Student*, cioè è tale che  $P(T < t) = 1 - \alpha/2$  ossia  $P(T > t) = \alpha/2$  in cui  $T$  è la *t di Student* con  $n-1$  gradi di libertà.

- Un intervallo di confidenza di livello approssimato  $1-\alpha$  per la media di una distribuzione normale con un campione di dimensione elevata è

$$\bar{X} \pm z \cdot \sqrt{s^2/n}$$

dove  $z$  è tale che  $P(Z < z) = 1 - \alpha/2$  ossia  $P(Z > z) = \alpha/2$  e  $Z$  è normale standard.

- Un intervallo di confidenza di livello approssimato  $1-\alpha$  per la proporzione di una popolazione dicotomica in cui  $n\hat{p}(1-\hat{p}) > 9$  è

$$\hat{p} \pm z \cdot \sqrt{\hat{p}(1-\hat{p})/n}$$

dove  $z$  è tale che  $P(Z < z) = 1 - \alpha/2$  ossia  $P(Z > z) = \alpha/2$  e  $Z$  è normale standard.

## 8.1

Un'azienda produce un modello di auto la cui percorrenza  $X$  (in km con 1 litro di benzina) ha distribuzione normale, media 25 km/l e deviazione standard 2 km/l. Supponiamo di avere un campione casuale di 4 auto prodotte in serie.

- La percorrenza media campionaria che distribuzione ha?
- Qual è la probabilità che la percorrenza media sia superiore a 26 km/l?
- Ricalcolare la probabilità precedente con una dimensione campionaria di 25 auto.

### Soluzione

$X \sim N(25, \sigma = 2)$ . Campione =  $(X_1, X_2, X_3, X_4)$  indipendenti e identicamente distribuiti come  $X$ .

- La percorrenza media è  $\bar{X} = (X_1 + X_2 + X_3 + X_4)/4$  ed ha distribuzione normale  $N(25, \sigma_{\bar{X}} = 2/\sqrt{4} = 1)$ .
- $P(\bar{X} > 26) = P(Z > (26 - 25)/1) = P(Z > 1) = 1 - P(Z < 1) = 1 - 0.8413 = 0.1587$ .
- $P(\bar{X} > 26) = P(Z > (26 - 25)/(2/5)) = P(Z > 2.5) = 1 - P(Z < 2.5) = 1 - 0.9938 = 0.0062$ .

## 8.2

Una popolazione di studenti è composta dal 40% di femmine e dal 60 % di maschi. Se si estrae un campione casuale con ripetizione di 25 studenti qual è la distribuzione di probabilità della proporzione di femmine nel campione? Qual è la varianza della proporzione di femmine nel campione?

### Soluzione

$X \sim Bernoulli$  con  $p = 0.4, q = 0.6$ . Perciò  $\text{var}(X) = pq = 0.24$ .

Campione di  $n = 25$  elementi  $(X_1, \dots, X_{25})$ .

$\$P = \$$  Proporzione di femmine =  $\#femmine / 25 = (X_1 + \dots + X_{25})/25$ .

$P * 25$  è distribuita come una Binomiale(25,  $p = 0.4$ ). Sappiamo allora che

$$\text{var}(P) = \text{var}(X)/n = 0.24/25 = 0.0096.$$

## 8.3

Sia  $X$  la distribuzione dell'età di una popolazione con  $E(X) = 50$  anni e  $\sigma(X) = 10$  anni. Se seleziono un campione di  $n = 4$  persone e calcolo la media:

- Si conosce la distribuzione campionaria dell'età media?
- Si conosce il valore atteso della distribuzione campionaria?
- Si conosce la varianza della distribuzione campionaria?

Giustificare.

### Soluzione

$X = \text{età} \sim \text{incognita} ? (\mu = 50, \sigma = 10)$ .

Campione:  $(X_1, X_2, X_3, X_4)$  indipendenti e identicamente distribuiti come  $X$ .

$\bar{X} = (X_1 + X_2 + X_3 + X_4)/4$ .

- Si conosce la distribuzione campionaria della media? NO è incognita.
- Si conosce il valore atteso della distribuzione campionaria della media? SÌ, è  $E(\bar{X}) = \mu = 50$  anni.
- Si conosce la varianza della distribuzione campionaria? SÌ, è  $\text{var}(\bar{X}) = \sigma^2/n = 100/4 = 25$ .

## 8.4

Rispondere all'esercizio precedente se  $n = 100$ .

### Soluzione

- Si conosce la distribuzione campionaria della media? SÌ, poiché la dimensione del campione è grande ha approssimativamente distribuzione normale.
- Si conosce il valore atteso della distribuzione campionaria della media? SÌ, è  $E(\bar{X}) = \mu = 50$  anni.
- Si conosce la varianza della distribuzione campionaria? SÌ, è  $\text{var}(\bar{X}) = \sigma^2/n = 100/100 = 1$ .

## 8.5

Il numero di televisori che escono ogni giorno da una certa linea di produzione si distribuisce come una variabile casuale con deviazione standard (nota) di 17.4. La media giornaliera della linea di produzione determinata su un campione di 20 giorni è 452.3. Quale dei seguenti intervalli rappresenta un intervallo di confidenza al 95% per la media della produzione in un giorno?

- A)  $453 \pm 9.4$
- B)  $452.3 \pm 13.8$
- C)  $452.3 \pm 11.3$
- D)  $452.3 \pm 7.63$

### Soluzione

$X = \#$  televisioni  $\sim N(\mu = ?, \sigma = 17.4)$ .

NOTA: il sigma fornito dal testo è la deviazione standard **della popolazione**.

La stima di  $\mu$  con un campione di  $n = 20$  elementi è 452.3 con un errore standard  $ES = \sigma/\sqrt{20} = 17.4/\sqrt{20} = 3.890758$ . L'intervallo di confidenza (IC) al 95% è

$$452.3 \pm ME$$

con un margine di errore

$$ME = 1.96 ES = (1.96)(3.890758) = 7.625886.$$

Quindi la risposta è (arrotondando a 2 decimali) la D).

## 8.6

L'errore di stima è la differenza tra il valore di una statistica determinata su un campione ed il corrispondente valore del parametro determinato nella popolazione. Vero o falso?

### Soluzione

L'errore di stima è la differenza tra uno stimatore e il parametro. Il termine "statistica" è sinonimo di stimatore, e vuol dire un indice calcolato sul campione.

Quindi la risposta è: Vero.

## 8.7

Il tempo che gli studenti dedicano allo studio segue una distribuzione normale con deviazione standard di 8 ore. Si estrae un campione casuale di 4 studenti. La probabilità che la media campionaria **differisca dalla media della popolazione** per più di 4 ore è

A) 0.2987 B) 0.3080 C) 0.3174 D) 0.3085

**Soluzione**

Tempo =  $X \sim N(\mu = ?, \sigma = 8)$ .

Se ho un campione di dimensione  $n = 4$ , la media campionaria  $\bar{X} = (X_1 + X_2 + X_3 + X_4)/4$  ha distribuzione

$$\bar{X} \sim N(\mu = ?, \sigma_{\bar{X}} = 8/2 = 4).$$

La probabilità che  $\bar{X}$  differisca da  $\mu$  per più di 4 ore è

$$P(|\bar{X} - \mu| > 4) = 1 - P(-4 < \bar{X} - \mu < 4) = 1 - P(\mu - 4 < \bar{X} < \mu + 4)$$

NOTA: Fate attenzione a queste disuguaglianze, studiatele con calma.

Quindi se si standardizza  $\bar{X}$  rispetto alla sua media  $\mu$  e alla sua deviazione standard = 4 si ha

$$P(|\bar{X} - \mu| > 4) = 1 - P(-1 < Z < 1) = 1 - (0.8413 - (1 - 0.8413)) = 0.3174$$

e quindi la risposta corretta è la C).

### 8.8

Uno stimatore è una variabile casuale calcolata su un campione casuale che fornisce la stima puntuale per il parametro della popolazione. Vero o falso?

**Soluzione** È esattamente così: la stima è un numero, mentre lo stimatore è una variabile casuale calcolata sul campione che fornisce una stima del parametro della popolazione. Quindi: Vero.

### 8.9

Un intervallo di confidenza al 95% per la media della popolazione  $\mu$  è stimato da 65.48 a 76.52. Se ora viene stimato un intervallo di confidenza al 90% per  $\mu$  sarà:

- A) più ampio di quello al 95%.
- B) lo stesso dell'intervallo al 95%.
- C) più stretto di quello al 95%.
- D) Non c'è abbastanza informazione per rispondere.

**Soluzione**

Un intervallo di confidenza è

$$\bar{X} \pm ME$$

dove  $ME$ , il margine di errore è

$$ME = z_{\alpha/2} \sigma / \sqrt{n}.$$

L'ampiezza dell'intervallo cresce o decresce con  $ME$ . Se il livello di confidenza cambia,  $ME$  cambia perché cambia  $z_{\alpha/2}$ . Allora, ad esempio abbiamo:

Livello	$1 - \alpha$	$\alpha/2$	$z_{\alpha/2}$
95%	95%	2.5%	1.96
90%	90%	5%	1.64

Quindi se il livello è 90% il  $ME$  è più piccolo e l'intervallo di confidenza è meno ampio. La risposta giusta è C).

### 8.10

Un'agenzia turistica è interessata all'ammontare medio di denaro speso al giorno da un tipico studente universitario durante le vacanze estive. Un'indagine condotta su 30 studenti mette in luce che la somma media spesa è 63.57 Euro con una deviazione standard di 17.32 Euro. Determinare l'intervallo di confidenza al 95% per la spesa media nella popolazione.

#### Soluzione

Spesa giornaliera di uno studente =  $X \sim N(\mu=?, \sigma=?)$ . Da un campione di  $n = 30$  studenti si sa che

$$\bar{X} = 63.57 \text{ Euro} = 17.32 \text{ Euro}$$

NOTA: la deviazione standard fornita è quella del campione NON quella della popolazione.

Quindi l'IC per  $\mu$  è basato sulla  $t$  di Student con  $n - 1 = 29$  gradi di libertà:

$$\bar{X} \pm t_{\alpha/2} s / \sqrt{n}$$

ossia

$$63.57 \pm (2.045)17.32/\sqrt{30}$$

Cioè IC = (57.10333, 70.03667).

### 8.11

Da una popolazione infinita con media pari a 80 e deviazione standard 18, vengono selezionati campioni casuali di dimensione  $n = 36$ . La media e l'errore standard della relativa distribuzione campionaria della media sono rispettivamente:

- A) 80 e 18. B) 80 e 3. C) 36 e 2. D) 80 e 2.

#### Soluzione

La media  $\bar{X}$  di un campione iid di  $n = 36$  elementi da una QUALSIASI distribuzione  $X$  è tale che

$$E(\bar{X}) = \mu, \quad \text{var}(\bar{X}) = \sigma^2(X)/n, \quad ES(\bar{X}) = \sigma(X)/\sqrt{n}$$

Quindi in questo caso

$$E(\bar{X}) = 80, \quad ES(\bar{X}) = 18/6 = 3.$$

Quindi la soluzione è B).

### 8.12

Se l'errore standard della distribuzione della proporzione campionaria è 0.0229 per campioni di dimensione 400, allora la vera proporzione nella popolazione deve essere:

- A) 0.2 o 0.8.  
B) 0.3 o 0.7.  
C) 0.4 o 0.6.  
D) 0.5 o 0.5.

#### Soluzione

Una proporzione campionaria  $P$  ha un errore standard

$$ES(P) = \sqrt{pq/n}$$

poichè questo deve essere 0.0229 basta provare nei casi A), B), C), D) che cosa otteniamo:

- A)  $ES = \sqrt{(0.2 \cdot 0.8/400)} = 0.02$
- B)  $ES = \sqrt{(0.3 \cdot 0.7/400)} = 0.02291288$
- C)  $ES = \sqrt{(0.4 \cdot 0.6/400)} = 0.0244949$
- D)  $ES = \sqrt{(0.5 \cdot 0.5/400)} = 0.025$

Quindi, approssimando, la risposta giusta è B).

### 8.13

Nel costruire un intervallo di confidenza per la media della popolazione è stato utilizzato un campione di 40 osservazioni. La stima intervallare risultante è stata  $28.76 \pm 1.48$ . Se la numerosità campionaria fosse stata 160 invece che 40, la stima intervallare sarebbe stata:

- A)  $28.76 \pm 0.74$ .
- B)  $28.76 \pm 0.37$ .
- C)  $7.19 \pm 0.37$ .
- D)  $7.19 \pm 1.48$ .

**Soluzione** Siccome 160 è una numerosità quadrupla di quella di partenza di  $n = 40$ , l'ampiezza dell'intervallo di confidenza è la metà perché l'errore standard è

$$ES = \sigma/\sqrt{4n} = (1/2)\sigma/\sqrt{n}.$$

quindi il margine di errore con  $n = 160$  è  $ME = 1.48/2 = 0.74$ . Quindi la risposta giusta è A).

NOTA: Naturalmente qui si suppone che sia nel caso  $n = 40$  che nel caso  $n = 160$  la media campionaria sia sempre la stessa.

### 8.14

Siano  $X_1, X_2, X_3$  e  $X_4$  le osservazioni di un campione casuale semplice estratto da una popolazione  $X$  con media  $\mu$  e varianza  $\sigma^2$ . Si consideri il seguente stimatore di  $\mu$ :

$$T = 0.15X_1 + 0.35X_2 + 0.20X_3 + 0.30X_4.$$

Qual è la varianza di  $T$ ?

**Soluzione** La varianza di  $T$  è

$$\begin{aligned} \text{var}(T) &= \text{var}(0.15X_1 + 0.35X_2 + 0.20X_3 + 0.30X_4) \\ &= 0.15^2 \text{var}(X_1) + 0.35^2 \text{var}(X_2) + 0.20^2 \text{var}(X_3) + 0.30^2 \text{var}(X_4) \\ &= (0.15^2 + 0.35^2 + 0.20^2 + 0.30^2) \text{var}(X) \\ &= 0.275\sigma^2. \end{aligned}$$

### 8.15

La distribuzione campionaria della media avrà la stessa media della popolazione dalla quale sono stati estratti i campioni che l'hanno generata. Vero o falso?

**Soluzione**

È vero perché la media campionaria è uno stimatore corretto di  $\mu$ :

$$E(\bar{X}) = \mu.$$

## 8.16

Trova il quantile della  $t$  di Student con 8 gradi di libertà che lascia a destra una probabilità di 0.025.

**Soluzione**

$$t(0.025) = 2.306$$

## 8.17

Si supponga che il tempo medio che un ragazzo passa su Facebook sia distribuito come una variabile normale con una deviazione standard di 1.5 ore. In un campione di 100 ragazzi è stata rilevata una media di 6.5 ore. Determinare l'intervallo di confidenza al 95% per il tempo medio passato su Facebook nella popolazione.

**Soluzione**

Tempo passato su Facebook =  $X \sim N(\mu = ?, \sigma = 1.5)$ . L'intervallo di confidenza è basato sulla normale perché la varianza della popolazione è nota.

$$IC = \bar{X} \pm ME$$

ossia

$$IC = 6.5 \pm 1.96(1.5)/10$$

da cui  $IC = (6.206, 6.794)$ .

## 8.18

Un ricercatore, incaricato di stimare la percentuale di famiglie italiane che hanno più di un computer, dopo aver rilevato che il 27% di un campione costituito da 492 famiglie ha dichiarato di possedere più di un computer, fornisce l'intervallo di confidenza (0.2308; 0.3092), ma omette di dire il livello di confidenza. Qual è il livello di confidenza associato a questo intervallo?

**Soluzione** Un intervallo di confidenza per  $p$  (approssimato per grandi campioni) è

$$\hat{p} \pm z_{\alpha/2} ES \quad \text{dove} \quad ES = \sqrt{\hat{p}\hat{q}/n}.$$

Il margine di errore è  $ME = z_{\alpha/2} ES$  e quindi  $z_{\alpha/2} = ME/ES$ .

Qui  $\hat{p}$  è la proporzione stimata che ovviamente è il punto centrale dell'intervallo cioè

$$\hat{p} = (0.2308 + 0.3092)/2 = 0.27.$$

L'errore standard è

$$ES = \sqrt{0.27 * 0.73/492} = 0.02.$$

Il margine di errore è la lunghezza di mezzo intervallo e lo sappiamo:

$$ME = 0.3092 - 0.27 = 0.0392.$$

Perciò

$$z_{\alpha/2} = ME/ES = 0.0392/0.02 = 1.96$$

Siamo fortunati! A occhio sappiamo che il livello di confidenza è il 95%.

Ecco tutti i passaggi:

$$1 - \alpha/2 = P(Z < 1.96) = 0.975$$

Quindi  $\alpha = 0.05$  e  $1 - \alpha =$  livello di confidenza  $= 0.95$ .

## 8.19

Determinare l'ampiezza campionaria necessaria per stimare la proporzione  $p$  nella popolazione se  $ME = 0.05$  e il livello di confidenza è il 99%.

### Soluzione

Deve essere  $ME = 2.58 \sqrt{p(1-p)/n}$ .

Siccome  $p$  non è noto si prende il caso peggiore (variabilità massima) con  $p = 0.5$ .

Quindi

$$ME = 2.58 \cdot \sqrt{0.5^2/n}$$

da cui  $0.05^2 = (2.58^2)(0.5^2)/n$  e quindi

$$n = (2.58^2)(0.5^2)/(0.05^2) = 665.64.$$

## 8.20

Si intervista un campione casuale di 220 famiglie e si rileva che il 58.7% legge la pubblicità postale. Trovare l'intervallo di confidenza al 99% per la proporzione di destinatari che legge la pubblicità postale nella popolazione.

### Soluzione

Con 220 famiglie si usa l'approssimazione normale. Il quantile appropriato è  $z_{\alpha/2} = 2.58$  dove  $\alpha = 0.01$ . Quindi

$$ME = 2.58 \cdot \sqrt{0.587 * (1 - 0.587)/220} = 0.0856$$

Quindi l'intervallo è  $0.587 \pm 0.0856$  ovvero  $(0.501, 0.673)$ .

## 8.21

La quantità di stoffa usata per produrre poltrone è distribuita come una variabile casuale normale. Su un campione casuale di 15 poltrone, si è riscontrato che l'ammontare medio del materiale è 912 centimetri quadrati, con una deviazione standard di 64 centimetri quadrati. Quali dei seguenti intervalli rappresenta l'intervallo di confidenza al 99% per la media della quantità di materiale?

- A)  $912 \pm 44.3$
- B)  $912 \pm 42.6$
- C)  $912 \pm 49.2$
- D)  $912 \pm 46.8$

Campione di  $n = 15$ .  $\bar{X} = 912$  ed  $s = 64$  (deviazione standard campionaria).

L'intervallo è basato sulla  $t$  di Student con  $n - 1 = 14$  gradi di libertà. Quindi con  $\alpha = 0.01$  il quantile è  $t_{\alpha/2} = 2.977$ .

L'errore standard è  $ES = s/\sqrt{15} = 64/\sqrt{15} = 16.52473$ .

Quindi il margine di errore è

$$ME = 2.977 \cdot 16.52473 = 49.19 \simeq 49.2$$

Dunque la risposta è la C).

NOTA: l'errore tipico nei compiti è

$$ME = 2.58 \cdot 16.52473 = 42.6$$

dove 2.58 è il quantile della normale (che in questo caso è sbagliato).

## 8.22

Si supponga che il tempo trascorso dai clienti in un negozio sia distribuito in modo normale con media incognita e deviazione standard pari a 6 minuti. Si supponga di aver stimato il tempo medio della popolazione tramite un intervallo di confidenza al 95% e di aver ottenuto il seguente risultato: (22.06, 27.94). Qual è stata la dimensione del campione necessaria ad ottenere il precedente intervallo di confidenza?

**Soluzione** Come visto il margine di errore per un intervallo di livello 95% è  $ME = 1.96ES$  e qui

$$ME = \text{semiampiezza dell'intervallo} = 2.94.$$

e inoltre

$$ES = \sigma/\sqrt{n} = 6/\sqrt{n}$$

Quindi si imposta l'equazione

$$2.94 = 1.966/\sqrt{n}$$

e si risolve con

$$2.94^2 = (1.96^2 \cdot 6^2)/n$$

da cui

$$n = (1.96^2 \cdot 6^2)/(2.94^2) = 16.$$

## 9 Test delle ipotesi

- Introduzione ai test delle ipotesi
- Errore del I e II tipo
- test sulla media con varianza nota
- test sulla media con varianza incognita
- Test su una proporzione
- p-value
- Potenza del test

### 9.1

Se si rifiuta l'ipotesi nulla contro l'ipotesi alternativa ad un livello di significatività del 5% , allora, con gli stessi dati deve essere rifiutata anche ad un livello di significatività dell'1%. Vero o Falso?

**Soluzione**

Falso. Per esempio può capitare che in un test con la normale la statistica stia tra 1.96 e 2.56. In questo caso si rifiuta al 5% ma non all'1%.

### 9.2

Un idraulico afferma di poter completare l'installazione di un box doccia in meno di un'ora. Per un campione di 24 interventi, l'idraulico impiega una media di 63.2 minuti con una deviazione standard di 7.7 minuti. Qual è la statistica test osservata?

- A)  $t = 1.79$
- B)  $t = 2.04$
- C)  $Z = 2.04$
- D)  $Z = 1.79$

### Soluzione

Ipotesi:  $H_0 : \mu \leq 60$  contro  $H_1 : \mu > 60$ . Il parametro  $\mu_0 = 60$ . Abbiamo le statistiche  $\bar{X} = 63.2$  e  $s = 7.7$ , con  $n = 24$ .

La statistica test è

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{63.2 - 60}{7.7/\sqrt{24}} = 2.03594$$

che arrotondato a 2 decimali è  $t = 2.04$ . La risposta è B) e non C) perché la statistica è una t di Student con 23 gradi di libertà.

### 9.3

Il valore atteso della media della popolazione è dato dalla media campionaria. Vero o Falso?

#### Soluzione

Falso. La verità è che  $E(\bar{X}) = \mu$ . Qui si dice invece che  $E(\mu) = \bar{X}$  e questo non è assolutamente vero.

### 9.4

La ditta produttrice di un nuovo macchinario afferma che il suo macchinario incrementerà la produzione per macchina di almeno 29 unità di prodotto all'ora. Vengono acquistate 15 nuove macchine e si trova che l'incremento di produzione medio ottenuto è pari a 26 pezzi per macchina all'ora con una deviazione standard di 4.2. C'è evidenza empirica sufficiente per dubitare dell'affermazione fatta dal produttore dei nuovi macchinari considerando un livello di significatività  $\alpha = 0.05$ ?

#### Soluzione

Affermazione:  $\mu \geq 29$ . Dato che c'è l'uguale (almeno...) questa è  $H_0$ . Quindi

$$H_0 : \mu \geq 29 \text{ contro } H_1 : \mu < 29$$

Inoltre  $X \sim N(\mu=?, \sigma=?)$  entrambi incogniti.

Il test è unilaterale. Al livello del 5% si rifiuta se

$$t = (\bar{x} - \mu_0)/(s/\sqrt{n}) < -1.761$$

distribuita come una t di Student con  $n - 1 = 14$  gradi di libertà e probabilità a sinistra = 0.05.

Dati:  $n = 15$ ;  $\bar{x} = 26$ ,  $s = 4.2$ .

Quindi  $t = (26 - 29)/(4.2/\sqrt{15}) = -2.766$  e si rifiuta perché  $t < -1.761$ . Quindi la risposta è SÌ c'è evidenza empirica per dubitare al livello del 5%.

### 9.5

Un professore sostiene che il punteggio medio in un certo test è stato almeno 83. Si assuma che il punteggio al test si distribuisca normalmente. Tu ritieni che invece il punteggio medio sia inferiore ad 83, per cui decidi di chiedere ad un campione casuale di studenti il loro voto e risulta:

$$82, 77, 85, 76, 81, 91, 70, 82.$$

Verifica che la media e la varianza corretta sono: 80.5 e  $s^2 = 39.71429$ .

Ritieni sia lecito dubitare dell'affermazione del professore ad un livello di significatività del 5%?

### Soluzione

Si ha  $n = 8$ , la media è

$$\bar{x} = (82 + 77 + 85 + 76 + 81 + 91 + 70 + 82)/8 = 80.5.$$

La varianza campionaria corretta è

$$s^2 = ((82-80.5)^2 + (77-80.5)^2 + (85-80.5)^2 + (76-80.5)^2 + (81-80.5)^2 + (91-80.5)^2 + (70-80.5)^2 + (82-80.5)^2)/(8-1) = 39.71429$$

Quindi la deviazione standard è  $s = \sqrt{39.71429} = 6.3$ .

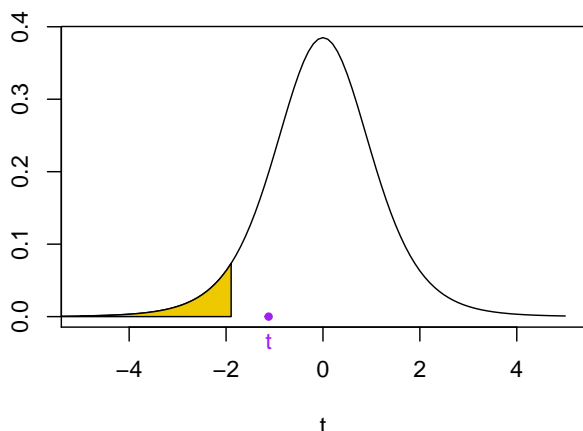
Il sistema di ipotesi è

$$H_0 : \mu \geq 83 \text{ contro } H_1 : \mu < 83$$

La regione critica unilaterale al livello del 5% è  $t < -1.895$  (quantile inferiore t di Student con 7 gradi di libertà al 5%)

$$\text{Statistica test} = (80.5 - 83)/(6.3/\sqrt{8}) = -1.122.$$

Quindi non si rifiuta al livello del 5% perché  $-1.122 > -1.895$ .



Non si dubita dell'affermazione al livello del 5%.

## 9.6

Supponi di voler effettuare con un livello di significatività  $\alpha = 0.10$  il seguente test sulla media di una popolazione:

$$H_0 : \mu = 277 \text{ contro } H_1 : \mu \neq 277.$$

Supponi inoltre di sapere che la deviazione standard della popolazione è  $\sigma = 13.5$ . Se selezioni un campione casuale di 20 osservazioni, per quale valore della media campionaria rifiuterai l'ipotesi nulla?

### Soluzione

È un test per la media di una normale con deviazione standard nota = 13.5. Campione di dimensione  $n = 20$ . L'errore standard è

$$ES = \sigma/\sqrt{n} = 13.5/\sqrt{20} = 3.018692.$$

Sia  $z^* = 1.645$  il valore tale che  $P(Z > z^*) = 0.05$  (vedi le tavole della t di Student, ultima riga).

Si rifiuta se

$$(\bar{X} - \mu_0)/ES > 1.645 \quad \text{oppure} \quad (\bar{X} - \mu_0)/ES < -1.645$$

cioè se

$$\bar{X} > \mu_0 + 1.645ES \quad \text{oppure} \quad \bar{X} < \mu_0 - 1.645ES$$

ossia se

$$\bar{X} > 277 + 1.645(3.018692) = 281.9657$$

oppure

$$\bar{X} < 277 - 1.645(3.018692) = 272.0343$$

## 9.7

Si verifica un errore del I tipo quando viene rifiutata un'ipotesi nulla vera. Vero o Falso?

Vero. È la definizione: errore del I tipo = rifiutare  $H_0$  quando  $H_0$  è vera. Si tratta di sbagliare gli innocenti per colpevoli.

## 9.8

Un'azienda produttrice di caffè asserisce che ciascun lotto contiene almeno 50.1 kg di prodotto. Si assuma che la deviazione standard della quantità di caffè contenuta in ciascun lotto sia 1.2 kg. La regola di decisione adottata dall'azienda è di fermare le consegne se la media campionaria della quantità di caffè in un campione di 40 lotti è inferiore a 49.7. Qual è la probabilità di commettere un errore del primo tipo?

### Soluzione

Ipotesi nulla  $H_0 : \mu \geq 50.1$ , alternativa  $H_1 : \mu < 50.1$ .

Siccome il testo dice "si assuma" vuol dire che  $1.2 = \sigma$  la deviazione standard della popolazione. Si usano campioni di numerosità  $n = 40$ .

La regola di decisione definisce la regione critica: ferma le consegne se si rifiuta  $H_0$  cioè se  $\bar{X} < 49.7$ .

$$P(I) = P(\text{Rifiutare } H_0, \text{ quando } \mu = 50.1) = P(\bar{X} < 49.7 \text{ quando } \mu = 50.1)$$

Quindi poiché sotto  $H_0$   $\bar{X} \sim N(\mu = 50.1, \sigma = 1.2/\sqrt{40} = 0.1897367)$  risulta

$$P(I) = P(\bar{X} < 49.7) = P(Z < (49.7 - 50.1)/0.1897367) = P(Z < -2.11) = 1 - P(Z < 2.11) = 0.0174.$$

Cioè il livello del test è 0.0174.

## 9.9

Associa al simbolo  $\beta$  la definizione opportuna.

- A) La potenza del test.
- B) La probabilità dell'errore di II tipo.
- C) La probabilità dell'errore di I tipo.
- D) La probabilità di rifiutare  $H_0$ .

### Soluzione

È la probabilità di errore del II tipo cioè di accettare  $H_0$  quando è falsa. Cioè la probabilità di scambiare un colpevole per innocente. Quindi la risposta giusta è B).

### 9.10

Aumentando il livello di significatività di un test, la probabilità dell'errore del II tipo aumenta. Vero o Falso?

#### Soluzione

Falso. Infatti il livello del test è  $\alpha$  e la probabilità di errore di II tipo è  $\beta$ . Ma è noto che se  $\alpha$  aumenta  $\beta$  diminuisce.

### 9.11

Quale delle seguenti frasi NON è vera?

- A) La regione di rifiuto è l'insieme di tutti i valori della statistica test per cui l'ipotesi alternativa viene rifiutata.
- B) Una statistica test è una funzione dei dati campionari sulla base della quale si decide se rifiutare o meno l'ipotesi nulla.
- C) La regione di rifiuto è l'insieme di tutti i valori della statistica test per cui l'ipotesi nulla viene rifiutata.
- D) Una buona procedura di test delle ipotesi deve comportare una probabilità dell'errore del I tipo e del II tipo piccola.

#### Soluzione

- A) è falsa: La regione di rifiuto è l'insieme di tutti i valori della statistica test per cui l'ipotesi NULLA viene rifiutata.
- B) è vera
- C) è vera (vedi sopra)
- D) Ovviamente è vera.

Quindi la risposta è A)

### 9.12

Quale deve essere la dimensione del campione necessaria per stimare la media di una popolazione distribuita normalmente se  $ME = 5$ ,  $\sigma = 40$ , livello di confidenza = 99% ?

#### Soluzione

Deve essere  $ME = 2.58\sigma/\sqrt{n}$  quindi

$$5 = 2.58 * 40/\sqrt{n}$$

ossia

$$25 = (2.58^2)(40^2)/n$$

che risolto dà

$$n = (2.58^2)(40^2)/25 = 426.$$

### 9.13

Il livello di significatività di un test è la probabilità che l'ipotesi nulla sia vera. Vero o Falso?

#### Soluzione

FALSO. Il livello è la probabilità di rifiutare  $H_0$  quando è vera non la probabilità che  $H_0$  sia vera.

## 9.14

Si supponga di voler effettuare un test su  $H_0 : \mu \geq 0.54$  contro  $H_1 : \mu < 0.54$  basato su un campione iid di  $n = 25$  da  $N(\mu, \sigma^2)$  sapendo che nel campione  $s = 13.2$ . Quale dovrebbe essere la statistica test? A)  $(\bar{X} - 0.54)/2.64$  C)  $(\bar{X} - 0.54)/0.528$  B)  $(\bar{X} - 0.54)/34.848$  D)  $(\bar{X} - 0.54)/0.2789$

### Soluzione

Nel test t di Student per la media la statistica test è

$$t = (\bar{X} - \mu_0)/ES$$

dove  $ES = s/\sqrt{n}$ . Qui  $\mu_0 = 0.54$  e  $ES = 13.2/5 = 2.64$ .

Quindi la soluzione è A).

## 9.15

Un professore asserisce che il punteggio medio conseguito ad un recente esame è stato 83. Si assuma che la variabile punteggio conseguito si distribuisca normalmente. Tu chiedi ad alcuni in classe quale punteggio abbiano conseguito ed ottieni le seguenti risposte: 82, 77, 85, 76, 81, 91, 70 e 82. Supponi di voler verificare se l'affermazione del professore è corretta contro un'alternativa bilaterale.

Quale affermazione tra le seguenti è più appropriata per il p-value?

- A) p-value < 0.10
- B) p-value < 0.01
- C) p-value < 0.05
- D) p-value > 0.10

### Soluzione

Il sistema di ipotesi è  $H_0 : \mu = 83$  contro  $H_1 : \mu \neq 83$ . La dimensione campionaria è  $n = 8$  e la media campionaria è 80.5 e la deviazione standard è  $s = 6.3$ . Si usa la statistica t di Student (dato che la varianza della popolazione incognita è stimata dai dati) è

$$t = (80.5 - 83)/(6.3/\sqrt{8}) = -1.12$$

(distribuzione t di Student con 7 gradi di libertà sotto  $H_0$ ).

Il p-value è la probabilità che la statistica test assuma un valore più estremo di quello osservato sotto  $H_0$ . Cioè è

$$p = P(T < -1.12) + P(T > 1.12) = 1 - P(-1.12 < T < 1.12).$$

Questa probabilità non si può calcolare esattamente dalle tavole della t. Tuttavia si nota che le regioni critiche classiche sono

al livello dell'1%  $t < -3.499, t > +3.499$  al livello dell'5%  $t < -2.365, t > 2.365$  al livello del 10%  $t < -1.895, t > 1.895$

quindi la probabilità di  $t < -1.12$  o  $t > 1.12$  deve essere per forza maggiore del 10%. Conclusione: la risposta giusta è la D).

## 9.16

Supponiamo di voler effettuare una procedura di verifica delle ipotesi su una proporzione  $p$ , e che la proporzione campionaria  $\hat{p}$  sia approssimativamente normale. Se l'ipotesi alternativa è  $H_1 : p \neq p_0$ , allora la regione di rifiuto al livello  $\alpha = 0.05$  è  $Z < -1.96$  o  $Z > 1.96$ . Vero o Falso?

**Soluzione** La regione critica di livello  $\alpha = 0.05$  è  $Z > z_{\alpha/2}$  o  $Z < -z_{\alpha/2}$ . Dove  $Z = (\hat{p} - p_0)/ES(\hat{p})$ . Quindi se  $\alpha = 0.05$ ,  $z_{\alpha/2} = 1.96$ . Quindi è vero.

### 9.17

Associa al simbolo  $1 - \beta$  la definizione opportuna.

- A) La probabilità di rifiutare  $H_0$ .
- B) La probabilità dell'errore di I tipo.
- C) La probabilità di corretto rifiuto di  $H_0$ .
- D) La probabilità dell'errore di II tipo.

**Soluzione** Poiché  $\beta$  è la probabilità P(II) di errore di secondo tipo cioè la probabilità di accettare  $H_0$  quando è falsa,  $1 - \beta$  è la probabilità di rifiutare  $H_0$  quando è falsa, cioè la potenza del test. Soluzione C).

### 9.18

La Regione Veneto ha dichiarato che il reddito medio familiare annuo della regione è superiore a 37000 Euro. Si assuma che la distribuzione del reddito medio familiare della Regione Veneto sia distribuito come una variabile normale con deviazione standard di 5756 euro. Si supponga che in un campione di 25 famiglie si sia rilevato un reddito medio annuo pari a 36243 euro. Quale affermazione tra le seguenti è più appropriata per il p-value?

- A) p-value < 0.01
- B) p-value < 0.10
- C) p-value < 0.05
- D) p-value > 0.10

#### Soluzione

Il problema suppone che il reddito  $X \sim N(\mu = ?, \sigma = 5756)$ . L'ipotesi della Regione Veneto è che  $\mu = E(X) = E(\text{reddito annuo}) > 37000$  Euro. Poiché non è specificata un'uguaglianza questa è l'ipotesi alternativa. L'ipotesi da verificare è

$$H_0 : \mu \leq 37000 \quad \text{contro} \quad H_1 : \mu > 37000 \quad (\text{l'ipotesi della Regione})$$

La statistica test è (con  $n = 25$ )

$$z = (36243 - 37000) / (5756 / \sqrt{25}) = -0.657.$$

Le regioni critiche classiche sono

- 1%  $z < -2.58$  o  $z > 2.58$
- 5%  $z < -1.96$  o  $z > 1.96$
- 10%  $z < -1.64$  o  $z > 1.64$

Quindi il p-value è sicuramente  $> 0.10$ . Infatti  $z = -0.657$  porterebbe ad accettare  $H_0$  al livello del 10%.

### 9.19

Un'ipotesi nulla è rifiutata a livello di significatività 0.025, ma non ad un livello di 0.01. Ciò significa che il p-value del test è compreso tra 0.01 e 0.025. Vero o Falso?

**Soluzione** Se un'ipotesi è rifiutata al livello del 2.5% vuol dire che il p-value è  $<$  del 2.5%.

Se un'ipotesi non è rifiutata al livello dell' 1% vuol dire che il p-value è  $>$  dell' 1%.

Il p-value  $p$  è  $0.01 < p < 0.025$ . Quindi è vero.

## 9.20

Un commercialista afferma di poter completare una dichiarazione dei redditi standard in meno di un'ora. Per un campione di 24 dichiarazioni, il commercialista impiega una media di 63.2 minuti con una deviazione standard di 7.7 minuti. Quale affermazione tra le seguenti è più appropriata per il p-value?

- A)  $0.025 < \text{p-value} < 0.05$
- B)  $\text{p-value} < 0.025$
- C)  $\text{p-value} > 0.05$
- D)  $\text{p-value} < 0.01$

### Soluzione

Il tempo  $X$  impiegato ha distribuzione  $N(\mu = ?, \sigma = ?)$ . L'affermazione è  $\mu < 60$  min e quindi è l'ipotesi alternativa. L'ipotesi nulla (che contiene il segno di uguaglianza) è  $H_0 : \mu \geq 60$ .

Abbiamo un campione di dimensione  $n = 24$ ,  $\bar{x} = 63.2$ ,  $s = 7.7$ .

Quindi la statistica test è  $t$  di Student (con 23 gradi di libertà):

$$t = (63.2 - 60) / (7.7 / \sqrt{24}) = 2.034.$$

Le regioni critiche standard sono unilaterali sinistre:

- 1%  $t < -2.500$
- 2.5%  $t < -2.069$
- 5%  $t < -1.714$
- 10%  $t < -1.319$

Quindi  $t = 2.034 > -1.319$  e quindi il p-value deve essere  $> 10\%$  e quindi anche maggiore del 5%. La risposta quindi è C).

## 9.21

Aumentando il livello di significatività di un test, la probabilità dell'errore del II tipo aumenta. Vero o Falso?

**Soluzione** Se il livello  $\alpha$  del test aumenta la probabilità di errore di secondo tipo  $\beta$  diminuisce a parità di altri elementi. Quindi è falso.

## 9.22

Associa al simbolo  $1 - \alpha$  la definizione opportuna.

- A) La probabilità dell'errore di II tipo.
- B) La probabilità dell'errore di I tipo.
- C) La probabilità di non rifiutare l'ipotesi nulla quando questa è vera.
- D) La potenza del test.

### Soluzione

$\alpha$  è la  $P(I) =$  probabilità rifiutare  $H_0$  quando  $H_0$  è vera. Quindi  $1 - \alpha$  è la probabilità di accettare  $H_0$  quando è vera. Quindi la risposta è C).

## 9.23

Avete un campione da una normale la cui media può essere 10 o 12 (non si sa quale delle due) e deviazione standard 2. Con un campione di 4 elementi dalla popolazione dovete verificare  $H_0 : \mu = 10$  contro  $H_1 : \mu = 12$  e pensate di rifiutare  $H_0$  quando la media campionaria è maggiore di 11.5.

- Qual è la probabilità di errore del I tipo?
- Qual è la probabilità di corretto rifiuto (ossia la potenza del test)?

**Soluzione**

La  $P(I)$  = probabilità che la media campionaria sia  $> 11.5$  quando  $H_0$  è vera cioè se  $\mu = 10$ .

La potenza = probabilità che la media campionaria sia  $> 11.5$  quando  $H_0$  è falsa cioè se  $\mu = 12$ .

Quindi si possono calcolare sapendo che la media campionaria è

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n} = 2/2 = 1).$$

Abbiamo

$$P(I) = P(\bar{X} > 11.5, \bar{X} \sim N(10, 1)) = P(Z > (11.5 - 10)) = P(Z > 1.5) = 1 - P(Z < 1.5) = 0.0668.$$

Inoltre

$$\text{Potenza} = P(\bar{X} > 11.5, \bar{X} \sim N(12, 1)) = P(Z > (11.5 - 12)) = P(Z > -0.5) = P(Z < 0.5) = 0.6915.$$

**9.24**

Associa al simbolo  $\beta$  la definizione opportuna:

- A) La probabilità dell'errore di I tipo.
- B) La probabilità di non rifiutare  $H_0$  vera.
- C) La probabilità di accettare un'ipotesi nulla falsa.
- D) La potenza del test.

**Soluzione**

È la probabilità di errore di II tipo e cioè la probabilità di accettare un'ipotesi nulla falsa. Risposta C).

**9.25**

L'errore del II tipo può essere definito come:

- A) Rifiutare un'ipotesi alternativa vera.
- B) Non rifiutare un'ipotesi alternativa falsa.
- C) Non rifiutare un'ipotesi nulla falsa.
- D) Rifiutare un'ipotesi nulla vera.

**Soluzione**

La definizione di errore del II tipo è accettare (ossia non rifiutare)  $H_0$  quando è falsa. Risposta C). È l'errore di scambiare un colpevole con un innocente.

**9.26**

L'azienda produttrice di sacchi di farina afferma che ciascun sacco contiene almeno 50.1 kg di farina. Si assuma che la deviazione standard della quantità di farina contenuta in ciascun sacco sia 1.21 kg. La regola di decisione adottata dall'azienda è di mettere in manutenzione una macchina riempitrice se la media campionaria della quantità di farina in un campione di 40 sacchi è inferiore a 49.7. Qual è la probabilità di commettere un errore del primo tipo?

**Soluzione**

- $X$  = quantità di farina nel sacco = aleatoria  $\sim N(\mu = ?, \sigma = 1.21)$ .

- Test:  $H_0 : \mu \geq 50.1$ ,  $H_1 : \mu < 50.1$
- Regione critica:  $\bar{X}$  = media di 40 sacchi (campione casuale)  $< 49.7$
- $P(I) = P(\bar{X} < 49.7 \text{ quando } \mu = 50.1)$

Questa si calcola sapendo che se  $\mu = 50.1$  allora

$$\bar{X} \sim N(50.1, \sigma_{\bar{X}} = 1.21/\sqrt{40} = 0.1913178)$$

Quindi

$$P(I) = P(Z < (49.7 - 50.1)/0.1913178) = P(Z < -2.09) = 1 - P(Z < 2.09) = 0.02.$$

## 9.27

Un produttore di lenti per occhiali sostiene che almeno l'80% degli oculisti preferisce il suo tipo di lenti per occhiali. Decidi di verificare la sua affermazione e, su un campione di 200 oculisti, trovi che il 74.1% preferisce quelle lenti. C'è sufficiente evidenza per dubitare dell'affermazione del produttore? Usa il livello  $\alpha = 0.025$ .

### Soluzione

L'ipotesi nulla è  $H_0 : p \geq 0.8$  contro  $H_1 : p < 0.8$ . In un campione di dimensione  $n = 200$ ,  $\hat{p} = 0.741$ . La statistica test è

$$z = (0.741 - 0.8)/ES$$

con

$$ES = \sqrt{(p_0(1 - p_0)/n)} = \sqrt{(0.8(1 - 0.8)/200)} = 0.02828427$$

NOTA: l'errore standard viene calcolato SOTTO  $H_0$ . Pertanto  $z = -2.085$  che sotto  $H_0$  ha una distribuzione approssimata normale (la binomiale è approssimata bene da una normale se  $n = 200$ ).

La regione critica unilaterale al livello 0.025 è (i valori della normale sono all'ultima riga della tavola della t di Student)  $z < -1.96$ .

Perciò con un valore osservato di  $z = -2.085$  si rifiuta  $H_0$ . Quindi c'è sufficiente evidenza per dubitare dell'affermazione al livello del 2.5%.