INFERENZA STATISTICA

La Statistica Inferenziale studia come estendere i risultati e le conclusioni che provengono dall'analisi di un campione di osservazioni alla popolazione a cui il campione appartiene.

Vi sono molteplici motivi per ricorrere ad una indagine campionaria:

- costi di un'indagine censuaria;
- tempi prolungati di un'indagine censuaria;
- l'osservazione è distruttiva;
- in moltissimi casi, la popolazione di riferimento è **virtuale** e infinita.

IL CAMPIONAMENTO

Si intuisce che il campione deve essere rappresentativo della popolazione di riferimento e quindi libero da qualsiasi elemento soggettivo.

Esistono vari metodi di campionamento (*Tecniche di Campionamento*), ma noi considereremo solo il **campionamento casuale**: le unità statistiche che entrano a far parte del campione sono estratte in modo casuale dalla popolazione di riferimento.
Se le estrazioni delle unità statistiche che entrano nel campione sono indipendenti si parla di **campione casuale semplice**. Affinché le estrazioni siano indipendenti

- se la popolazione è finita le estrazioni devono essere con reinserimento
- se la popolazione è infinita oppure se la popolazione è molto più ampia della numerosità del campione, le estrazioni possono essere con o **senza** reinserimento

Nel seguito ci limiteremo a considerare campioni casuali semplici, che verranno abbreviati con **c.c.s**.

RAPPRESENTAZIONE PROBABILISTICA DI UN CAMPIONE CASUALE SEMPLICE

Supponiamo di estrarre a caso una unità dalla popolazione di riferimento e di osservare su di essa la caratteristica di interesse.
Sia x il valore osservato sull'unità.
A priori rispetto all'esperimento, ossia prima dell'estrazione dell'unità e della rilevazione della caratteristica, non siamo in grado di prevedere quale valore osserveremo. Pertanto, x è la realizzazione di una variabile casuale X la cui distribuzione di probabilità descrive quali sono i valori che possono essere assunti dalla caratteristica su un'unità estratta a caso e con quale probabilità (se X è discreta) o densità (se X è continua) verranno osservati.

Generalizziamo il ragionamento da una sola unità a n unità, per ottenere un c.c.s. di dimensione n.

Se $x_1, x_2, \ldots, x_n$ sono le n osservazioni numeriche raccolte sulle n unità estratte, ciascuna è la realizzazione di una variabile casuale. Più precisamente, x_1 è la realizzazione della variabile casuale X_1 che descrive il risultato che osserveremo sulla prima unità estratta, x_2 è la realizzazione della variabile casuale X_2 che descrive il risultato che osserveremo sulla seconda unità estratta e così via.

Trattandosi di repliche della variabile X queste n variabili sono identicamente distribuite, tutte con la stessa distribuzione di X. Inoltre, essendo le estrazioni indipendenti, le variabili stesse sono indipendenti.

In sintesi,

- **PRIMA** di effettuare l'osservazione, un c.c.s. può essere rappresentato da una n–pla di variabili casuali, X_1, X_2, ..., X_n, **i.i.d.**;
- **DOPO** aver effettuato l'osservazione, un c.c.s. è rappresentabile da una n–pla di valori, x_1, x_2, ..., x_n, ciascuno realizzazione di una variabile casuale.

Esempio 1
Supponiamo di voler studiare l'altezza della popolazione italiana. Estraiamo un c.c.s. di dimensione 8, osservando le seguenti altezze (in cm):

$$x_1 = 170, x_2 = 181, x_3 = 189, x_4 = 160,$$

$$x_5 = 182, x_6 = 171, x_7 = 158, x_8 = 186$$

Questi valori sono realizzazioni di altrettante variabili casuali, $X_1, \ldots, X_8$, i.i.d. che descrivono il c.c.s. **prima** di fare l'osservazione.
La distribuzione di caratteristiche antropometriche è spesso approssimabile con una distribuzione gaussiana. È quindi ragionevole assumere che $X \sim N(\mu, \sigma^2)$, dove X è la variabile casuale che descrive l'osservazione su una generica unità estratta a caso. Analogamente, $X_1, \ldots, X_n$ sono i.i.d. con distribuzione $N(\mu, \sigma^2)$.

In questo esempio, $\mu = E(X) = E(X_i)$ è l'**ignota** altezza media della popolazione italiana e $\sigma^2 = V(X) = V(X_i)$ è l'**ignota** varianza dell'altezza nella popolazione italiana. Quindi, entrambi i parametri della normale, μ e σ^2, rappresentano delle quantità di interesse della popolazione di riferimento e sono **ignoti**.
Potremmo pensare di "stimare" le ignote quantità di popolazione, μ e σ^2, con le corrispondenti quantità campionarie:

$$\bar{x} = 175 \text{ cm} \qquad \text{e} \qquad s^2 = 10,85^2$$

Tuttavia, se ripetessimo l'esperimento ed estraessimo un nuovo c.c.s. di 8 individui otterremmo, in generale, altezze diverse dalle precedenti e quindi stime diverse delle quantità di popolazione. Ci domandiamo, pertanto,

- quanto sono accurate le stime della media e della varianza dell'altezza della popolazione italiana?
- possiamo individuare un intervallo di valori ragionevoli per μ e per σ^2?
- in base ai risultati ottenuti, è ragionevole affermare, ad esempio, che l'altezza media della popolazione italiana è superiore a 170 cm?

Esempio 2
Un'industria che produce pompe idrauliche acquista guarnizioni in plastica da un fornitore. Per garantire la qualità delle pompe prodotte, l'industria ha la necessità di verificare quale è la percentuale di guarnizioni acquistate che sono difettose. A questo fine, si estraggono casualmente 20 guarnizioni tra quelle acquistate e per ciascuna si

verifica la conformità alle specifiche, ottenendo i seguenti risultati

$$x_1 = 0, x_2 = 0, x_3 = 0, x_4 = 0, x_5 = 0, x_6 = 1, x_7 = 0,$$

$$x_8 = 0, x_9 = 0, x_{10} = 0, x_{11} = 0, x_{12} = 0, x_{13} = 0, x_{14} = 0,$$

$$x_{15} = 1, x_{16} = 0, x_{17} = 0, x_{18} = 0, x_{19} = 0, x_{20} = 0$$

dove 0= "non difettosa", 1= "difettosa".
Si noti che la popolazione di riferimento è costituita da tutte le guarnizioni acquistate, presenti e future (popolazione "virtuale").
Prima di estrarre il campione non siamo in grado di prevedere i risultati, per cui ciascuna x_i può essere vista come realizzazione di una variabile casuale X_i, con $X_1, X_2, \ldots, X_{20}$ i.i.d.. Poiché il risultato è dicotomico (difettosa/non difettosa), è naturale assumere $X_i \sim Be(\pi)$. In questo esempio, $\pi = P(X_i = 1)$ rappresenta la frazione di guarnizioni difettose nell'intera popolazione (virtuale) di riferimento, ed è **ignota**. Come nell'Esempio 1, anche in questo contesto il parametro della distribuzione delle X_i è una quantità di interesse della popolazione: è da π, che è possibile giudicare la qualità delle guarnizioni acquistate.

Sembra ragionevole stimare l'ignoto π tramite la corrispondente frazione campionaria di pezzi difettosi, ossia tramite

$$p = \frac{\sum_{i=1}^{20} x_i}{20} = \frac{2}{20} = 0,1$$

Tuttavia, se ripetessimo il campionamento, estraendo un nuovo c.c.s. di dimensione 20, potremmo ottenere un

numero differente di pezzi difettosi e quindi una stima differente. Le domande che ci poniamo sono allora

- Quanto accurata è la nostra valutazione della **vera, ma ignota** frazione di guarnizioni difettose acquistate?
- Possiamo individuare un intervallo di valori ragionevoli per π?
- Se l'industria fornitrice ha garantito una percentuale di guarnizioni difettose non superiore a 5%, mentre il campione ne contiene 10%, ci sono elementi per contestare la fornitura e chiedere il risarcimento?

Esempio 3

Per valutare l'uso di uno sportello bancomat, è stato rilevato il numero di utilizzi su un campione di 10 giorni scelti casualmente, ottenendo i seguenti valori

$$x_1 = 10, x_2 = 6, x_3 = 5, x_4 = 7, x_5 = 7,$$

$$x_6 = 6, x_7 = 7, x_8 = 8, x_9 = 11, x_{10} = 9$$

Si noti che, come nell'Esempio 2, la popolazione di riferimento è infinita e virtuale, essendo costituita da tutti i giorni passati, presenti e futuri. Se gli utilizzi tra i vari giorni sono tra loro indipendenti e non vi è stato nessun cambiamento sistematico durante il periodo della rilevazione, le variabili casuali $X_1, \ldots, X_{10}$, di cui le 10 osservazioni campionarie sono realizzazioni, possono essere assunte i.i.d.. In aggiunta, trattandosi di conteggi, potrebbe essere ragionevole assumere $X_i \sim Po(\lambda)$. In

questo esempio, $\lambda = E(X_i)$ rappresenta il numero medio giornaliero di utilizzi del bancomat nella popolazione di interesse ed è **ignoto**. Anche in questo caso, il parametro della distribuzione delle X_i è una quantità di interesse della popolazione: tramite λ è possibile valutare l'uso del bancomat. È ragionevole stimare l'ignoto λ attraverso la corrispondente media campionaria $\bar{x} = 7,6$. Siamo, però, consapevoli che se prendessimo un nuovo campione di 10 giorni otterremmo valori diversi e quindi una media diversa. Pertanto, valgono le usuali considerazioni:

- quanto affidabile è la nostra stima?
- possiamo identificare un intervallo di valori ragionevoli per λ?
- i dati campionari danno sostegno, ad esempio, all'ipotesi che il numero medio giornaliero di utilizzi sia inferiore a 10?

I FONDAMENTI DELL'INFERENZA STATISTICA (PARAMETRICA)

I tre esempi visti hanno in comune alcune caratteristiche che costituiscono i fondamenti dell'Inferenza Statistica (parametrica).

Dato un c.c.s. di osservazioni $x_1, x_2, \ldots, x_n$, ogni x_i è realizzazione di una variabile casuale X_i. Le variabili $X_1, X_2, \ldots, X_n$ sono i.i.d. con una funzione di probabilità (se le X_i sono discrete) o di densità (se le X_i sono continue) che indicheremo in modo generico con $f(x)$. La funzione $f(x)$ dipende da uno o più parametri **ignoti** che rappresentano caratteristiche di interesse della popolazione di riferimento.

- Nell'Esempio 1, $X_i \sim N(\mu, \sigma^2)$ e i parametri ignoti sono μ e σ^2, che rappresentano, rispettivamente, la media e la varianza dell'altezza della popolazione italiana:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \quad x \in \mathbb{R}$$

- Nell'Esempio 2, $X_i \sim Be(\pi)$ e il parametro ignoto è π che rappresenta la frazione di guarnizioni difettose nella popolazione di tutte le guarnizioni acquistate:

$$f(x) = \pi^x(1-\pi)^{1-x}, \quad x = 0 \text{ oppure } 1$$

- Nell'Esempio 3, $X_i \sim Po(\lambda)$ e il parametro ignoto è λ che rappresenta il numero medio giornaliero di

utilizzi del bancomat nella popolazione costituita da tutti i giorni:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

Obiettivo dell'Inferenza Statistica è utilizzare le osservazioni campionarie per risalire al vero valore dei parametri ignoti e da qui alla distribuzione di probabilità del fenomeno di interesse. Ad esempio, nell'Esempio 2, variando $\pi \in (0, 1)$ si genera una famiglia di distribuzioni bernoulliane di parametro π. Attraverso i dati, si vuole identificare il vero valore di π e da questo la vera distribuzione di probabilità delle X_i.

Le domande a cui l'Inferenza Statistica cerca di dare una risposta rientrano in 3 principali categorie

1. Si vuole stimare il vero valore dei parametri ignoti in base alle osservazioni del campione e capire quanto accurata è la stima proposta (**STIMA PUNTUALE**).

2. Si vuole identificare un insieme di valori ragionevoli per i parametri ignoti (**STIMA INTERVALLARE**).

3. Si formula un'ipotesi sul vero valore dei parametri ignoti e si vuole verificare se tale ipotesi è vera oppure no, in base alle osservazioni capionarie. (**VERIFICA DI IPOTESI**).

LE STATISTICHE CAMPIONARIE: IL CASO DELLA MEDIA CAMPIONARIA

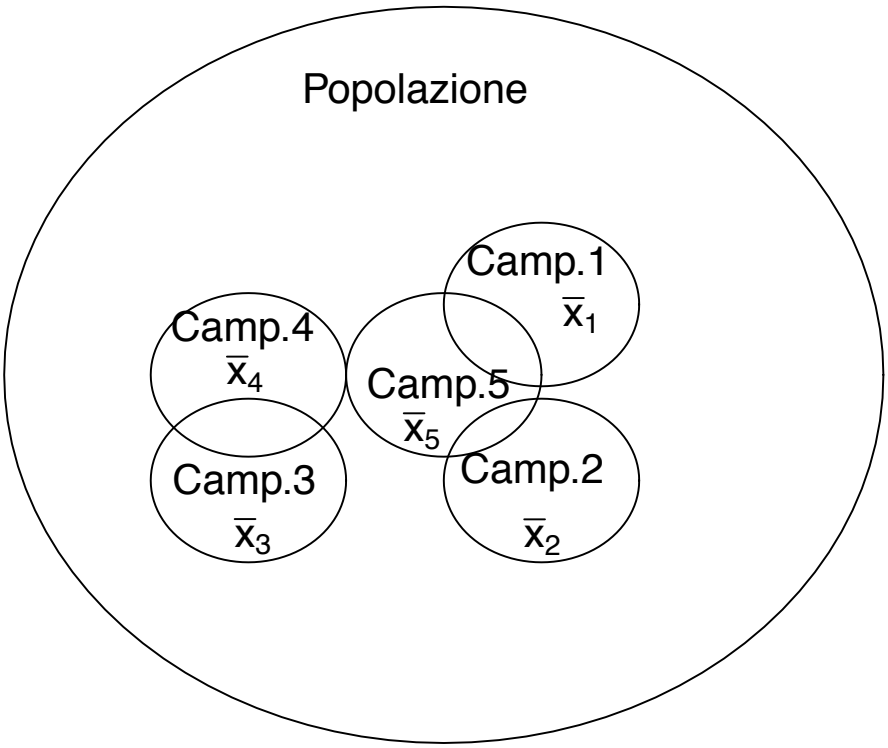
Prendiamo un c.c.s. $x_1, x_2, \ldots, x_n$, realizzazione della n-pla di variabili casuali X_1, X_2, $\ldots$, X_n, i.i.d.. Supponiamo di voler fare inferenza sulla media ignota di popolazione (o, in generale, del fenomeno di interesse) che indichiamo con $m = E(X_i)$.
Una scelta naturale è stimare m tramite la corrispondente media campionaria:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Tuttavia, ogni c.c.s. di dimensione n produce "potenzialmente" un valore diverso di $\bar{x}$. Poiché le osservazioni x_i sono realizzazioni delle variabili casuali X_i, la stessa media $\bar{x}$ può essere vista come la realizzazione della variabile casuale

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

La distribuzione di probabilità di $\bar{X}$, chiamata **distribuzione campionaria**, rappresenta la distribuzione delle medie campionarie di tutti i c.c.s. di dimensione n estraibili dalla popolazione di riferimento. In altre parole, la distribuzione di $\bar{X}$ determina il modo in cui la media campionaria $\bar{x}$ varia passando da un c.c.s. di dimensione n all'altro.



Esempio 4
Supponiamo che la nostra popolazione di riferimento sia un'urna con 3 palline numerate da 1 a 3 e indichiamo con X la variabile casuale che descrive il valore di una pallina estratta casualmente dall'urna.

Prendiamo $n = 2$. Tutti i possibili c.c.s. di dimensione 2 sono (estrazioni con reinserimento): (1,2), (2,1), (1,3), (3,1), (2,3), (3,2), (1,1), (2,2), (3,3).

Poniamo X_1=valore della prima pallina estratta, X_2=valore della seconda pallina estratta e

$$\bar{X} = (X_1 + X_2)/2$$

La distribuzione campionaria di $\bar{X}$ è

valori di $\bar{X}$	1	1,5	2	2,5	3	Tot
probabilità	1/9	2/9	3/9	2/9	1/9	1

Ad esempio, la probabilità di osservare un c.c.s. tale che $\bar{x} = 2$ è 3/9 (ossia la probabilità di estrarre (2,2) o (1,3) o (3,1)).

N.B.: La media di popolazione è $m = E(X_1) = E(X_2) = 2$.

Esempio 1
Riprendiamo l'Esempio 1 e supponiamo che $\mu = 170$ cm e $\sigma^2 = 15^2$ (ma media e varianza di popolazione sono in generale ingnote). Sappiamo che $X_1, \ldots, X_8$ sono i.i.d. con distribuzione $N(\mu = 170; \sigma^2 = 15^2)$. Per noti risultati della probabilità, la distribuzione campionaria di $\bar{X}$ è (si confronti con pag. 101 della Probabilità)

$$\bar{X} = \frac{1}{8}\sum_{i=1}^{8} X_i \sim N\left(170, \frac{15^2}{8}\right)$$

Pertanto, la probabilità, per esempio, di osservare un campione per il quale $\bar{x} < 175$ è

$$P(\bar{X} < 175) = \Phi\left(\frac{175 - 170}{15/\sqrt{8}}\right) = 0,631$$

LE STATISTICHE CAMPIONARIE

- Quanto visto per $\bar{X}$ può essere esteso a qualunque altra funzione di $X_1, \ldots, X_n$.
- In generale, si definisce **STATISTICA CAMPIONARIA** una qualsiasi funzione $T = g(X_1, X_2, \ldots, X_n)$ delle n variabili casuali $X_1, \ldots, X_n$ che generano i dati.
- Essendo funzione di variabili casuali, T stessa è una variabile casuale dotata di una propria distribuzione di probabilità, chiamata **distribuzione campionaria**.
- Sul campione effettivamente osservato $x_1, x_2, \ldots, x_n$, T avrà come realizzazione il valore numerico $t = g(x_1, x_2, \ldots, x_n)$ (si pensi alla relazione tra la variabile casuale $\bar{X}$ e la media numerica $\bar{x}$).
- La media campionaria è un esempio di statistica campionaria. Un altro esempio molto importante di statistica campionaria è la varianza campionaria

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

la cui realizzazione sul c.c.s. effettivamente osservato è

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

- **Esempio 4**
  Riprendiamo l'esempio di una popolazione costituita da un'urna con 3 palline numerate da 1 a 3.

  Ricordiamo che tutti i possibili c.c.s di dimensione 2 sono: (1,2), (2,1), (1,3), (3,1), (2,3), (3,2), (1,1), (2,2), (3,3).

  La distribuzione campionaria della varianza campionaria

$$S^2 = \frac{1}{2}\left\{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2\right\}$$

  è

valori di S^2	0	0,25	1	Tot
probabilità	3/9	4/9	2/9	1

  Ad esempio, la probabilità di estrarre un c.c.s tale che $s^2 = 0$ è 3/9 (è la probabilità di estrarre o (1,1) o (2,2) o (3,3)).

LA STIMA PUNTUALE

Sulla base di un c.c.s. di osservazioni, l'obiettivo è identificare il vero valore del parametro ignoto in modo puntuale, ossia attraverso un unico valore.

In modo formale, il problema della stima puntuale può ricondursi alla scelta di una statistica campionaria $t = g(x_1, \ldots, x_n)$ che quantifichi il vero valore del parametro ignoto. La statistica campionaria t usata per stimare il parametro ignoto è chiamata **STIMA**. La variabile casuale $T = g(X_1, \ldots, X_n)$, di cui t è la realizzazione sul campione effettivamente estratto, viene chiamata **STIMATORE**. Lo stimatore è quindi una variabile casuale dotata di una sua distribuzione campionaria.

Esistono metodi generali di stima puntuale che permettono di costruire stimatori/stime per parametri di interesse in qualunque situazione. Tuttavia, inizieremo ad affrontare il problema della stima puntuale in modo semplice, usando criteri di "ragionevolezza".

1. Se vogliamo stimare una media di popolazione, $m = E(X_i)$, è ragionevole usare come stima la media campionaria $\bar{x}$, che è realizzazione dello stimatore $\bar{X}$. È quello che abbiamo fatto nell'Esempio 1 per stimare μ, l'altezza media degli italiani, ottenendo come stima $\bar{x} = 175$ cm.

2. Se vogliamo stimare la varianza di popolazione, $v^2 = V(X_i)$, è ragionevole usare come stima la varianza campionaria s^2, che è realizzazione dello stimato-

re S^2. È quello che abbiamo fatto nell'Esempio 1 per stimare σ^2, la varianza dell'altezza degli italiani, ottenendo come stima s^2=10,85^2.

3. Se vogliamo stimare una probabilità π di "successo" in una popolazione bernoulliana (ossia, $X_i \sim Be(\pi)$), è ragionevole usare come stima la frazione di "successi" nel campione p

$$p = \frac{1}{n}\sum_{i=1}^{n} x_i$$

che è realizzazione dello stimatore

$$p = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

È quello che abbiamo fatto nell'Esempio 2, ottenendo la stima p=0,1.

N.B.: Per la frazione campionaria di successi, si userà la stessa notazione per la variabile casuale (lo stimatore) e per la sua realizzazione (la stima).

Sono tutti e tre stimatori/stime suggeriti dal "buon senso", ma hanno qualche proprietà? Il punto fondamentale è che lavoriamo con stime, ossia approssimazioni del vero valore; per questo vorremmo sapere

quanto buona è la stima? quanto grande è l'errore di approssimazione?

Risponderemo a queste domande tenendo conto che uno stimatore è una variabile casuale con una sua distribuzione di probabilità (la distribuzione campionaria). Le proprietà dello stimatore possono, quindi, essere studiate analizzando le caratteristiche della sua distribuzione campionaria, come il valore atteso, la varianza, ecc..

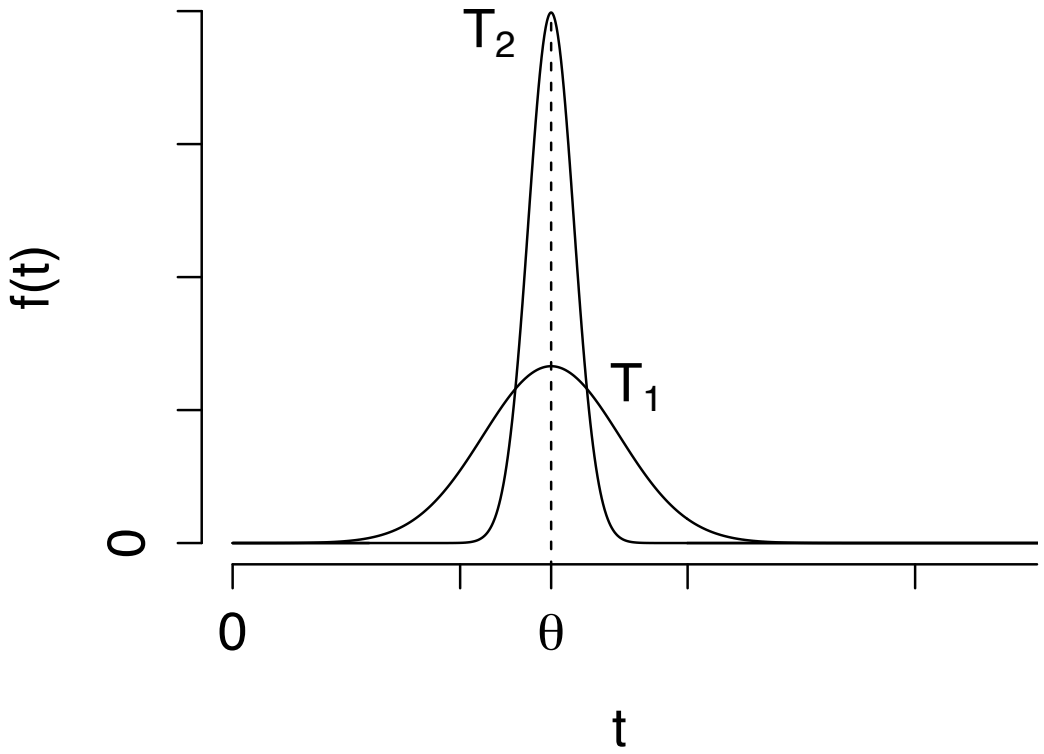
Questo è noto come il **principio del campionamento ripetuto** in base al quale si valutano le proprietà di uno stimatore pensando ad ipotetiche repliche del campionamento. Si ragiona, cioè, come se si estraessero tutti i c.c.s. di dimensione n dalla popolazione di riferimento, e su ciascuno di essi si calcolasse una nuova stima. Viene poi studiato il comportamento complessivo di tutte le stime ottenute, le quali sono realizzazioni della stessa variabile casuale, lo stimatore. Di fatto l'esperimento non viene replicato, ma si ragiona come se lo fosse.

Di seguito, vediamo le più importanti proprietà formali che possono essere richieste ad uno stimatore.

PROPRIETA' DEGLI STIMATORI: 1) L'ERRORE QUADRATICO MEDIO

Indichiamo genericamente con θ il parametro ignoto che vogliamo stimare (ad esempio, $\theta = \mu$ o $\theta = \pi$, ecc.) e con T lo stimatore di θ (ad esempio, $T = \bar{X}$ o $T = p$, ecc.).

Supponiamo che T_1 e T_2 siano due stimatori per lo stesso parametro ignoto θ, le cui funzioni di densità di probabilità sono rappresentate nel grafico sottostante; quale dei due preferireste?



- Idealmente, vorremmo che la distribuzione campionaria dello stimatore T di θ fosse il più possibile concentrata attorno a θ stesso. Se così fosse, infatti, qualunque sia il c.c.s. estratto, la stima t non potrebbe discostarsi molto da θ.

- Per esempio, nel caso molto irrealistico che $T = \theta$ con probabilità 1, avremmo la certezza che qualunque sia il c.c.s. estratto la stima t sia esattamente uguale al parametro da stimare θ.

- Una quantità che viene frequentemente utilizzata per misurare la concentrazione della distribuzione di T attorno a θ è l'**ERRORE QUADRATICO MEDIO (EQM)** definito come

$$\text{EQM}(T) = E\left[(T-\theta)^2\right].$$

- Poiché $(T-\theta)^2$ può essere interpretato come una distanza di T da θ, EQM è la distanza media da θ. In altre parole, EQM misura di quanto in media le realizzazioni di T su tutti i possibili c.c.s. di dimensione n estraibili dalla popolazione di riferimento distano da θ.

- Un valore piccolo di EQM implica che qualunque sia il c.c.s. di dimensione n estratto dalla popolazione, con alta probabilità, la stima t sarà vicina a θ.

- È auspicabile, pertanto, scegliere uno stimatore con EQM piccolo. In particolare, tra due stimatori T_1 e T_2 dello stesso parametro θ, dovremmo scegliere quello con EQM più piccolo.

- L'EQM può essere scomposto nel modo seguente:

$$\text{EQM}(T) = V(T) + [E(T) - \theta]^2$$

Infatti,

$$E[(T-\theta)^2] = E(T^2+\theta^2-2\theta T) = E(T^2)+\theta^2-2\theta E(T) =$$

$$= E(T^2) - [E(T)]^2 + [E(T)]^2 + \theta^2 - 2\theta E(T) =$$

$$= V(T) + [E(T) - \theta]^2.$$

- La quantità

$$E(T) - \theta$$

è chiamata **DISTORSIONE** dello stimatore T.

- Pertanto, l'EQM può essere espresso come

$$\text{EQM}(T) = [\text{distorsione}(T)]^2 + V(T)$$

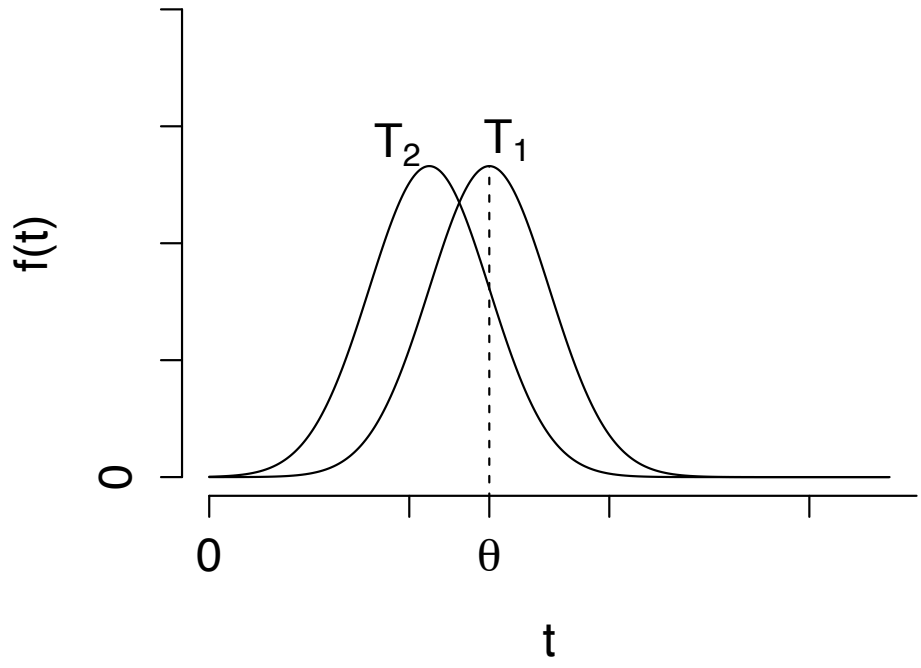
Le due componenti dell'EQM sono

1. la distorsione di T, che è legata alla **posizione** della distribuzione di T;
2. la **variabilità** della distribuzione di T.

- Per avere uno stimatore con EQM piccolo possiamo lavorare su entrambe le componenti. Inizieremo considerando la distorsione.

PROPRIETA' DEGLI STIMATORI: 2) LA CORRETTEZZA

Supponiamo che T_1 e T_2 siano due stimatori per lo stesso parametro ignoto θ, le cui funzioni di densità di probabilità sono rappresentate nel grafico sottostante; quale dei due preferireste?



Uno stimatore T per θ è detto **corretto** o **non distorto** se

$$E(T) = \theta,$$

ossia se la distorsione $E(T) - \theta$ è pari a 0.

- La distribuzione campionaria di uno stimatore corretto è "centrata" attorno a θ.
- La correttezza di uno stimatore garantisce che la media delle stime ottenute su tutti i c.c.s. di dimensione n che possiamo estrarre dalla popolazione sia uguale al parametro ignoto.
- Seppure la correttezza è una proprietà auspicabile, essa fornisce solo delle garanzie in media e non ci assicura che la stima ottenuta sul c.c.s. effettivamente estratto sia uguale a θ (o anche solo vicina).

ESEMPI DI STIMATORI CORRETTI E NON CORRETTI

1. La media campionaria $\bar{X}$ è **sempre** uno stimatore corretto per la media di popolazione $m = E(X_i)$.

   Dimostrazione: Dato $m = E(X_i)$, per ogni $i = 1, \ldots, n$, si ha

   $$E(\bar{X}) = E\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{nm}{n} = m$$

   Esempio 4
   La correttezza della media campionaria è un risultato generale. Tuttavia, per fissare le idee, riprendiamo l'esempio dell'urna con 3 palline numerate da 1 a 3 e verifichiamo per questo caso specifico per via diretta la non distorsione di $\bar{X}$.

   La media di popolazione è $m = 2 = E(X_1) = E(X_2)$.

   Usando la distribuzione campionaria di $\bar{X}$ specificata a pag. 12, abbiamo

   $$E(\bar{X}) = 1 \cdot \frac{1}{9} + 1,5 \cdot \frac{2}{9} + 2 \cdot \frac{3}{9} + 2,5 \cdot \frac{2}{9} + 3 \cdot \frac{1}{9} = 2$$

   Si noti però che se avessimo estratto il campione (1,1) la stima di m sarebbe stata $\bar{x} = 1 \neq 2$.

2. La frazione campionaria di successi p è uno stimatore corretto della frazione π di "successi" nella popolazione.

Dimostrazione: Segue dalla correttezza della media campionaria, notando che p stessa è una media di variabili casuali $Be(\pi)$, ciascuna con valore atteso π.

3. La varianza campionaria S^2 non è uno stimatore corretto della varianza di popolazione v^2.

   Sia $x_1, \ldots, x_n$ un c.c.s. da una popolazione di media m e varianza v^2, ossia $E(X_i) = m$ e $V(X_i) = v^2$, per ogni $i = 1, \ldots, n$. Allora,

$$E(S^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\bar{X}^2)$$

   Tenendo conto del fatto che $E(X_i^2) = V(X_i) + [E(X_i)]^2 = v^2 + m^2$ e, analogamente, $E(\bar{X}^2) = V(\bar{X}) + [E(\bar{X})]^2$, si ha

$$E(S^2) = \frac{1}{n} n(v^2 + m^2) - \left\{V(\bar{X}) + [E(\bar{X})]^2\right\}.$$

   Sappiamo inoltre che $E(\bar{X}) = m$ e

$$V(\bar{X}) = V\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{nv^2}{n} = \frac{v^2}{n}$$

   da cui

$$E(S^2) = v^2 + m^2 - \left\{\frac{v^2}{n} + m^2\right\} = \frac{n-1}{n} v^2.$$

È facile "correggere" S^2 in modo da derivare uno stimatore corretto di v^2. Per i precedenti risultati, infatti, lo stimatore

$$S'^2 = S^2 \frac{n}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

è tale che

$$E(S'^2) = E\left(S^2 \frac{n}{n-1}\right) = E(S^2) \frac{n}{n-1} = v^2.$$

S'^2 è chiamata **varianza campionaria corretta**. Indicheremo la realizzazione campionaria di S'^2 con s'^2. Ad esempio, nell'Esempio 1, la varianza campionaria corretta osservata è $s'^2 = 10,85^2 \times 8/7 = 134,54$

L'EQM PER STIMATORI CORRETTI

- Si noti che se T è uno stimatore corretto di θ, allora

$$\text{EQM}(T) = V(T)$$

e la qualità dello stimatore può semplicemente essere valutata in base alla sua varianza.

Esempio 5
Valutiamo l'EQM della media campionaria $\bar{X}$ come stimatore di una media di popolazione $m = E(X_i)$.
Abbiamo visto che la media campionaria è uno stimatore corretto di m, quindi

$$\text{EQM}(\bar{X}) = V(\bar{X}) = V\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{1}{n^2} \cdot nv^2 = \frac{v^2}{n}$$

dove $v^2 = V(X_i)$ è la varianza di popolazione.
Si noti che l'EQM decresce al crescere di n. In altre parole, più grande è il campione e meglio si comporta lo stimatore. È questa una condizione che deve essere soddisfatta da un qualunque stimatore "sensato".

- Se T_1 e T_2 sono entrambi stimatori corretti per θ, in base al criterio dell'EQM, tra i due è preferibile quello con varianza più piccola.

Esempio 6
Consideriamo $T_1 = X_1$ e $T_2 = \bar{X}$ come stimatori per una media $m = E(X_i)$ di popolazione.
Sono entrambi stimatori corretti, infatti $E(T_1) = E(X_1) = m$, ma

$$\text{EQM}(T_1) = V(T_1) = V(X_1) = v^2$$

e dall'Esempio 5 sappiamo che

$$\text{EQM}(T_2) = V(\bar{X}) = \frac{v^2}{n},$$

dove $v^2 = V(X_i)$. Da cui si deduce che, per ogni $n \geq 1$, EQM$(T_1) \geq$EQM(T_2) e che T_2 è preferibile a T_1.

- Una parte della Statistica Inferenziale (la Teoria ottimale) si dedica a identificare per ciascun problema di stima lo stimatore migliore, ossia con EQM più piccolo.

- Non entreremo nei dettagli della Teoria ottimale, ci limitiamo a dire (senza dimostrazione) che

  - se il c.c.s. $x_1, \ldots, x_n$ è tratto da una **popolazione normale**, ossia $X_i \sim N(\mu, \sigma^2)$ per ogni $i = 1, \ldots, n$, allora la media campionaria $\bar{X}$ e la varianza campionaria corretta S'^2 sono gli stimatori ottimali, rispettivamente, per μ e σ^2, nella classe degli stimatori corretti;

  - se il c.c.s. $x_1, \ldots, x_n$ è tratto da una **popolazione bernoulliana**, ossia $X_i \sim Be(\pi)$ per ogni $i = 1, \ldots, n$, allora la frazione campionaria di "successi" p è lo stimatore ottimale di π nella classe degli stimatori corretti.

PROPRIETÀ DEGLI STIMATORI: 3) LA CONSISTENZA

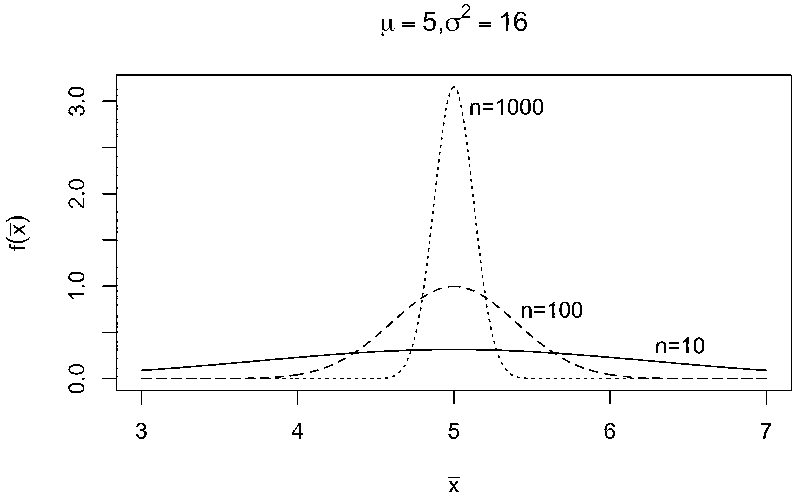
- Tranne in problemi di stima molto semplici, è in generale difficile identificare uno stimatore con proprietà di ottimalità, ad esempio, non distorto e/o con EQM minimo.
- Questo ha indotto gli statistici ad analizzare le proprietà **asintotiche** degli stimatori, ossia come si comporta la distribuzione campionaria dello stimatore per $n \to \infty$.
- Verificare le proprietà di uno stimatore per $n \to \infty$ non è un mero esercizio teorico: l'idea di base è che una proprietà asintotica sarà almeno approssimativamente soddisfatta in campioni di dimensione finita, purché **sufficientemente grandi**.
- La **consistenza** è una delle **principali** proprietà asintotiche che possono essere richieste ad uno stimatore.
- Per sottolineare la dipendenza dalla numerosità campionaria n, indichiamo lo stimatore con $T_n = g(X_1, \ldots, X_n)$.
- Idealmente, si vorrebbe che $T_n \to \theta$, per $n \to \infty$ (se il campione è così grande da coprire l'intera popolazione, dovremmo essere in grado di quantificare le quantità di popolazione di interesse in modo esatto).
- Tuttavia, non possiamo applicare l'usuale definizione di limite, perché T_n non è una successione numerica, bensì una successione di variabile casuali. Abbiamo,

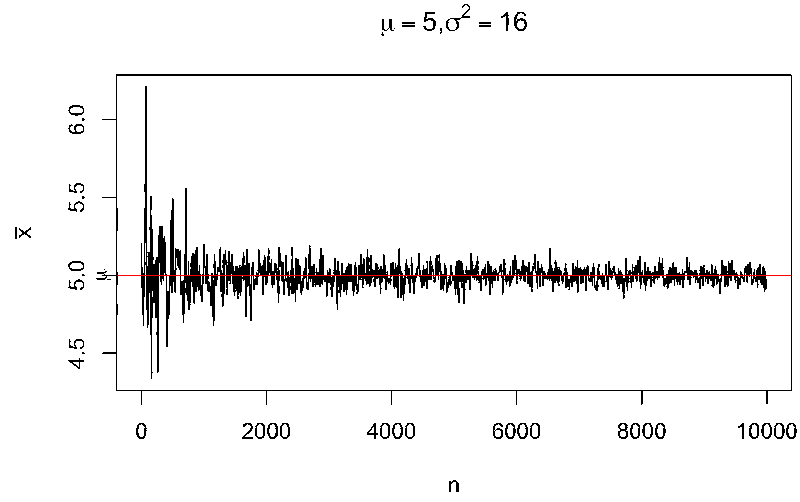
per questo, bisogno di una definizione di convergenza diversa.

- Il problema viene risolto richiedendo che, al crescere di n, la distribuzione campionaria di T_n si concentri sempre di più attorno a θ. Più precisamente, uno stimatore T_n di θ si dice **consistente** se

$$\lim_{n\to\infty} \text{EQM}(T_n) = \lim_{n\to\infty} E[(T_n - \theta)^2] = 0$$

In sostanza, uno stimatore è consistente se la sua accuratezza migliora all'aumentare della numerosità campionaria, sino a diventare perfetta.





- Se T_n è uno stimatore corretto di θ, allora $\text{EQM}(T_n) = V(T_n)$ e lo stimatore è consistente se

$$\lim_{n\to\infty} V(T_n) = 0,$$

- **Esempio 6**
  Riprendiamo gli stimatori $T_1 = X_1$ e $T_2 = \bar{X}$ per una media di popolazione m. Mentre T_2 è uno stimatore consistente, dato che

$$\text{EQM}(T_2) = \frac{v^2}{n} \to 0, \quad \text{per } n \to \infty,$$

  T_1 non è uno stimatore consistente, poiché

$$\text{EQM}(T_1) = v^2 \nrightarrow 0, \quad \text{per } n \to \infty.$$

  dove v^2 è la varianza di popolazione.

- Per gli stimatori $\bar{X}$, S'^2 e p si ha
  - la media campionaria $\bar{X}$ è uno stimatore consistente per una media di popolazione $m = E(X_i)$, sia che la popolazione sia gaussiana che non gaussiana. Segue dagli sviluppi dell'Esempio 6.
  - la varianza campionaria corretta S'^2 è uno stimatore consistente per una varianza di popolazione $v^2 = V(X_i)$, sia che la popolazione sia gaussiana che non gaussiana (non dimostrato).
  - la frazione campionaria di "successi" p è uno stimatore consistente per una probabilità π. Segue dagli sviluppi per la media campionaria, ricordando che p stesso è una media di variabili bernoulliane.

- **Esercizio (per casa)**
  Siano $X_1, \ldots, X_n$ i.i.d. con distribuzione $N(\mu, \sigma^2)$. Dati i due stimatori di μ

$$T_1 = 0,9X_1 + 0,1\left(\frac{1}{n-1}\sum_{i=2}^{n} X_i\right), \quad T_2 = \frac{X_1 + X_n}{2}$$

  1. dimostrare che sono entrambi stimatori corretti per μ;
  2. calcolare l'EQM di entrambi gli stimatori;
  3. verificare se sono stimatori consistenti.

SINTESI DELLE PROPRIETÀ DI $\bar{X}$, S'^2 e p

1. LA MEDIA CAMPIONARIA

- La media campionaria

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

con $X_1, \ldots, X_n$ i.i.d. tali che $E(X_i) = m$ e $V(X_i) = v^2$, è uno stimatore corretto e consistente per m. Il suo EQM è

$$\text{EQM}(\bar{X}) = V(\bar{X}) = \frac{v^2}{n}$$

che fornisce una misura dell'accuratezza di $\bar{X}$ come stimatore di m. Se v^2 non è nota, l'EQM può essere valutato sostituendo v^2 con s'^2.

- Se $X_1, \ldots, X_n$ sono i.i.d. $N(\mu, \sigma^2)$, allora $\bar{X}$ è anche lo stimatore migliore (con EQM più piccolo) tra tutti gli stimatori corretti di μ. Per noti risultati della probabilità (si veda pag. 101 della Probabilità), vale inoltre

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Se il campione ha numerosità n grande, allora, per il teorema del limite centrale, qualunque sia la distribuzione (non gaussiana) delle variabili X_i

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} \dot{\sim} N\left(m, \frac{v^2}{n}\right)$$

dove $m = E(X_i)$ e $v^2 = V(X_i)$ (si confronti con pag. 114 della Probabilità).

2. LA FRAZIONE CAMPIONARIA DI SUCCESSI

- La frazione campionaria di successi

$$p = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

con $X_1, \ldots, X_n$ i.i.d. tali che $X_i \sim Be(\pi)$, è uno stimatore corretto e consistente di una probabilità di "successo" π. È anche lo stimatore con il più piccolo EQM nella classe degli stimatori corretti di π. Il suo EQM è

$$\text{EQM}(p) = V(p) = \frac{\pi(1-\pi)}{n}$$

che si deriva facilmente dall'EQM di $\bar{X}$ sostituendo la varianza generica v^2 con $V(X_i) = \pi(1-\pi)$. L'EQM può essere valutato sostituendo a π la sua stima p.

- Per il teorema del limite centrale, per n grande vale

$$p \dot{\sim} \; N\left(\pi, \frac{\pi(1-\pi)}{n}\right),$$

(si confronti con pag. 115 della Probabilità).

3. LA VARIANZA CAMPIONARIA CORRETTA

- La varianza campionaria corretta

$$S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n}{n-1}S^2$$

con $X_1, \ldots, X_n$ i.i.d. tali che $E(X_i) = m$ e $V(X_i) = v^2$, è uno stimatore corretto e consistente per v^2. (Non viene precisato l'EQM, perché non utilizzato nel seguito).

- Se $X_1, \ldots, X_n$ sono i.i.d. $N(\mu, \sigma^2)$, allora S'^2 è anche lo stimatore migliore (con EQM più piccolo) tra tutti gli stimatori corretti di σ^2. Inoltre,

$$\frac{(n-1)S'^2}{\sigma^2} \sim \chi^2_{n-1}$$

(risultato non dimostrato)

UN METODO GENERALE DI STIMA PUNTUALE: IL METODO DEI MOMENTI

- Gli stimatori $\bar{X}$, p e S'^2 sono stati proposti sulla base di criteri di "ragionevalezza", ma è importante poter disporre di metodi generali di costruzione di stimatori, applicabili, in linea di principio, a qualunque problema di stima e per qualunque parametro.

- Esistono diversi metodi generali per la stima puntuale. Noi copriremo uno solo di questi, **il metodo dei momenti**.

- Sia $x_1, \ldots, x_n$ un c.c.s. e θ il parametro da stimare. Il primo passo del metodo dei momenti consiste nel determinare $E(X_i)$ che sarà in generale una funzione di θ. Al secondo passo, si risolve rispetto a θ l'equazione

  $$E(X_i) = \bar{x}.$$

  La soluzione dell'equazione, che indichiamo con $\tilde{\theta}$, è la stima con il metodo dei momenti di θ.

- L'idea di base del metodo dei momenti è eguagliare la media teorica (anche chiamata **momento primo** teorico) $E(X_i)$ alla corrispondente media campionaria $\bar{x}$ e da questa equazione derivare una stima per θ.

- **Esempio 1**
  Se vogliamo stimare l'altezza media degli italiani μ, in base al metodo dei momenti, dovremmo risolvere rispetto a μ l'equazione

$$E(X_i) = \mu = \bar{x} = 175,$$

  ottenendo la stima $\tilde{\mu} = \bar{x} = 175$.

- **Esempio 2**
  Se vogliamo stimare la frazione π di guarnizioni difettose nell'intera popolazione, in base al metodo dei momenti dovremmo risolvere rispetto a π l'equazione

$$E(X_i) = \pi = \bar{x} = p = 0,1$$

  dove si tiene conto del fatto che, essendo le x_i realizzazioni di variabili casuali bernoulliane, $\bar{x} = p$. La stima con il metodo dei momenti di π è quindi $\tilde{\pi} = p = 0,1$.

- **Esempio 3**
  Se vogliamo stimare il numero medio giornaliero di utilizzi del bancomat λ, in base al metodo dei momenti dovremmo risolvere rispetto a λ l'equazione

$$E(X_i) = \lambda = \bar{x} = 7,6.$$

  La stima con il metodo dei momenti di λ è quindi $\tilde{\lambda} = \bar{x} = 7,6$.

- **Esempio 7**
  È stato osservata la durata di 15 lampadine scelte casualmente tra quelle prodotte da una fabbrica. La durata media delle 15 lampadine del campione è

$\bar{x} = 65$ ore. Si può ipotizzare che la durata delle lampadine abbia distribuzione esponenziale di parametro λ ignoto. Si vuole stimare λ.

Con queste ipotesi le variabili casuali che descrivono il campione, $X_1, \ldots, X_{15}$, sono i.i.d. $Exp(\lambda)$. Applicando il metodo dei momenti, si ottiene una stima puntuale per λ risolvendo l'equazione

$$E(X_i) = 1/\lambda = \bar{x} = 65.$$

La stima con il medoto dei momenti per λ è quindi $\tilde{\lambda} = 1/\bar{x} = 1/65$.

- Il metodo dei momenti è applicabile anche al caso in cui si debbano stimare due o più parametri ignoti. Vediamo il caso di due parametri.

  Supponiamo che la distribuzione delle X_i dipenda da due parametri ignoti, θ_1 e θ_2, che devono essere stimati. Al primo passo il metodo dei momenti richiede il calcolo di $E(X_i)$ e $E(X_i^2)$ che saranno, in generale, funzioni di θ_1 e θ_2. Al secondo passo, si risolve rispetto a θ_1 e θ_2 il sistema di equazioni

  $$\begin{cases} E(X_i) = \bar{x} \\ E(X_i^2) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{cases}$$

  Le soluzioni del sistema, $\tilde{\theta}_1$ e $\tilde{\theta}_2$, sono le stime con il metodo dei momenti di θ_1 e θ_2, rispettivamente.

- L'idea di base è ancora eguagliare le medie teoriche, $E(X_i)$ e $E(X_i^2)$, con la corrispondenti medie campionarie. Lo stesso meccanismo può essere applicato

a più di due parametri. Il nome del metodo deriva dal fatto che $E(X^j)$ è anche chiamato il **momento j-esimo** di X_i.

- **Esempio 1**
  Nell'esempio sull'altezza degli italiani, sia la media di popolazione μ che la varianza di popolazione σ^2 sono ignote e devono essere stimate dalle osservazioni. Applicando il metodo dei momenti, abbiamo che $E(X_i) = \mu$ e $E(X_i^2) = \sigma^2 + \mu^2$ e risolvendo rispetto a μ e σ^2 il sistema di equazioni

$$\begin{cases} \mu = \bar{x} \\ \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{cases}$$

  si ottengo le stime $\tilde{\mu} = \bar{x} = 175$ e $\tilde{\sigma}^2 = s^2 = 10,85^2$, rispettivamente, per μ e σ^2. Le stime con il metodo dei momenti della media e della varianza di popolazione sono quindi la media campionaria e la varianza campionaria **non corretta**, rispettivamente.

- **Proprietà del metodo dei momenti:**
  - È un metodo molto semplice e quasi sempre applicabile con campioni casuali semplici.
  - Produce stimatori consistenti, ma non necessariamente corretti e/o ottimali (con EQM minimo). Può produrre stimatori corretti e/o ottimali, ma le due proprietà non sono garantite e devono essere verificate caso per caso.

STIMA INTERVALLARE

L'obiettivo della stima intervallare è identificare, sulla base delle osservazioni del campione, un intervallo di valori che, con un certo grado di attendibilità, contenga l'ignoto parametro di popolazione θ.

Esempio 2
Avevamo ottenuto la stima puntuale $p = 0,1$. Possiamo dire che l'intervallo (0,05,0,15) è un intervallo di valori ragionevoli per π, la vera frazione di guarnizioni difettose nella popolazione? quali garanzie ci dà questo intervallo di includere π?

GLI INTERVALLI DI CONFIDENZA

Il problema di individuare un intervallo di valori ragionevoli per un parametro θ ignoto può essere affrontato trovando due statistiche campionarie, $T_I = g_I(X_1, \ldots, X_n)$ e $T_S = g_S(X_1, \ldots, X_n)$, tali che

$$P\left(T_I \leq \theta \leq T_S\right) = 1 - \alpha$$

dove α è un valore prefissato.

In questo modo, la probabilità che θ sia incluso nell'intervallo (T_I, T_S) è $1 - \alpha$.

L'intervallo (T_I, T_S) viene chiamato **intervallo di confidenza di livello 1-α o di livello (1-α)·100%** per θ.

CONSIDERAZIONI

1. α è la probabilità che l'intervallo di confidenza **non** includa θ e, idealmente, vorremmo che fosse più piccola possibile. Ma attenzione, se ponessimo $\alpha = 0$, l'intervallo dovrebbe comprendere tutti i valori che possono essere assunti da θ e questo intervallo non ha alcuna utilità pratica. In generale, più piccolo è α più ampio è l'intervallo di confidenza, identificando così in modo meno preciso dove si colloca θ. Esiste un *trade-off* tra la garanzia di includere θ e la precisione dell'intervallo. Come compromesso, tipicamente si sceglie $\alpha \in [0,01; 0,1]$.
2. Essendo T_I e T_S due variabili casuali, anche l'intervallo (T_I, T_S) **è casuale**. Dopo aver osservato il c.c.s. $x_1, \ldots, x_n$ avremo che $t_I = g_I(x_1, \ldots, x_n)$ e $t_S = g_S(x_1, \ldots, x_n)$ sono le realizzazioni effettivamente osservate di T_I e T_S. Pertanto, l'intervallo di confidenza osservato (t_I, t_S) non è più casuale, ma **numerico**.
3. Non è corretto dire che l'intervallo numerico (t_I, t_S) contiene θ con probabilità $1 - \alpha$. Questa affermazione vale solo per l'intervallo casuale (T_I, T_S). Richiamando il principio del campionamento ripetuto, la corretta interpretazione di un intervallo di confidenza di livello $1 - \alpha$ è la seguente. Se ripetiamo il campionamento e su tutti i possibili c.c.s. di dimensione n calcoliamo un nuovo intervallo osservato (t_I, t_S), allora $(1 - \alpha) \cdot 100\%$ di questi contiene θ. Tuttavia, sul singolo campione effettivamente estratto non possiamo sapere se contenga θ oppure no.

METODI DI COSTRUZIONI DI INTERVALLI DI CONFIDENZA

Esistono varie procedure di costruzione di intervalli di confidenza per un parametro θ di popolazione ignoto.
In generale, esse si fondano sui seguenti passi.

1. Si sceglie il livello di confidenza $1 - \alpha$.
2. Si identifica uno stimatore T di θ.
3. Si cerca una funzione di T e di θ, $Q(T, \theta)$ (**quantità pivotale**), la cui distribuzione di probabilità sia completamente nota.
4. Se $q_{\alpha/2}$ e $q_{1-\alpha/2}$ sono, rispettivamente, i quantili $\alpha/2$ e $1 - \alpha/2$ di $Q(T, \theta)$, allora
$$P(q_{\alpha/2} \leq Q(T, \theta) \leq q_{1-\alpha/2}) = 1 - \alpha$$
5. Isolando θ all'interno della disuguaglianza, si deriva un intervallo di confidenza per θ.

ESEMPI DI INTERVALLI DI CONFIDENZA

1. INTERVALLI DI CONFIDENZA PER I PARAMETRI DI UNA POPOLAZIONE NORMALE

Sia $x_1, x_2, \ldots, x_n$ un c.c.s. da una popolazione normale, ossia X_1, X_2, ..., X_n sono i.i.d. con distribuzione $N(\mu, \sigma^2)$.

INTERVALLO DI CONFIDENZA PER μ

- **CASO CON σ^2 NOTA**
  Supponiamo di conoscere σ^2, la varianza di popolazione, ma di non conoscere μ, la media di popolazione. Vogliamo costruire un intervallo di confidenza di livello $1-\alpha$ per μ.
  Sappiamo che

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

  da cui

$$\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1).$$

  Allora,

$$P\left(-z_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq z_{1-\alpha/2}\right) = 1-\alpha$$

ricordando che $z_{\alpha/2} = -z_{1-\alpha/2}$. Ora, isolando μ si ottiene

$$P\left(\underbrace{\bar{X} - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}}_{T_I} \leq \mu \leq \underbrace{\bar{X} + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}}_{T_S}\right) = 1-\alpha$$

Allora, un intervallo di confidenza di livello $1-\alpha$ per la media μ quando la varianza σ^2 è nota è

$$\left(\bar{X} - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \bar{X} + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right)$$

o più brevemente

$$\left(\bar{X} \pm z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right)$$

L'ampiezza dell'intervallo è

$$2z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}$$

A parità di tutto il resto, al crescere di n l'ampiezza dell'intervallo diminuisce, ossia l'intervallo diventa più preciso. A parità di tutto il resto, se diminuisce α, $z_{1-\alpha/2}$ aumenta e quindi l'ampiezza dell'intervallo aumenta. A parità di tutto il resto, se diminuisce σ^2, diminuisce l'ampiezza dell'intervallo.

L'intervallo numerico, basato sulle osservazioni $x_1, \ldots, x_n$, realizzazione dell'intervallo casuale trovato è

$$\left(\bar{x} \pm z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right)$$

dove $\bar{x}$ è la media calcolata sul campione estratto.

Esempio 8
Si supponga che il peso (in kg) di certe confezioni abbia distribuzione normale di media μ ignota e deviazione standard pari a 2,5kg. Su un campione casuale di 100 confezioni è stato calcolato un peso medio di 11,5kg. Si costruisca un intervallo di confidenza di livello 95% per μ.

E' chiaro dal testo che la varianza σ^2 può considerarsi nota e pari a $2{,}5^2\text{kg}^2$. Si tratta allora di usare l'espressione per l'intervallo di confidenza per μ appena trovata:

$$\left(\bar{x} \pm z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right)$$

In questo caso

$$\bar{x} = 11,5\text{kg}$$

$$1-\alpha = 0,95 \text{ e quindi } \alpha = 0,05$$

$$z_{1-\alpha/2} = z_{0,975} = 1,96$$

$$n = 100$$

L'intervallo di confidenza (numerico) di livello 95% per μ è

$$\left(11,5 \pm 1,96\sqrt{\frac{2,5^2}{100}}\right) = (11,02\text{kg}; 11,98\text{kg})$$

Interpretazione ...

- **CASO CON σ^2 IGNOTA**

  Siamo nella stessa situazione di prima, ma supponiamo ora che sia μ che σ^2 siano ignoti. Il nostro parametro di interesse rimane la media di popolazione μ e, in particolare, vogliamo costruire un intervallo di confidenza per μ.
  Nel caso precedente abbiamo usato la quantità

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1)$$

  per costruire l'intervallo per μ. Ora però σ^2 non è nota e questa quantità non è direttamente utilizzabile. Possiamo procedere in questo modo: stimare σ^2 tramite il suo stimatore corretto S'^2 e sostituire questo stimatore all'interno della precedente quantità, ottenendo

$$\frac{\bar{X} - \mu}{\sqrt{\frac{S'^2}{n}}} \sim t_{n-1}$$

  Infatti, $\bar{X}$ e S'^2 sono indipendenti (non dimostrato). Inoltre,

$$\frac{\bar{X} - \mu}{\sqrt{\frac{S'^2}{n}}} = \frac{(\bar{X} - \mu)}{\sqrt{\frac{S'^2(n-1)\sigma^2}{n(n-1)\sigma^2}}} = \frac{(\bar{X} - \mu)/\sqrt{\frac{\sigma^2}{n}}}{\sqrt{\frac{S'^2(n-1)}{(n-1)\sigma^2}}} \sim t_{n-1},$$

  dato che il numeratore dell'ultima frazione ha distribuzione $N(0, 1)$ e al denominatore abbiamo $\sqrt{\chi^2_{n-1}/(n-1)}$.

Da questo risultato, si ha che

$$P\left(-t_{n-1;1-\alpha/2} \leq \frac{\bar{X}-\mu}{\sqrt{\frac{S'^2}{n}}} \leq t_{n-1;1-\alpha/2}\right) = 1-\alpha$$

tenendo conto del fatto che $t_{n-1;\alpha/2} = -t_{n-1,1-\alpha/2}$. Ora, isolando μ, si ottiene

$$P\left(\underbrace{\bar{X} - t_{n-1;1-\alpha/2}\sqrt{\frac{S'^2}{n}}}_{T_I} \leq \mu \leq \underbrace{\bar{X} + t_{n-1;1-\alpha/2}\sqrt{\frac{S'^2}{n}}}_{T_S}\right) = 1-\alpha$$

Allora, un intervallo di confidenza per μ di livello $1-\alpha$, quando anche σ^2 non è nota, è

$$\left(\bar{X} - t_{n-1;1-\alpha/2}\sqrt{\frac{S'^2}{n}}, \bar{X} + t_{n-1;1-\alpha/2}\sqrt{\frac{S'^2}{n}}\right)$$

o, più brevemente,

$$\left(\bar{X} \pm t_{n-1;1-\alpha/2}\sqrt{\frac{S'^2}{n}}\right)$$

Si noti che, valendo $S'^2 = nS^2/(n-1)$, l'intervallo ottenuto è equivalente a

$$\left(\bar{X} \pm t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n-1}}\right)$$

dove S^2 la varianza campionaria non corretta.

L'intervallo numerico, basato sulle osservazioni $x_1, \ldots, x_n$, corrispondente all'intervallo casuale trovato è

$$\left(\bar{x} \pm t_{n-1;1-\alpha/2} \sqrt{\frac{s'^2}{n}} \right)$$

dove $\bar{x}$ e s'^2 sono, rispettivamente, la media e la varianza corretta calcolate sul campione.

Si ricordi infine, che, per n sufficientemente grande, diciamo n maggiore di 31, possiamo approssimare $t_{n-1;1-\alpha/2}$ con $z_{1-\alpha/2}$.

Esempio 9

Una macchina è stata predisposta per riempire bottiglie di olio dal contenuto nominale di 1 litro. Si sa che la macchina commette degli errori in fase di riempimento e che la quantità effettivamente versata su ciascuna bottiglia segue una distribuzione gaussiana. La macchina viene fermata e controllata ogni qualvolta si verifica che la media giornaliera esce dai limiti di ± 5 ml.
Se l'osservazione per un dato giorno viene fatta solo su un campione di 10 bottiglie, ottenendo i seguenti dati (in litri)

$$1,000 \quad 0,998 \quad 1,003 \quad 1,002 \quad 0,999 \quad 0,997$$

$$0,999 \quad 1,001 \quad 1,000, \quad 1,010$$

utilizzare i dati per vedere se la macchina deve essere fermata trovando un intervallo di confidenza al 99% per la media giornaliera.

Poichè gli errori nel riempimento sono normalmente distribuiti, possiamo vedere le 10 osservazioni x_1, x_2, ..., x_{10} come un c.c.s. da una $N(\mu, \sigma^2)$, dove μ rappresenta la media della quantità di olio versata nelle bottiglie. Sia μ che σ^2 sono ignote. Vogliamo costruire un intervallo di confidenza di livello 0,99 per μ. Calcoliamo gli ingredienti dell'intervallo di confidenza.

$$\bar{x} = 1,0009 \text{ litri}$$

$$s'^2 = 0,00001343 \text{ litri}^2$$

$$1-\alpha = 0,99 \quad \alpha = 0,01 \quad \alpha/2 = 0,005 \quad 1-\alpha/2 = 0,995$$

$$t_{9;0,995} = 3,2498$$

L'intervallo cercato è

$$\left(1,0009 - 3,2498\sqrt{\frac{0,00001343}{10}}; 1,0009 + 3,2498\sqrt{\frac{0,00001343}{10}}\right) =$$

$$= (0,99614; 1,00466) \text{ litri}$$

In base all'intervallo ottenuto, fermeremmo la macchina? La tolleranza è ± 5 ml sul contenuto nominale di 1 lt, ossia riteniamo la macchina "fuori controllo" se la media giornaliera esce al di fuori dell'intervallo (0,995,1,005) lt. In base all'intervallo ottenuto, la media giornaliera cade all'interno dei limiti di tolleranza e quindi non fermeremmo la macchina.

INTERVALLO DI CONFIDENZA PER σ^2

Spostiamo ora l'interesse da μ, che è stato il parametro di interesse nei precedenti intervalli, a σ^2, la varianza della popolazione. Vogliamo costruire un intervallo di confidenza di livello $1-\alpha$ per σ^2, quando anche μ è ignoto.

Così come la media campionaria $\bar{X}$ è stato il punto di partenza per costruire un intervallo di confidenza per μ, la varianza campionaria (nella sua versione corretta) S'^2 è il punto di partenza per costruire un intervallo di confidenza per σ^2. Sappiamo che, per una popolazione normale,

$$\frac{(n-1)S'^2}{\sigma^2} \sim \chi^2_{n-1}$$

Pertanto,

$$P\left(\chi^2_{n-1;\alpha/2} \leq \frac{(n-1)S'^2}{\sigma^2} \leq \chi^2_{n-1;1-\alpha/2}\right) = 1-\alpha$$

Ora isoliamo σ^2, ottenendo

$$P\left(\frac{(n-1)S'^2}{\chi^2_{n-1;1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S'^2}{\chi^2_{n-1;\alpha/2}}\right) = 1-\alpha$$

Allora, un intervallo di confidenza di livello $1-\alpha$ per σ^2 è

$$\left(\frac{(n-1)S'^2}{\chi^2_{n-1;1-\alpha/2}}, \frac{(n-1)S'^2}{\chi^2_{n-1;\alpha/2}}\right)$$

Se volessimo un intervallo per σ piuttosto che per σ^2 basterà prendere la radice quadrata di entrambi gli estremi:

$$\left(\sqrt{\frac{(n-1)S'^2}{\chi^2_{n-1;1-\alpha/2}}}, \sqrt{\frac{(n-1)S'^2}{\chi^2_{n-1;\alpha/2}}} \right)$$

Infatti,

$$P\left(\frac{(n-1)S'^2}{\chi^2_{n-1;1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S'^2}{\chi^2_{n-1;\alpha/2}} \right) =$$

$$= P\left(\sqrt{\frac{(n-1)S'^2}{\chi^2_{n-1;1-\alpha/2}}} \leq \sigma \leq \sqrt{\frac{(n-1)S'^2}{\chi^2_{n-1;\alpha/2}}} \right) = 1-\alpha$$

Esempio 10
Una nuova terapia è stata sperimentata su un campione di 12 pazienti e i tempi di guarigione osservati sono stati (in giorni)
15 23 32 18 25 16 27 22 30 41 18 29

(a) Assumendo una distribuzione normale per il tempo di guarigione, trovare un intervallo di confidenza di livello 95% per il tempo medio di guarigione dei pazienti sottoposti a terapia.

Le 12 osservazioni rappresentano un c.c.s. da $N(\mu, \sigma^2)$, con μ e σ^2 entrambi ignoti. In particolare, μ è il tempo medio (in giorni) di guarigione per il quale vogliamo costruire un intervallo di confidenza. Gli ingredienti dell'intervallo sono

$$\bar{x} = 24,67 \text{ giorni}$$

$$s^2 = 53,39$$

da cui

$$s'^2 = \frac{S^2 \cdot 12}{11} = 58,24$$

$$1 - \alpha = 0,95 \quad \alpha = 0,05 \quad 1 - \alpha/2 = 0,975$$

$$t_{11;0,975} = 2,2$$

L'intervallo cercato è

$$\left(24,67 \pm 2,2\sqrt{\frac{58,24}{12}}\right) = (19,82; 29,52) \text{ giorni}$$

(b) Si dia un intervallo di confidenza al 90% per la varianza σ^2

$$1-\alpha = 0,9 \quad \alpha = 0,1 \quad 1-\alpha/2 = 0,95 \quad \alpha/2 = 0,05$$

$$\chi^2_{11;0,05} = 4,57 \quad \chi^2_{11,0,95} = 19,68$$

L'intervallo cercato per σ^2 è

$$\left(\frac{11 \cdot 58,24}{19,68}; \frac{11 \cdot 58,24}{4,57}\right) = (32,55; 140,18)$$

2. INTERVALLO DI CONFIDENZA APPROSSIMATO PER UNA PROBABILITÀ

Sia $x_1, x_2, \ldots, x_n$ un c.c.s. da una popolazione bernoulliana, ossia $X_1, X_2, \ldots, X_n$ sono i.i.d. $Be(\pi)$.

Non conosciamo la probabilità π di "successo" nella popolazione e vogliamo costruire un intervallo di confidenza per π di livello $1-\alpha$. Partiamo da

$$p = \frac{1}{n}\sum_{i=1}^{n} X_i$$

che sappiamo essere uno stimatore ottimale per π. Sappiamo inoltre che, per n grande,

$$p \dot{\sim} \; N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Allora,

$$\frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \dot{\sim} \; N(0,1)$$

da cui

$$P\left(-z_{1-\alpha/2} \leq \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \leq z_{1-\alpha/2}\right) \doteq 1-\alpha$$

Ora, isolando π, otteniamo

$$P\left(p - z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \leq p \leq p + z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right) \doteq 1-\alpha$$

da cui un intervallo di confidenza di livello **approssimato** $1-\alpha$ per π è

$$\left(p - z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}, p + z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

Il problema è che non conoscendo π non siamo in grado di calcolare $\sqrt{\pi(1-\pi)/n}$. Tuttavia, poiché stiamo lavorando con grandi campioni, possiamo sostituire a π il suo stimatore consistente p ottenendo l'intervallo

$$\left(p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)$$

Il corrispondente intervallo numerico si ottiene ponendo p uguale alla frazione osservata di "successi" nel campione.

OSSERVAZIONE: Affinché l'approssimazione risulti buona, dovremmo avere $n\pi, n(1-\pi) \geq 5$, ma non conoscendo π, possiamo verificare se $np, n(1-p) \geq 5$. Dobbiamo avere quindi a disposizione almeno 10 osservazioni, di cui almeno 5 "successi" e 5 "insuccessi".

Esempio 11
Il direttore di una banca di una piccola città intende investigare la proporzione di depositanti che vengono pagati mensilmente. Per compiere tale studio vengono scelti in modo casuale 200 depositanti e di questi 23 affermano di essere pagati mensilmente. Stimare la vera proporzione di depositanti della banca pagati mensilmente e costruire un intervallo di confidenza di livello 90% per tale proporzione.

Abbiamo due modi equivalenti di vedere il problema.
1. Osserviamo 200 valori, $x_1, \ldots, x_{200}$, con

$$x_i = \begin{cases} 1 & \text{se l'i-esimo cliente è pagato mensilmente} \\ 0 & \text{se l'i-esimo cliente non è pagato mensilmente} \end{cases}$$

23 x_i sono uguali a 1 e 200-23 x_i sono uguali a 0. In questo caso abbiamo un c.c.s di dimensione 200 da $Be(\pi)$.

2. Osserviamo un unico valore $x = 23$ da una $Bin(200, \pi)$.

Qualunque sia il modo di interpretare l'indagine, sappiamo che

$$p = \frac{23}{200} = 0,115 = \text{frazione di successi osservati}$$

è una stima della frazione π nella popolazione.

Per costruire l'intervallo di confidenza,

$$1 - \alpha = 0,9 \quad \alpha = 0,1 \quad 1 - \alpha/2 = 0,95$$

$$z_{0,95} = 1,64$$

L'intervallo di confidenza cercato è

$$\left(0,115 - 1,64\sqrt{\frac{0,115(1-0,115)}{200}}, (0,115 + 1,64\sqrt{\frac{0,115(1-0,115)}{200}}\right) =$$

$$= (0,082; 0,158)$$

3. INTERVALLO DI CONFIDENZA APPROSSIMATO PER LA MEDIA DI UNA POPOLAZIONE NON-NORMALE

Sia $x_1, \ldots, x_n$ un c.c.s. da una popolazione di media m e varianza v^2, ossia $E(X_i) = m$ e $V(X_i) = v^2$ per ogni $i = 1, \ldots, n$.

Si vuole costruire un intervallo di confidenza di livello $1 - \alpha$ per m. In generale, per poter costruire tale intervallo è necessario conoscere la distribuzione delle X_i che non è però qui specificata. Tuttavia, se n è **grande**, generalizzando quanto visto per π, è possibile costruire un intervallo che abbia almeno livello **approssimato** $1 - \alpha$, usando argomentazioni asintotiche. Infatti, per il teorema del limite centrale, per n grande, si ha che

$$\bar{X} \dot{\sim} \; N\left(m, \frac{v^2}{n}\right)$$

da cui

$$\frac{\bar{X} - m}{\sqrt{v^2/n}} \dot{\sim} \; N\left(0, 1\right).$$

e

$$P\left(-z_{1-\alpha/2} \leq \frac{\bar{X} - m}{\sqrt{v^2/n}} \leq z_{1-\alpha/2}\right) \doteq \; 1 - \alpha.$$

Isolando m all'interno della disuguaglianza, si deriva

$$P\left(\bar{X} - z_{1-\alpha/2}\sqrt{v^2/n} \leq m \leq \bar{X} + z_{1-\alpha/2}\sqrt{v^2/n}\right) \doteq \; 1-\alpha,$$

e quindi un intervallo di confidenza di livello **approssimato** $1-\alpha$ per m per grandi campioni è dato da

$$\left(\bar{X} - z_{1-\alpha/2}\sqrt{v^2/n}, \bar{X} + z_{1-\alpha/2}\sqrt{v^2/n}\right).$$

Se, come spesso accade, v^2 non è noto, poiché stiamo lavorando con grandi campioni, possiamo sostituirlo con il suo stimatore consistente S'^2, ottenendo l'intervallo

$$\left(\bar{X} - z_{1-\alpha/2}\sqrt{S'^2/n}, \bar{X} + z_{1-\alpha/2}\sqrt{S'^2/n}\right).$$

Il corrispondente intervallo numerico è

$$\left(\bar{x} - z_{1-\alpha/2}\sqrt{s'^2/n}, \bar{x} + z_{1-\alpha/2}\sqrt{s'^2/n}\right).$$

N.B.:

- L'espressione dell'intervallo ottenuto per v^2 noto è la stessa di quello per una popolazione normale. Tuttavia, nel caso di una popolazione normale l'intervallo ha esatto livello $1-\alpha$ indipendentemente dalla numerosità campionaria, mentre nel caso di una popolazione non-normale l'intervallo ha livello approssimato $1-\alpha$ solo in grandi campioni. La qualità dell'approssimazione migliora con la dimensione del campione.
- L'intervallo di confidenza ottenuto per π è un caso particolare di questa procedura, in cui $\bar{X} = p$ e, sfruttando il fatto che per una popolazione bernoulliana $v^2 = \pi(1-\pi)$, lo stimatore utilizzato per v^2 non è S'^2, bensì $p(1-p)$.

Esempio 12
Si è interessati a studiare il reddito medio mensile delle famiglie di un certo paese. A questo scopo si estrae un campione di 196 famiglie il cui reddito medio mensile risulta pari a $\bar{x} = 1864$ euro con varianza campionaria corretta pari a $s'^2 = 141,61$. Si costruisca un intervallo di confidenza di livello $1-\alpha = 0,95$ per il reddito medio mensile della popolazione.

La distribuzione del reddito è generalmente caratterizzata da una forte asimmetria positiva, determinata dalla presenza di pochi redditi molto elevati. Per questo non è ragionevole assumere che le X_i siano normalmente distribuite. Tuttavia, essendo la numerosità campionaria piuttosto elevata possiamo ricorrere ad argomenti asintotici e utilizzare l'intervallo approssimato

$$\left(1864 \pm z_{0,975}\sqrt{141,61/196}\right) = (1865,67; 1862,33).$$

4. INTERVALLO DI CONFIDENZA PER LA DIFFERENZA DELLE MEDIE DI DUE POPOLAZIONI NORMALI

Supponiamo di voler confrontare due popolazioni in relazione ad un fenomeno di interesse. Ad esempio, vogliamo confrontare il reddito delle famiglie italiane con il reddito delle famiglie francesi, oppure la redditività di una strategia aziendale A contro la redditività di una strategia B. A questo fine prendiamo due c.c.s.:
$x_1, x_2, \ldots, x_{n_1}$ dalla popolazione I
$y_1, y_2, \ldots, y_{n_2}$ dalla popolazione II
e supponiamo che
$x_1, x_2, \ldots, x_{n_1}$ siano realizzazioni di $X_1, X_2, \ldots, X_{n_1}$ i.i.d. $N(\mu_1, \sigma_1^2)$;
$y_1, y_2, \ldots, y_{n_2}$ siano realizzazioni di $Y_1, Y_2, \ldots, Y_{n_2}$ i.i.d. $N(\mu_2, \sigma_2^2)$;
e i due campioni siano tra loro indipendenti.

Ad esempio, per confrontare il reddito dei francesi e il reddito degli italiani, prendiamo n_1 famiglie francesi e ne osserviamo il reddito e n_2 famiglie italiane e ne osserviamo il reddito. Quindi assumiamo che sia per i francesi che per gli italiani il reddito sia normalmente distribuito, ma con medie e varianze che possono essere diverse.
Siamo interessati a confrontare il livello medio del fenomeno nelle due popolazioni, ossia μ_1 e μ_2 (ad esempio, il reddito medio dei francesi contro il reddi-

to medio degli italiani) e, per questo, costruiamo un intervallo di confidenza per $\mu_1 - \mu_2$.

- **CASO CON VARIANZE NOTE**
  Supponiamo inizialmente che σ_1^2 e σ_2^2 siano note. Il punto di partenza per costruire un intervallo di confidenza per $\mu_1 - \mu_2$ è la differenza tra le due medie campionarie

$$\bar{X} - \bar{Y} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i - \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

  Sappiamo che

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$

  e

$$\bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

BREVE RICHIAMO DI PROBABILITA'

Se $X \sim N(\mu_1, \sigma_1^2)$ e $Y \sim N(\mu_2, \sigma_2^2)$ indipendenti, allora,

$$aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$$

Dal richiamo di probabilità,

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

e

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Allora,

$$P\left(-z_{1-\alpha/2} \leq \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_2^2}{n_1}+\frac{\sigma_1^2}{n_2}}} \leq z_{1-\alpha/2}\right) = 1-\alpha$$

Se isoliamo $\mu_1 - \mu_2$, otteniamo

$$P\left((\bar{X}-\bar{Y}) - z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \leq \mu_1-\mu_2 \leq \right.$$

$$\left.(\bar{X}-\bar{Y}) + z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right) = 1-\alpha$$

Abbiamo quindi ottenuto l'intervallo di confidenza di livello $1-\alpha$ per $\mu_1 - \mu_2$, che è dato da

$$\left((\bar{X}-\bar{Y}) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$$

Il corrispondente intervallo numerico, si ottiene sostituendo alle variabili casuali $\bar{X}$ e $\bar{Y}$ le medie campionarie osservate $\bar{x}$ e $\bar{y}$.

Esempio 13
Per valutare la diversità della preparazione degli studenti di due corsi di laurea A e B della facoltà di Economia, 50 studenti del corso A e 100 studenti del corso B vengono sottoposti ad un test. Gli studenti del corso A e B ottengono un punteggio medio di 3 e 2,5, rispettivamente. Si supponga inoltre che le variabili casuali X e Y che generano i punteggi nei due corsi di laurea siano indipendenti e tali che

$$X \sim N(\mu_1, (0,2)^2)$$
$$Y \sim N(\mu_2, (0,5)^2)$$

Si trovi un intervallo di confidenza di livello 0,99 per $\mu_1 - \mu_2$. I dati indicano che il corso di laurea non influenza il punteggio, ossia $\mu_1 = \mu_2$?

Gli ingredienti di cui abbiamo bisogno per costruire l'intervallo sono

$$\bar{x} = 3$$
$$\bar{y} = 2,5$$
$$1 - \alpha = 0,99 \quad \alpha = 0,01 \quad 1 - \alpha/2 = 0,995$$
$$z_{0,995} = 2,58$$

L'intervallo cercato è

$$\left((3 - 2,5) \pm 2,58 \sqrt{\frac{0,2^2}{50} + \frac{0,5^2}{100}} \right) = (0,35; 0,65)$$

Poiché l'intervallo contiene solo valori positivi, abbiamo l'indicazione che $\mu_1 - \mu_2 > 0$, ossia $\mu_1 > \mu_2$.

- **CASO CON VARIANZE IGNOTE**
  Nel caso in cui le varianze di popolazione, σ_1^2 e σ_2^2, non siano note, si riesce a costruire un intervallo di confidenza per $\mu_1 - \mu_2$ solo se assumiamo che

$$\sigma_1^2 = \sigma_2^2 = \sigma^2,$$

  ossia che le due popolazioni normali abbiano stessa varianza σ^2 (**omoschedasticità)**.

  Nel caso precedente, l'intervallo veniva costruito sulla base della quantità

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

  che, se $\sigma_1^2 = \sigma_2^2 = \sigma^2$, diventa

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  Non conosciamo però σ^2 e non possiamo usare questa quantità direttamente, ma possiamo sostituire σ^2 con un suo stimatore.
  Costruiamo la varianza campionaria (corretta) sul primo campione

$$S'^2_1 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$

  e sul secondo campione

$$S'^2_2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2.$$

Sono entrambi stimatori corretti della stessa quantità, σ^2, e possiamo combinare le informazioni sulla variabilità che provengono dai due campioni in questo modo

$$S_p^2 = \frac{(n_1-1)S'^2_1 + (n_2-1)S'^2_2}{n_1+n_2-2}$$

Quello che abbiamo ottenuto è uno stimatore corretto di σ^2, infatti

$$E(S_p^2) = \frac{1}{n_1+n_2-2}\left((n_1-1)E(S'^2_1) + (n_2-1)E(S'^2_2)\right) =$$

$$= \frac{1}{n_1+n_2-2}\left((n_1-1)\sigma^2 + (n_2-1)\sigma^2\right) = \sigma^2$$

Indicheremo con s_p^2 la realizzazione campionaria di S_p^2.

Possiamo ora sostituire a σ^2 il suo stimatore corretto S_p^2, ottenendo

$$\frac{(\bar{X}-\bar{Y}) - (\mu_1-\mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

(risultato non dimostrato).

Allora,

$$P\left(-t_{n_1+n_2-2;1-\alpha/2} \leq \frac{(\bar{X}-\bar{Y}) - (\mu_1-\mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \leq t_{n_1+n_2-2;1-\alpha/2}\right) = 1-\alpha$$

Isolando $\mu_1 - \mu_2$ si deriva

$$P\left((\bar{X}-\bar{Y}) - t_{n_1+n_2-2;1-\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \leq \mu_1-\mu_2 \leq\right.$$

$$(\bar{X} - \bar{Y}) + t_{n_1+n_2-2;1-\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\Bigg) = 1-\alpha$$

e quindi un intervallo di livello $1-\alpha$ per $\mu_1 - \mu_2$ è

$$\left((\bar{X} - \bar{Y}) \pm t_{n_1+n_2-2;1-\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

Il corrispondente intervallo numerico si ottiene sostituendo alle variabili casuali le loro realizzazioni sul campione estratto.

N.B.: Per $n_1 + n_2$ sufficientemente grandi, diciamo $n_1 + n_2 \geq 32$, $t_{n_1+n_2-2;1-\alpha/2}$ può essere approssimato con $z_{1-\alpha/2}$.

Esempio 14

Il Sig. Bianchi va al lavoro con la sua vecchia macchina e ogni giorno segna il tempo X impiegato per arrivare in ufficio. Dopo 25 giorni lavorativi ha ottenuto un totale di

$$\sum_{i=1}^{25} x_i = 625 \text{ minuti} \quad \sum_{i=1}^{25} x_i^2 = 15849 \text{ minuti}^2$$

(a) Stimare il tempo medio e la deviazione standard.

Per stimare il tempo medio possiamo usare la media campionaria, che sappiamo essere una stima corretta,

$$\bar{x} = \frac{625}{25} = 25 \text{ minuti}$$

Per la deviazione standard, possiamo partire da una stima corretta per la varianza, e prendere la radice quadrata del valore così ottenuto:

$$s'^2 = \frac{s^2 25}{24} = \frac{25}{24}\left(\frac{15849}{25} - 25^2\right) = 9,33$$

e quindi

$$s' = \sqrt{9,33} = 3,1 \text{ minuti}$$

è una stima della deviazione standard del tempo impiegato per arrivare in ufficio.

(b) Dare un intervallo di confidenza di livello 95% per il tempo impiegato mediamente dal Sig. Bianchi per arrivare in ufficio sotto l'ipotesi di tempi di

percorrenza normali.

$1-\alpha = 0,95$ $\alpha = 0,025$ $1-\alpha/2 = 0,975$ $t_{24;0,975} = 2,06$

L'intervallo di confidenza per la media del tempo impiegato per andare in ufficio è

$$\left(25 \pm 2,06\frac{3,1}{\sqrt{25}}\right) = (23,72; 26,28) \text{ minuti}$$

(c) Il Sig. Rossi va al lavoro nella sua auto nuova e impiega un tempo Y. Su 50 giorni ha realizzato una media di $\bar{y} = 21$ minuti con deviazione standard di 3 minuti. Dare un intervallo di confidenza al livello 90% per il tempo che mediamente intercorre tra l'arrivo in ufficio del Sig. Bianchi e del Sig. Rossi se partono nello stesso momento.

Si tratta di costruire un intervallo di confidenza per la differenza tra il tempo medio impiegato per arrivare in ufficio dal Sig. Bianchi e dal Sig. Rossi. Per far questo, dobbiamo assumere che per i due soggetti i tempi di percorrenza siano normalmente distribuiti e che abbiano uguale varianza σ^2.

$$1 - \alpha = 0,9 \quad \alpha = 0,1 \quad 1 - \alpha/2 = 0,95$$

$$t_{25+50-2;0,95} \doteq z_{0,95} = 1,64$$

Assumiamo che la deviazione standard specificata dal testo sia la radice quadrata di s'^2 (non di

s^2). Allora,

$$s_p^2 = \frac{24 \cdot 9,33 + 49 \cdot 3^2}{25 + 50 - 2} = 6,15 \text{ minuti}^2$$

L'intervallo cercato è

$$\left((25 - 21) \pm 1,64 \sqrt{6,15 \left(\frac{1}{25} + \frac{1}{50} \right)} \right) = (2,76; 5,24) \text{ minuti}$$

Si noti che l'intervallo contiene solo valori positivi, quindi siamo indotti a concludere che il tempo impiegato mediamente dal Sig. Bianchi per arrivare in ufficio sia maggiore del tempo impiegato mediamente dal Sig. Rossi.

5. INTERVALLO DI CONFIDENZA APPROSSIMATO PER LA DIFFERENZA DI DUE PROBABILITÀ

Consideriamo ora il problema di voler confrontare tra loro due popolazioni bernoulliane, con l'obiettivo di valutare la differenza esistente tra le rispettive probabilità "successo". Ad esempio, un'industria vuole confrontare le frazioni di difettosità nelle popolazioni di pezzi acquistati da due differenti fornitori, per decidere a quali dei due affidarsi per le forniture future.

Supponiamo di estrarre due c.c.s. indipendenti, $x_1, \ldots, x_{n_1}$ dalla prima popolazione e $y_1, \ldots, y_{n_2}$ dalla seconda popolazione, tali che:
$X_1, \ldots, X_{n_1}$ siano i.i.d. con distribuzione $Be(\pi_1)$
$Y_1, \ldots, Y_{n_2}$ siano i.i.d. con distribuzione $Be(\pi_2)$.
Lo scopo è costruire un intervallo di confidenza di livello $1 - \alpha$ per la differenza $\pi_1 - \pi_2$.

Come sappiamo, le frazioni campionarie di "successo"

$$p_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i \quad \text{e} \quad p_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i$$

rappresentano gli stimatori ottimali delle corrispondenti probabilità di popolazione, π_1 e π_2. In aggiunta, per n_1 e n_2 **sufficientemente grandi**,

$$p_1 \dot{\sim} N\left(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1}\right) \quad \text{e} \quad p_2 \dot{\sim} N\left(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2}\right)$$

Per l'indipendenza dei due campioni, la differenza tra le due frazioni campionarie $p_1 - p_2$, è approssimabile con

$$p_1 - p_2 \dot{\sim} \; N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right)$$

e quindi

$$\frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \dot{\sim} \; N(0,1).$$

Pertanto

$$P\left(-z_{1-\alpha/2} \leq \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \leq z_{1-\alpha/2}\right) \doteq 1-\alpha$$

da cui, isolando $\pi_1 - \pi_2$, si deriva l'intervallo

$$\left((p_1 - p_2) \pm z_{1-\alpha/2}\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}\right)$$

Analogamente al caso di una sola popolazione, le frazioni ignote sotto radice quadrata vengono sostituite con i loro stimatori consistenti, p_1 e p_2, ottenendo l'intervallo di livello approssimato $1 - \alpha$

$$\left((p_1 - p_2) \pm z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

Il corrispondente intervallo numerico si ottiene sostituendo a p_1 e p_2 le corrispondenti frequenze campionarie di successo osservate nei due campioni.

Esempio 15
A tre mesi dalle elezioni politiche nazionali in cui la coalizione A è risultata vincitrice con il 54% dei voti, si vuole verificare attraverso un'indagine campionaria l'eventuale consenso verso le politiche adottate inizialmente dal governo. A riguardo sono stati estratti casualmente 100 elettori, classificati in base alla residenza ed all'attuale fiducia nei confronti del governo:

	Nord	Centro	Sud e Isole
Favorevoli	20	16	24
Contrari	14	12	14

Costruire un intervallo di confidenza di livello approssimato 95% per la differenza tra le percentuali di favorevoli all'operato del governo tra gli elettori del nord rispetto a quelli residenti al Sud e nelle Isole.

Se il c.c.s. dal Nord e quello dal Sud e Isole possono considerarsi tra loro indipendenti, si deriva l'intervallo richiesto da

$$p_1 = \frac{20}{34} = 0,59, \qquad p_2 = \frac{24}{38} = 0,63$$

e

$$\left(-0,04 \pm z_{0,975} \sqrt{\frac{0,59(1-0,59)}{34} + \frac{0,63(1-0,63)}{38}} \right)$$

ossia

$$(-0,04 \pm 1,96 \cdot 0,114) = (-0,26\,;\, 0,18)$$

VERIFICA DI IPOTESI

Esempio 1

L'altezza media sul c.c.s. di 8 individui è $\bar{x} = 175$ cm, mentre dai dati provenienti dall'ultimo censimento l'altezza media della popolazione italiana era pari a 173 cm. Alla luce del campione osservato dobbiamo ritenere che l'altezza media della popolazione italiana sia rimasta inalterata, quindi il fatto di aver osservato $\bar{x} = 175 > 173$ è dovuto all'incertezza legata alla stima campionaria (se avessi estratto un c.c.s. diverso sarebbe cambiata la media campionaria, che in generale è diversa da quella di popolazione), oppure è lecito affermare che nella popolazione italiana l'altezza media sia aumentata rispetto all'ultima rilevazione censuaria?

Se vogliamo verificare sulla base del c.c.s. di 8 osservazioni l'ipotesi secondo cui attualmente l'altezza media della popolazione italiana μ sia esattamente uguale a 173 cm, contro l'ipotesi che in realtà $\mu > 173$ cm, dobbiamo tener conto di due fattori:

- la media campionaria $\bar{x}$ osservata è diversa da 173 cm; tuttavia, anche se fosse vero che nella popolazione $\mu = 173$ cm, le medie sui singoli c.c.s. che possiamo estrarre non saranno in generale esattamente uguali a 173 cm
- se veramente $\mu = 173$ cm, allora sappiamo che

$$\bar{X} \sim N\left(173, \frac{\sigma^2}{8}\right)$$

Ci poniamo, quindi, la domanda: la media campionaria osservata $\bar{x} = 175$ cm è sufficientemente distante da 173 cm da far ritenere che l'ipotesi $\mu = 173$ cm sia inverosimile e quindi da rifiutare?

Esempio 2

La frazione campionaria di guarnizioni difettose è risultata pari a $p = 0,1$. Il fornitore sostiene che la vera frazione di guarnizioni difettose nell'intera popolazione è $\pi = 0,05$. La frazione campionaria è sufficientemente più grande di 0,05 per portarci a rifiutare l'ipotesi che $\pi = 0,05$ a favore dell'ipotesi $\pi > 0,05$ e quindi a contestare la fornitura?

LE IPOTESI STATISTICHE

In generale, sia θ il parametro di interesse. Indichiamo con H_0 l'ipotesi da noi formulata, chiamata **IPOTESI NULLA**, che assume la forma

$$H_0 : \theta = \theta_0$$

dove θ_0 è un valore da noi specificato. Indichiamo con H_1 l'ipotesi contrapposta a H_0, chiamata **IPOTESI ALTERNATIVA**. L'ipotesi H_1 può assumere la forma

$$H_1 : \theta \neq \theta_0$$

in questo caso si parla di ipotesi alternativa **BILATERALE** oppure la forma

$$H_1 : \theta > \theta_0 \quad \text{o} \quad H_1 : \theta < \theta_0.$$

Negli ultimi due casi si parla di ipotesi alternativa **UNILATERALE** (rispettivamente, destra e sinistra).

Combinando l'ipotesi nulla e l'ipotesi alternativa si ottiene un **SISTEMA DI IPOTESI**. Nell'Esempio 1, eravamo interessati al sistema di ipotesi

$$\begin{cases} H_0 : & \mu = 173 \\ H_1 : & \mu > 173 \end{cases}$$

mentre nell'Esempio 2

$$\begin{cases} H_0 : & \pi = 0,05 \\ H_1 : & \pi > 0,05 \end{cases}$$

entrambi con ipotesi alternativa unilaterale destra.

DECISIONE IN CONDIZIONI DI INCERTEZZA

Una **VERIFICA DI IPOTESI** è una procedura statistica che permette di utilizzare le informazioni campionarie per saggiare un sistema di ipotesi, ossia per decidere se accettare H_0 (e quindi rifiutare H_1) o se rifiutare H_0 (e quindi accettare H_1).

Ogni volta che decidiamo se accettare o rifiutare H_0 sulla base delle osservazioni campionarie possiamo commettere uno dei seguenti due errori.

- **ERRORE DI I TIPO:** Rifiutiamo H_0 quando H_0 è vera
- **ERRORE DI II TIPO:** Accettiamo H_0 quando H_1 è vera

Si indica con α la probabilità di commettere un errore di I tipo

$$\alpha = P(\text{Rifiuto } H_0 | H_0 \text{ vera})$$

e con β la probabilità di commettere un errore di II tipo

$$\beta = P(\text{Accetto } H_0 | H_1 \text{ vera})$$

Idealmente, vorremmo una procedura di verifica di ipotesi in cui entrambe le probabilità siano basse, ma ciò non è possibile in quanto al diminuire dell'una l'altra aumenta. Si pensi, ad esempio, ad una procedura che accetta H_0 qualunque sia il campione osservato. In questo caso $\alpha = 0$ ma β è evidentemente molto alto. Per contro, se decidiamo di rifiutare sempre H_0, allora $\beta = 0$ ma α è molto alto. Dobbiamo quindi trovare un compromesso.
Poiché H_0 è un'ipotesi che formuliamo noi, che nasce da una nostra congettura, le diamo una posizione privilegiata e ci proteggiamo contro un errore di I tipo, fissando a priori α ad un valore piuttosto piccolo.

Il valore scelto per α viene chiamato **LIVELLO DI SIGNIFICATIVITA'**. Si parla in questo caso di **TEST DI IPOTESI O VERIFICA DI IPOTESI AL LIVELLO DI SIGNIFICATIVITA'** α. Tipicamente, $\alpha \in [0,01; 0,1]$.

PROCEDURE PER LA VERIFICA DI IPOTESI

Le procedure di verifica di ipotesi che vedremo si basano, in generale, sui seguenti passi.

1. Si sceglie il livello di significatività del test α.
2. Si identifica uno stimatore T per θ
3. Se H_0 è vera T tenderà ad assumere valori che sono prossimi a θ_0. Pertanto, se la realizzazione campionaria di T è "sufficientemente distante" da θ_0 (nella direzione indicata da H_1) si deciderà di accettare H_1, altrimenti si accetta H_0.
4. Per stabilire in modo rigoroso che cosa si intende per "sufficientemente distante" da θ_0 si sfrutta il livello di significatività del test.

ESEMPI DI VERIFICA DI IPOTESI

1. VERIFICA DI IPOTESI SULLA MEDIA DI UNA POPOLAZIONE NORMALE

Sia $x_1, x_2, \ldots, x_n$ un c.c.s. da una popolazione normale, ossia X_1, X_2, ..., X_n sono i.i.d. con distribuzione $N(\mu, \sigma^2)$. Vogliamo verificare al livello di significatitività α il sistema di ipotesi

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu > \mu_0 \end{cases}$$

Partiamo dalla stima di μ tramite la media del c.c.s. osservato, $\bar{x}$. Se vale H_0 ci aspettiamo che $\bar{x}$ non si discosti di molto da μ_0; per contro, se vale H_1 ci aspettiamo che $\bar{x}$ sia più grande di μ_0. Allora, sembra ragionevole rifiutare H_0 (e quindi accettare H_1) se

$$\bar{x} - \mu_0 > c$$

dove c è una soglia critica con la quale definiamo di quanto $\bar{x}$ deve essere maggiore di μ_0 per decidere di rifiutare H_0.

Tramite la distribuzione campionaria di $\bar{X}$, la soglia c è univocamente determinata dal livello di significatività α del test. Si impone, infatti, che

$$\alpha = P(\text{Rifiuto } H_0 | H_0 \text{ vera})$$

$$= P(\bar{X} - \mu_0 > c | \mu = \mu_0).$$

Dobbiamo, quindi, scegliere c in modo tale che

$$P(\bar{X} - \mu_0 > c | \mu = \mu_0) = \alpha.$$

- **CASO CON σ^2 NOTA**

  Supponiamo di conoscere σ^2. Sappiamo che, in generale,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

  ma se H_0 è vera, e quindi $\mu = \mu_0$,

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$$

  da cui

$$\frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

  Allora,

$$\alpha = P(\bar{X} - \mu_0 > c | \mu = \mu_0) = P\left(\frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} > \frac{c}{\sqrt{\frac{\sigma^2}{n}}} | \mu = \mu_0\right) =$$

$$= 1 - \Phi\left(\frac{c}{\sqrt{\frac{\sigma^2}{n}}}\right)$$

  e quindi

$$\Phi\left(\frac{c}{\sqrt{\frac{\sigma^2}{n}}}\right) = 1 - \alpha$$

  ossia

$$\frac{c}{\sqrt{\frac{\sigma^2}{n}}} = z_{1-\alpha}$$

  da cui

$$c = z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}}$$

Concludiamo che si deve rifiutare l'ipotesi H_0 e accettare l'ipotesi H_1, al livello di significatività α, se

$$\bar{x} - \mu_0 > z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}}$$

o, equivalentemente, se

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} > z_{1-\alpha}$$

Si accetta invece H_0 e si rifiuta H_1 se

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \leq z_{1-\alpha}$$

La quantità

$$\frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

(a) è chiamata **statistica test** di cui

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

è il valore osservato sul campione.

(b) se sostituiamo a μ_0 il generico μ, coincide con la quantità pivotale usata per costruire l'intervallo di confidenza di livello $1 - \alpha$ per μ quando σ^2 è noto.

Per $H_1 : \mu > \mu_0$ rigettiamo H_0 se

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} > z_{1-\alpha}$$

ossia se si osservano valori elevati della statistica test.
Se $H_1 : \mu < \mu_0$, mentre rimane inalterata H_0, risulta naturale rifiutare H_0 se si osservano valori piccoli della statistica test. Più precisamente, il vincolo sul livello di significatività, impone di rigettare H_0 se

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} < -z_{1-\alpha}$$

Se $H_1 : \mu \neq \mu_0$, è ragionevole rifiutare H_0 se si osservano valori sia elevati che piccoli della statistica test, ossia se

$$\left| \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \right| > c$$

Ancora una volta dobbiamo scegliere la soglia c in modo da rispettare il vincolo sul livello di significatività del test:

$$P\left(\left| \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \right| > c \,|\mu = \mu_0 \right) = \alpha$$

e quindi

$$P\left(\left|\frac{\bar{X}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}}\right| \leq c \,|\mu=\mu_0\right) = 1-\alpha =$$

$$= P\left(-c \leq \frac{\bar{X}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}} \leq c \,|\mu=\mu_0\right)$$

e quindi $c = z_{1-\alpha/2}$.
Allora, per $H_1 : \mu \neq \mu_0$ rifiutiamo H_0 (e accettiamo H_1) se

$$\left|\frac{\bar{x}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}}\right| > z_{1-\alpha/2}$$

mentre accettiamo H_0 (e rifiutiamo H_1) se

$$\left|\frac{\bar{x}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}}\right| \leq z_{1-\alpha/2}$$

Esempio 16
L'ufficio qualità di uno stabilimento che produce pasta alimentare intende controllare se il peso dichiarato nella confezione di 500gr risponda al vero oppure se il processo di confezionamento dà luogo ad un peso medio superiore. Poiché sul processo influisce una pluralità di fattori è ragionevole assumere che il peso di una confezione sia una v.c. normale. Inoltre, da studi precedenti risulta che la varianza della popolazione è 42,5gr^2. In un campione di 25 confezioni, l'ufficio qualità trova che il peso medio è 503,7gr. Usare questi dati per sottoporre a verifica l'ipotesi di interesse dell'ufficio qualità ad un livello $\alpha = 0,01$.

Disponiamo di un c.c.s. di pesi (in gr) $x_1, x_2, \ldots, x_{25}$ da una $N(\mu; 42,5)$. Sappiamo che $\bar{x}$ =503,7gr. Il sistema di ipotesi di interesse è

$$\begin{cases} H_0: & \mu = 500 \\ H_1: & \mu > 500 \end{cases}$$

Il valore osservato della statistica test è

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} = \frac{503,7 - 500}{\sqrt{\frac{42,5}{25}}} = 2,83$$

Rifiutiamo H_0 se 2,83$> z_{1-\alpha} = z_{0,99} = 2,326$. La condizione è soddisfatta quindi possiamo accettare H_1 e concludere che il peso medio delle confezioni μ è maggiore di 500 gr.

Se fossimo stati interessati a verificare se il peso me-

dio è inferiore a 500gr, allora $H_1 : \mu < 500$ e avremmo rifiutato H_0 se 2,83$< -z_{0,99} = -2,326$. La condizione non è soddisfatta, quindi in questo caso H_0 viene accettata. Si intuisce da questo che accettare o rifiutare H_0 dipende anche dall'ipotesi alternativa contemplata. Con il sistema di ipotesi

$$\begin{cases} H_0 : & \mu = 500 \\ H_1 : & \mu < 500 \end{cases}$$

gli unici casi considerati sono $\mu \leq 500$ e all'interno di questi casi, H_0 è più verosimile, in base ai dati raccolti, di H_1.

Se fossimo stati interessati a verificare se il peso medio è diverso da 500gr, allora $H_1 : \mu \neq 500$ e avremmo rifiutato H_0 se $|2,83| = 2,83 > z_{1-\alpha/2} = z_{0,995} = 2,576$. La condizione è soddisfatta quindi avremmo rifiutato H_0 e accettato H_1.

RELAZIONE TRA VERIFICA DI IPOTESI BILATERALE E INTERVALLI DI CONFIDENZA

Si noti che l'intervallo di confidenza di livello $1 - 0,01 = 0,99$ per μ dell'Esempio 16 è in questo caso

$$\left(503,7 \pm 2,576\sqrt{\frac{42,5}{25}}\right) = (500,34; 507,06) \text{ gr}$$

L'intervallo non include il valore 500 gr e questo è sufficiente per rifiutare H_0, a favore di $H_1 : \mu \neq 500$, al livello $0,01$.

Questo ragionamento può essere generalizzato.
Se costruiamo un intervallo di confidenza **di livello** $1-\alpha$ per un parametro θ di interesse e l'intervallo ottenuto non include un valore prefissato θ_0, possiamo immediatamente rifiutare l'ipotesi nulla del sistema

$$\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta \neq \theta_0 \end{cases}$$

al livello α. Viceversa, se l'intervallo di confidenza di livello $1 - \alpha$ per θ include θ_0, allora possiamo accettare H_0 al livello α.

In altri termini, l'intervallo di confidenza di livello $1-\alpha$ per θ include tutti i valori θ_0 per cui accetteremmo l'ipotesi nulla $H_0 : \theta = \theta_0$ al livello α contro l'ipotesi $H_1 : \theta \neq \theta_0$.

IL LIVELLO DI SIGNIFICATIVITA' OSSERVATO (O p–VALUE)

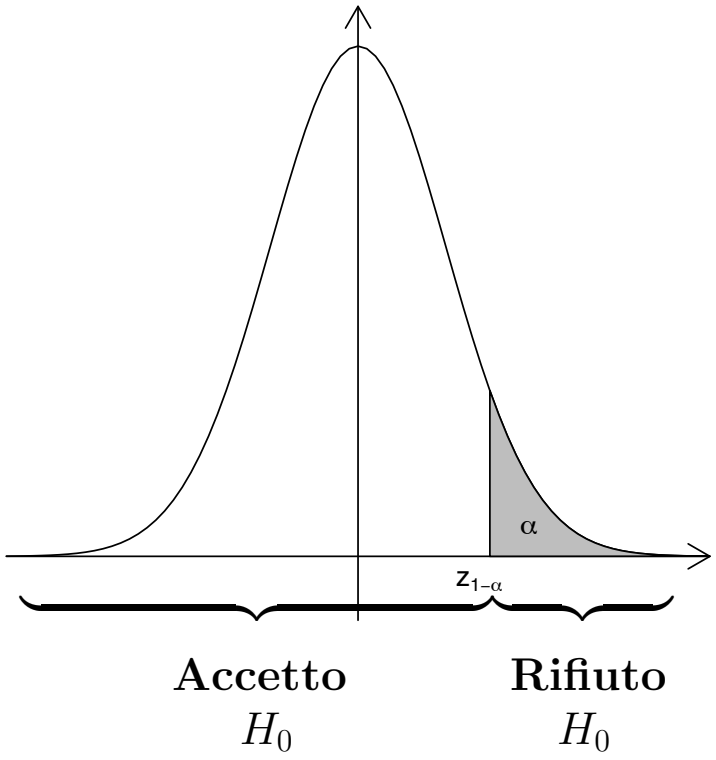
Riprendiamo il sistema di ipotesi

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu > \mu_0 \end{cases}$$

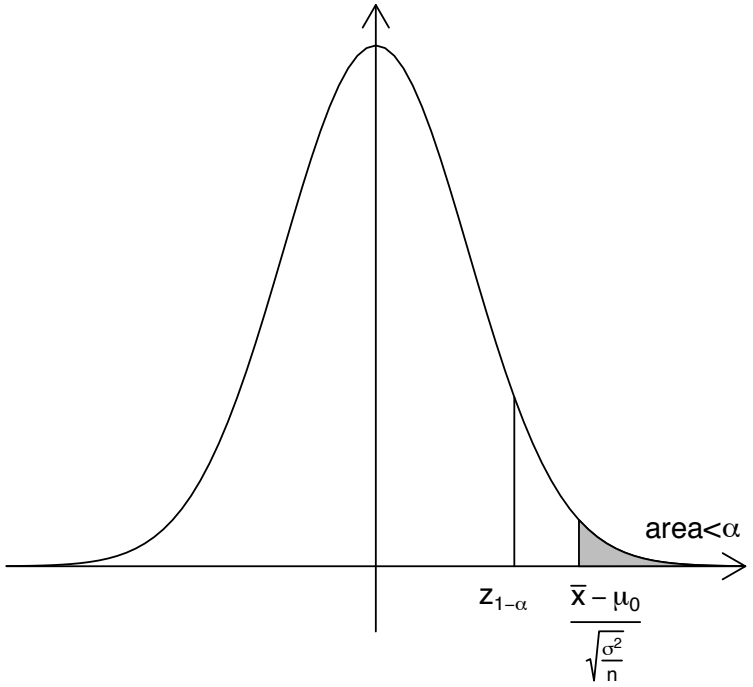
assumendo ancora che σ^2 sia nota. Come visto, si rifiuta H_0 al livello di significatività α se

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} > z_{1-\alpha}$$

ossia, graficamente,

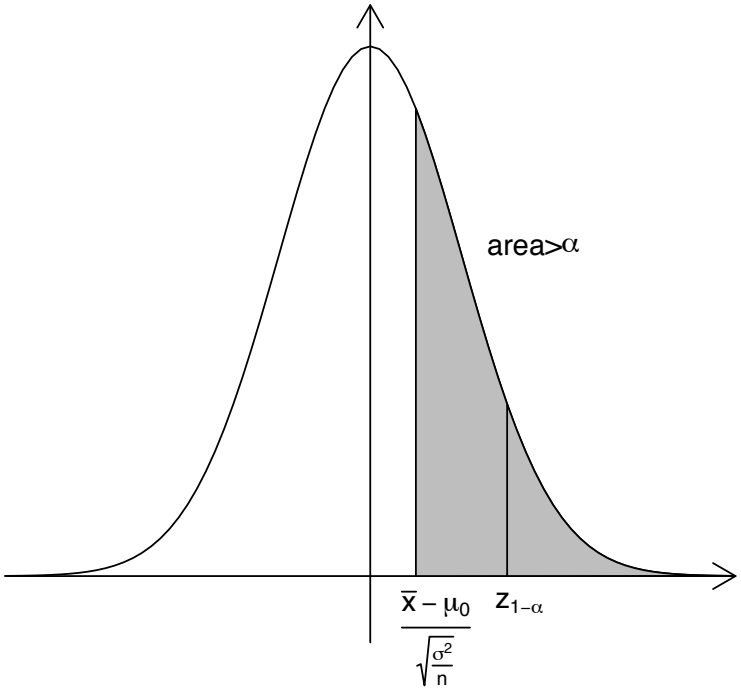


CASO A: RIFIUTO H_0



- il valore osservato della statistica test è superiore alla soglia $z_{1-\alpha}$;
- equivalentemente, l'area della densità della $N(0,1)$ a destra di $\frac{\bar{x}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}}$ è $< \alpha$

CASO B: ACCETTO H_0



- il valore osservato della statistica test è inferiore alla soglia $z_{1-\alpha}$
- equivalentemente, l'area della densità della $N(0,1)$ a destra di $\frac{\bar{x}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}}$ è $> \alpha$

L'area sottesa alla densità della $N(0,1)$ a destra di $\frac{\bar{x}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}}$ è

$$1 - \Phi\left(\frac{\bar{x}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}}\right) = p - \text{value}$$

Vale in generale:

(a) più piccolo è il p–value più H_0 è inverosimile.

(b) più grande è il p–value più i dati danno sostegno ad H_0.

(c) rifiutiamo H_0 al livello di significatività α se p–value$< \alpha$.

(d) accettiamo H_0 al livello di significatività α se p–value$> \alpha$.

Il concetto di p–value può essere esteso a qualsiasi sistema di ipotesi: è sempre dato dalla probabilità sotto H_0 che la statistica test assuma un valore più estremo (nella direzione specificata da H_1) di quello effettivamente osservato. Seppure la formula che restituisce il p–value cambia a seconda del sistema di ipotesi e della distribuzione di riferimento, l'interpretazione rimane quella data dai punti (a)–(d).

I software statistici conducono la verifica di ipotesi producendo come risultato il p–value del test, che dovrà essere interpretato come sopra specificato.

- **CASO CON σ^2 IGNOTA**

  Vogliamo saggiare ancora il sistema di ipotesi

$$\begin{cases} H_0 : & \mu = \mu_0 \\ H_1 : & \mu > \mu_0 \end{cases}$$

  al livello di significatività α, ma ora consideriamo il caso che σ^2 sia ignota.

  Nel caso precedente si rifiutava H_0 se

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} > z_{1-\alpha}$$

  ma ora questa condizione non è direttamente utilizzabile, dato che dipende dall'ignota varianza. Così come fatto per gli intervalli di confidenza, possiamo sostituire σ^2 con la sua stima ottimale s'^2 e rifiutare H_0 a favore di H_1 se

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{s'^2}{n}}} > c$$

  dove c è una costante che dobbiamo determinare in modo da rispettare il vincolo sul livello di significatività. Più precisamente, c deve essere tale che

$$P\left(\frac{\bar{X} - \mu_0}{\sqrt{\frac{S'^2}{n}}} > c | \mu = \mu_0\right) = \alpha$$

o, equivalentemente,

$$P\left(\frac{\bar{X}-\mu_0}{\sqrt{\frac{S'^2}{n}}} \leq c|\mu=\mu_0\right)=1-\alpha$$

Sappiamo che, in generale,

$$\frac{\bar{X}-\mu}{\sqrt{\frac{S'^2}{n}}} \sim t_{n-1}$$

ma se vale H_0, ossia $\mu=\mu_0$, si ha

$$\frac{\bar{X}-\mu_0}{\sqrt{\frac{S'^2}{n}}} \sim t_{n-1}$$

da cui concludiamo che $c = t_{n-1;1-\alpha}$. Allora, rifiutiamo H_0 (e accettiamo H_1) al livello α se

$$\frac{\bar{x}-\mu_0}{\sqrt{\frac{s'^2}{n}}} > t_{n-1;1-\alpha}$$

Se la condizione non è soddisfatta accettiamo H_0.

Se $H_1 : \mu < \mu_0$, allora rifiutiamo H_0 (e accettiamo H_1) al livello α, se è soddisfatta la condizione

$$\frac{\bar{x}-\mu_0}{\sqrt{\frac{s'^2}{n}}} < -t_{n-1;1-\alpha}$$

Se $H_1 : \mu \neq \mu_0$, allora rifiutiamo H_0 (e accettiamo H_1) al livello α, se è soddisfatta la condizione

$$\left|\frac{\bar{x}-\mu_0}{\sqrt{\frac{s'^2}{n}}}\right| > t_{n-1;1-\alpha/2}$$

Esempio 17
L'importo medio delle fatture emesse negli anni passati è di 33 euro. Un campione casuale di 330 fatture emesse quest'anno dalla stessa azienda fa riportare un importo medio di 30 euro. Tramite un test di livello 0,05 stabilire se questo risultato costituisce una prova che l'importo medio delle fatture è diminuito, sapendo che la deviazione standard campionaria (corretta) è di 2,2 euro e che gli importi delle fatture seguono una distribuzione normale.

Abbiamo un c.c.s. $x_1, x_2, \ldots, x_{330}$ di importi di fatture da una $N(\mu, \sigma^2)$, dove μ descrive la media degli importi di tutte le fatture emesse quest'anno e σ^2 la corrispondente varianza. Sappiamo che $\bar{x} = 30$ euro e s'=2,2 euro. Vogliamo verificare il sistema di ipotesi

$$\begin{cases} H_0: & \mu = 33 \\ H_1: & \mu < 33 \end{cases}$$

al livello 0,05.
Il valore osservato della statistica test risulta pari a

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{s'^2}{n}}} = \frac{30 - 33}{\sqrt{\frac{2{,}2^2}{330}}} = -24,47$$

Rifiutiamo H_0 se -24,47$< -t_{n-1;1-\alpha} = -t_{329;0,95} \doteq -z_{0,95} = -1,64$. La condizione è evidentemente soddisfatta, quindi rifiutiamo H_0 al livello 0,05 e accettiamo H_1 (si è verificata una riduzione dell'importo medio delle fatture emesse).

2. VERIFICA DI IPOTESI SU UNA PROBABILITÀ π

Sia $x_1, x_2, \ldots, x_n$ un c.c.s. da una popolazione bernoulliana, ossia $X_1, \ldots, X_n$ sono i.i.d. $Be(\pi)$, con π ignota. Vogliamo verificare il sistema di ipotesi

$$\begin{cases} H_0: & \pi = \pi_0 \\ H_1: & \pi > \pi_0 \end{cases}$$

al livello di significatività α.

È ragionevole rifiutare H_0 se

$$p - \pi_0 > c$$

dove c deve essere tale che

$$P(p - \pi_0 > c | \pi = \pi_0) = \alpha$$

Per n sufficientemente grande, se vale H_0 (ossia $\pi = \pi_0$),

$$p \dot{\sim} \; N\left(\pi_0, \frac{\pi_0(1-\pi_0)}{n}\right)$$

Allora,

$$\alpha = P(p - \pi_0 > c | \pi = \pi_0) =$$

$$P\left(\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} > \frac{c}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} | \pi = \pi_0\right) \doteq 1 - \Phi\left(\frac{c}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\right)$$

e quindi

$$\frac{c}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = z_{1-\alpha}$$

ossia

$$c = z_{1-\alpha}\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

Pertanto, rifiutiamo H_0 al livello **approssimato** α se

$$p - \pi_0 > z_{1-\alpha}\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

ossia se

$$\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} > z_{1-\alpha}$$

ATTENZIONE: Questo test vale solo per n grande (in pratica per $np, n(1-p) \geq 5$).

Se $H_1 : \pi < \pi_0$, si rifiuta H_0 al livello **approssimato** α se

$$\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} < -z_{1-\alpha}$$

Se $H_1 : \pi \neq \pi_0$, si rifiuta H_0 al livello **approssimato** α se

$$\left|\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\right| > z_{1-\alpha/2}$$

Esempio 18
Un partito politico ha ricevuto nelle ultime elezioni il 35% dei voti. Quattro anni dopo, da un sondaggio d'opinione basato su 300 intervistati si è trovato che il 32% degli intervistati ha dichiarato di essere disposto a votare per quel partito. Ci si chiede se, rispetto al risultato elettorale, la situazione del partito è peggiorata.

Abbiamo 300 osservazioni $x_1, \ldots, x_{300}$ da $Be(\pi)$, dove

$$x_i = \begin{cases} 1 & \text{se l'i-esimo intervistato vota per il partito} \\ 0 & \text{se l'i-esimo intervistato non vota per il partito} \end{cases}$$

e π rappresenta la vera frazione di elettori che votano per il partito. Si vuole verificare il sistema di ipotesi

$$\begin{cases} H_0: & \pi = 0,35 \\ H_1: & \pi < 0,35 \end{cases}$$

Il valore osservato della statistica test è

$$\frac{0,32 - 0,35}{\sqrt{\frac{0,35 \cdot 0,65}{300}}} = -1,09$$

Si rifiuta H_0 se -1,09$< -z_{1-\alpha}$. In questo esercizio, però, α non è stato specificato; decidiamo di fissarlo a $\alpha = 0,05$. Rifiutiamo quindi H_0 se $-1,09 < -1,645$. La condizione non è verificata quindi accettiamo H_0 e concludiamo che la situazione del partito non è peggiorata al livello di significatività $\alpha = 0,05$.

3. VERIFICA DI IPOTESI SULLA MEDIA DI UNA POPOLAZIONE NON-NORMALE

Sia $x_1, \ldots, x_n$ un c.c.s. da una popolazione di media m e varianza v^2, ossia $E(X_i) = m$ e $V(X_i) = v^2$ per ogni $i = 1, \ldots, n$.

Si vuole verificare al livello di significatività α il seguente sistema di ipotesi su m:

$$\begin{cases} H_0 : & m = m_0 \\ H_1 : & m > m_0 \end{cases}$$

Come nel caso di una popolazione normale, è ragionevole rifiutare H_0 se

$$\bar{x} - m_0 > c$$

dove c deve essere tale che

$$P(\bar{X} - m_0 > c | m = m_0) = \alpha$$

Per identificare c che soddisfi la condizione precedente, dovremmo conoscere la distribuzione delle X_i che non è qui specificata. Tuttavia, per n sufficientemente grande, qualunque sia la distribuzione delle X_i vale

$$\bar{X} \dot{\sim} \; N\left(m, \frac{v^2}{n}\right)$$

e, sotto H_0 (ossia, per $m = m_0$)

$$\bar{X} \dot{\sim} \; N\left(m_0, \frac{v^2}{n}\right).$$

Pertanto,

$$\alpha = P(\bar{X} - m_0 > c | m = m_0) =$$

$$= P\left(\frac{\bar{X}-m_0}{\sqrt{\frac{v^2}{n}}} > \frac{c}{\sqrt{\frac{v^2}{n}}} | m = m_0\right) \doteq 1-\Phi\left(\frac{c}{\sqrt{\frac{v^2}{n}}}\right)$$

e quindi

$$\frac{c}{\sqrt{\frac{v^2}{n}}} = z_{1-\alpha}$$

ossia

$$c = z_{1-\alpha}\sqrt{\frac{v^2}{n}}$$

Pertanto, rifiutiamo H_0 al livello **approssimato** α se

$$\bar{x} - m_0 > z_{1-\alpha}\sqrt{\frac{v^2}{n}}$$

ossia se

$$\frac{\bar{x}-m_0}{\sqrt{\frac{v^2}{n}}} > z_{1-\alpha}.$$

Se $H_1 : m < m_0$, si rifiuta H_0 al livello **approssimato** α se

$$\frac{\bar{x}-m_0}{\sqrt{\frac{v^2}{n}}} < -z_{1-\alpha}$$

Se $H_1 : m \neq m_0$, si rifiuta H_0 al livello **approssimato** α se

$$\left|\frac{\bar{x}-m_0}{\sqrt{\frac{v^2}{n}}}\right| > z_{1-\alpha/2}$$

N.B.:

- Benché il test precedente abbia la stessa forma di quello per una popolazione normale, al di fuori del caso gaussiano esso vale solo per grandi campioni e il suo livello di significatività è solo approssimativamente pari ad α. La qualità dell'approssimazione migliora con la dimensione campionaria.
- Se v^2 non è nota, può essere sostituita con la sua stima consistente s'^2, ottenendo come valore osservato della statistica test $(\bar{x} - m_0)/\sqrt{\frac{s'^2}{n}}$.

Esempio 12

Riprendiamo l'Esempio 12 e supponiamo di voler verificare il sistema di ipotesi

$$\begin{cases} H_0: & m = 2000 \\ H_1: & m < 2000 \end{cases}$$

al livello di significatività (approssimato) $\alpha = 0,01$. Il valore osservato della statistica test è

$$\frac{1864 - 2000}{\sqrt{\frac{141,61}{196}}} = -160$$

che, se confrontato con $-z_{0,99} = -2,36$, porta a rifiutare H_0 a favore di H_1.

4. VERIFICA DI IPOTESI SULLA DIFFERENZA TRA LE MEDIE DI DUE POPOLAZIONI NORMALI

Ritorniamo al problema del confronto tra due popolazioni normali, già affrontato nella stima intervallare. Abbiamo due c.c.s. indipendenti:
$x_1, x_2, \ldots, x_{n_1}$ realizzazioni di $X_1, X_2, \ldots, X_{n_1}$ i.i.d. $N(\mu_1, \sigma_1^2)$;
$y_1, y_2, \ldots, y_{n_2}$ realizzazioni di $Y_1, Y_2, \ldots, Y_{n_2}$ i.i.d. $N(\mu_2, \sigma_2^2)$.
Vogliamo valutare le differenze tra le due popolazioni, facendo un confronto tra μ_1 e μ_2. In particolare, si vuole sottoporre a verifica il seguente sistema di ipotesi

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 > \mu_2 \end{cases}$$

al livello di significatività α.

- **CASO CON VARIANZE NOTE**

  Assumiamo inizialmente che le due varianze σ_1^2 e σ_2^2 siano note. È ragionevole rifiutare H_0 se

  $$\bar{x} - \bar{y} > c.$$

  Per rispettare il vincolo sul livello di significatività, c deve essere tale che

  $$P(\bar{X} - \bar{Y} > c | \mu_1 = \mu_2) = \alpha$$

  In generale,

  $$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

ma sotto H_0 ($\mu_1 = \mu_2$)

$$\bar{X} - \bar{Y} \sim N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

e quindi, ancora sotto H_0,

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

Allora,

$$\alpha = P(\bar{X} - \bar{Y} > c | \mu_1 = \mu_2) =$$

$$= P\left(\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > \frac{c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} | \mu_1 = \mu_2\right) = 1 - \Phi\left(\frac{c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)$$

e quindi

$$\frac{c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = z_{1-\alpha}$$

ossia

$$c = z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Pertanto, rifiutiamo H_0 al livello α se

$$\bar{x} - \bar{y} > z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

o, equivalentemente, se

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{1-\alpha}$$

Se $H_1 : \mu_1 < \mu_2$, rifiutiamo H_0 al livello α se

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < -z_{1-\alpha}$$

Se $H_1 : \mu_1 \neq \mu_2$, rifiutiamo H_0 al livello α se

$$\left| \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| > z_{1-\alpha/2}$$

Esempio 19
Un ricercatore che lavora alle dipendenze di un'industria produttrice di lampadine elettriche afferma di aver trovato un nuovo tipo di filamento che prolunga la durata delle lampadine. Dato che il nuovo filamento è considerevolmente più costoso di quello attualmente in uso, l'industria intende, prima di adottarlo, avere il conforto di una verifica sperimentale. Viene allora formulata l'ipotesi nulla che la durata media, μ_1, delle lampadine dotate del nuovo filamento sia uguale alla durata media, μ_2, delle lampadine del vecchio tipo, con l'ipotesi alternativa $\mu_1 - \mu_2 > 0$. Per verificare le ipotesi, vengono osservati due campioni dei due tipi di lampadine, entrambi di ampiezza 31. Le medie dei due campioni risultano essere

$$\bar{x} = 1195,16 \text{ ore e } \bar{y} = 1180,05 \text{ ore}$$

Nell'ipotesi che le durate delle lampadine seguano una distribuzione normale con varianza pari a 118,13 per il nuovo filamento e 124,34 per il vecchio filamento si verifichi H_0 contro H_1 al livello $\alpha = 0,01$.

Il valore osservato della statistica test è

$$\frac{1195,16 - 1180,05}{\sqrt{\frac{118,13}{31} + \frac{124,34}{31}}} = 5,4$$

Rifiutiamo l'ipotesi nulla se 5,4> $z_{1-\alpha} = z_{0,99} = 2,326$. La condizione è soddisfatta, per cui concludiamo che il nuovo filamento migliora la qualità delle lampadine rispetto al vecchio.

- **CASO CON VARIANZE IGNOTE**

  Vogliamo verificare al livello α il sistema

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 > \mu_2 \end{cases}$$

  nel caso in cui le due varianze σ_1^2 e σ_2^2 non siano note. Per arrivare ad una soluzione "trattabile" di questo problema dobbiamo assumere (come già fatto nella stima intervallare) che $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (omoschedasticità). Il valore osservato della statistica test nel caso precedente era

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

  che per $\sigma_1^2 = \sigma_2^2 = \sigma^2$ diventa

$$\frac{\bar{x} - \bar{y}}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  Tuttavia, questa quantità non è direttamente utilizzabile, essendo σ^2 ignoto. Come fatto nella stima intervallare, sostituiamo σ^2 con la sua stima s_p^2. È, allora, ragionevole rifiutare H_0 se

$$\frac{\bar{x} - \bar{y}}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} > c,$$

  dove c deve essere tale che

$$P\left(\frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} > c | \mu_1 = \mu_2 \right) = \alpha$$

In generale,

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

ma sotto H_0 $(\mu_1 = \mu_2)$

$$\frac{(\bar{X}-\bar{Y})}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

e quindi $c = t_{n_1+n_2-2;1-\alpha}$.
Rifiutiamo H_0 al livello α se

$$\frac{\bar{x}-\bar{y}}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} > t_{n_1+n_2-2;1-\alpha}$$

Se $H_1 : \mu_1 < \mu_2$, si rifiuta H_0 al livello α se

$$\frac{\bar{x}-\bar{y}}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} < -t_{n_1+n_2-2;1-\alpha}$$

Se $H_1 : \mu_1 \neq \mu_2$, si rifiuta H_0 al livello α se

$$\left|\frac{\bar{x}-\bar{y}}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\right| > t_{n_1+n_2-2;1-\alpha/2}$$

Per $n_1+n_2 \geq 32$ possiamo approssimare $t_{n_1+n_2-2;1-\alpha}$ con $z_{1-\alpha}$.

Esempio 20
Un campione di 20 comuni governati dall'alleanza A mostra che essi spendono una somma media di 87,5€ annue per ciascun contribuente in spese di amministrazione, con una deviazione standard di 12,5€, mentre una simile indagine su un campione di 15 comuni governati dall'alleanza B trova una media di 79€ con deviazione standard campionaria di 15€. E' giustificabile l'ipotesi che non vi sia differenza significativa tra A e B per quanto riguarda le spese comunali di amministrazione?

Indichiamo con μ_1 la spesa media per contribuente nei comuni dell'alleanza A e con μ_2 la spesa media per contribuente nei comuni dell'alleanza B. Vogliamo verificare il sistema di ipotesi

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 \neq \mu_2 \end{cases}$$

Se assumiamo che le spese di amministrazione nei due comuni siano normalmente distribuite con uguale varianza σ^2, allora il valore osservato della statistica test è

$$\left| \frac{87,5 - 79}{\sqrt{s_p^2 \left(\frac{1}{20} + \frac{1}{15}\right)}} \right|$$

Non viene specificato se le due deviazioni standard riportate dal testo siano quelle associate alla varianza corretta oppure no. Supponiamo che siano quelle associate alla varianza campionaria non corretta. In

questo caso,

$$s_p^2 = \frac{20 \cdot 12,5^2 + 15 \cdot 15^2}{20 + 15 - 2} = 196,97$$

Sostituendo 196,97 nell'espressione della statistica test, si deriva il valore 1,77. Se conduciamo il test al livello $\alpha = 0,1$, allora $t_{33;0,95} = 1,6924$ e H_0 verrebbe rifiutata; mentre, se conduciamo la verifica di ipotesi al livello $\alpha = 0,05$, allora $t_{33,0,975} = 2,0345$ e H_0 sarebbe accettata. Questo implica che le osservazioni danno solo una moderata indicazione contro l'ipotesi nulla.

5. VERIFICA DI IPOTESI PER LA DIFFERENZA TRA DUE PROBABILITÀ

Consideriamo ora il problema del confronto tra due popolazioni bernoulliane:
$x_1, \ldots, x_{n_1}$ sono realizzazioni di $X_1, \ldots, X_{n_1}$ i.i.d. $Be(\pi_1)$
$y_1, \ldots, y_{n_2}$ sono realizzazioni di $Y_1, \ldots, Y_{n_2}$ i.i.d. $Be(\pi_2)$ e i due campioni sono tra loro indipendenti. Lo scopo è verificare ad un livello di significatività α il sistema di ipotesi

$$\begin{cases} H_0: & \pi_1 = \pi_2 \\ H_1: & \pi_1 > \pi_2 \end{cases}$$

Per n_1 ed n_2 sufficientemente grandi (e per l'indipendenza dei due campioni), si ha che

$$p_1 - p_2 \dot{\sim} \; N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right)$$

da cui

$$\frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \dot{\sim} \; N(0,1)$$

Sotto H_0, $\pi_1 = \pi_2 = \pi$ e pertanto abbiamo due campioni indipendenti provenienti dalla stessa distribuzione $Be(\pi)$ e quindi

$$\frac{(p_1 - p_2)}{\sqrt{\pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \dot{\sim} \; N(0,1)$$

Dato che la probabilità di successo π comune alle due popolazioni è incognita, risulta naturale stimarla attraverso la quantità

$$\bar{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

che combina le informazioni provenienti dai due campioni.

Pertanto, rifiutiamo H_0 al livello **approssimato** α se

$$\frac{(p_1 - p_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} > z_{1-\alpha}$$

Se $H_1 : \pi_1 < \pi_2$, rifiutiamo H_0 al livello **approssimato** α se

$$\frac{(p_1 - p_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < -z_{1-\alpha}$$

Se $H_1 : \pi_1 \neq \pi_2$, rifiutiamo H_0 al livello **approssimato** α se

$$\left| \frac{(p_1 - p_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| > z_{1-\alpha/2}$$

Esempio 21
La tabella seguente mostra i risultati di un'indagine sui giovani neolaureati, dove un campione di 65 maschi ed uno di 90 femmine sono stati classificati in base alla condizione occupazionale a 6 mesi dalla laurea:

	Numero Disoccupati
Maschi	20
Femmine	18

Verificare ad un livello di significatività del 10% l'ipotesi di uguaglianza della percentuale di disoccupati tra i due sessi, assumendo come alternativa che la disoccupazione incida maggiormente sulla popolazione maschile di neolaureati.

I risultati nella tabella si riferiscono a due campioni tra loro indipendenti: un campione di neolaureati maschi di numerosità $n_1 = 65$ ed un campione femminile dove $n_2 = 90$. Indicando ora con π_1 e π_2 le percentuali di disoccupati rispettivamente nella popolazione maschile e femminile di neolaureati, vogliamo verificare a livello 10% il seguente sistema di ipotesi

$$\begin{cases} H_0: & \pi_1 = \pi_2 \\ H_1: & \pi_1 > \pi_2 \end{cases}$$

Indicando con

$$p_1 = \frac{20}{65} = 0,308 \qquad \text{e} \qquad p_2 = \frac{18}{90} = 0,2$$

le frazioni campionarie di disoccupati dei due sessi, il test asintotico da adottare in questo ambito consiste nel rifiutare H_0 se il valore osservato della statistica test

$$\frac{p_1 - p_2}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

dove

$$\bar{p} = \frac{p_1 \times n_1 + p_2 \times n_2}{n_1 + n_2} = \frac{20 + 18}{155} = 0,245$$

rappresenta una stima della frazione di disoccupazione comune espressa da H_0, risulta superiore al quantile $z_{0,90} = 1,28$ di una normale standard. Dato che nel nostro caso il valore osservato della statistica test risulta pari a 1,54 dobbiamo rifiutare l'ipotesi nulla H_0, concludendo che l'incidenza della disoccupazione nella popolazione maschile di neolaureati è superiore a quella femminile.

6. TEST DI INDIPENDENZA IN UNA TABELLA A DOPPIA ENTRATA

Supponiamo di aver rilevato **su un campione** di n unità estratte casualmente da una popolazione di interesse due variabili X e Y e di aver riassunto in una tabella a doppia entrata le informazioni raccolte sul campione. In statistica descrittiva si è visto che un indice appropriato per misurare il grado di dipendenza tra X e Y è l'indice χ^2

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

dove n_{ij} sono le frequenze osservate e n^*_{ij} le frequenze teoriche sotto l'ipotesi di indipendenza tra le due variabili

$$n^*_{ij} = n_{i.} \cdot n_{.j}/n.$$

Abbiamo anche visto che se $\chi^2 = 0$ le due variabili sono indipendenti e che maggiore è il valore dell'indice maggiore è l'associazione tra X e Y. Dobbiamo però tener conto del fatto che stiamo lavorando solo con un campione di unità estratte dalla popolazione di riferimento e non con l'intera popolazione. Questo significa che, anche se X e Y sono esattamente indipendenti nella popolazione di riferimento, sul campione possiamo osservare un valore di $\chi^2 > 0$. Allora, dato un certo valore dell'indice χ^2 calcolato sul campione casuale estratto, come facciamo a decidere se X e Y sono indipendenti? In altri termini,

vogliamo verificare il sistema di ipotesi

$$\begin{cases} H_0: & X \text{ e } Y \text{ sono indipendenti} \\ H_1: & X \text{ e } Y \text{ non sono indipendenti} \end{cases}$$

al livello di significatività α.

E' ragionevole rifiutare H_0 se

$$\chi^2 > c$$

dove, per il livello di significatività del test, c deve essere tale che

$$P(\chi^2 > c | X \text{ e } Y \text{ sono indipendenti}) = \alpha$$

Si può dimostrare che per n sufficientemente grande, se vale H_0,

$$\chi^2 \dot{\sim} \chi^2_{(r-1)\cdot(s-1)}$$

dove r è il numero di righe della tabella a doppia entrata (il numero di modalità di X) e s è il numero di colonne della tabella a doppia entrata (il numero di modalità di Y). In pratica, si è visto che questa approssimazione è piuttosto buona se $n^*_{ij} \geq 5$.
Deduciamo che

$$c = \chi^2_{(r-1)\cdot(s-1);1-\alpha}$$

Rifiutiamo, quindi, l'ipotesi H_0 di indipendenza tra le due variabili al livello **approssimato** α se

$$\chi^2 > \chi^2_{(r-1)\cdot(s-1);1-\alpha}$$

Se la condizione non è verificata, accettiamo H_0.

Esempio 15
Riprendiamo l'Esempio 15 relativo al consenso verso le politiche del governo. Erano stati estratti casualmente 100 elettori e successivamente classificati in base alla residenza ed all'attuale fiducia nei confronti del governo, ottenendo

	Nord	Centro	Sud e Isole
Favorevoli	20	16	24
Contrari	14	12	14

Verificare ad un livello di significatività dell'1% se vi è o meno indipendenza tra l'atteggiamento nei confronti del governo ed il luogo di residenza.

Per verificare l'ipotesi di indipendenza tra le due variabili al livello $\alpha = 0,01$ utilizziamo il test basato sull'indice χ^2. Innanzitutto calcoliamo le frequenze teoriche n^*_{ij}:

	Nord	Centro	Sud e Isole	Tot
Favorevoli	20,4	16,8	22,8	60
Contrari	13,6	11,2	15,2	40
Tot	34	28	38	100

da cui si ricava il valore $\chi^2 = 0,2727$. Il test prevede di rifiutare l'ipotesi nulla di indipendenza se il valore del χ^2 risulta superiore al valore critico:

$$\chi^2_{(2-1)\times(3-1)\,;\,0,99} = \chi^2_{2\,;\,0,99} = 9,2104.$$

Tale condizione non è verificata e pertanto dobbiamo accettare l'ipotesi di indipendenza tra le variabili.