# INFERENZA STATISTICA

(Camussi, A., Mölle, F., Ottaviano, E., Sari Gorla, M., M*etodi Statistici per la sperimentazione biologica*, Zanichelli, 1995 II ed.
Cicchitelli, G., *Probabilità e Statistica*, Maggioli Editore, 1984 I ed.; 2001 II ed.)

Si riprendano le considerazioni fatte nella parte di Introduzione alla Statistica.

Si vuole studiare una **popolazione**, reale o virtuale, con riferimento a particolari caratteristiche di interesse.

La popolazione viene esaminata in modo parziale, considerando un **campione** di unità statistiche, cioè un aggregato di unità, appartenenti alla popolazione di riferimento, selezionate mediante l'**esperimento di campionamento**.

La **Statistica inferenziale** fornisce strumenti e metodi per ricavare dai dati campionari informazioni sulla popolazione e per quantificare la fiducia da accordare a tali informazioni.

**Esempio.** Si vuole studiare la distribuzione dei redditi delle famiglie italiane, mediante un'indagine campionaria. L'insieme delle famiglie italiane costitusce la *popolazione reale (finita)*, l'insieme ristretto delle famiglie su cui viene condotto lo studio è il *campione*. ◇

**Esempio.** Per studiare l'efficacia di un determinato farmaco si studiano i suoi effetti su un gruppo di pazienti affetti da una certa malattia. Il gruppo di pazienti è il *campione*, ma non risulta ben definita la popolazione di riferimento. Si può pensare ad una *popolazione virtuale (potenzialmente infinita)* costituita dalle potenziali osservazioni ricavabili dalla indefinita replicazione dell'esperimento, nelle stesse condizioni. L'interesse sulla popolazione, in questo caso, è sinonimo di interesse sull'efficacia del farmaco. ◇

**Esempio.** Per effettuare un controllo di qualità si analizzano le caratteristiche di un determinato gruppo di oggetti prodotti da un certo macchinario. Il gruppo di oggetti analizzati è il *campione*, mentre la *popolazione virtuale (potenzialmente infinita)* è costituita da tutti i pezzi che il macchinario può produrre, nelle stesse condizioni. ◇

In questa sede non si considerano le problematiche legate al *campionamento da popolazioni finite*.

Le osservazioni, i dati, riferiti al campione vengono interpretati come *sperimentali* anche se provengono da *osservazione* e il campione viene selezionato da una popolazione reale, finita.

Per la potenziale replicabilità, nelle stesse condizioni, dell'esperimento di campionamento, si suppone che la *popolazione* di riferimento sia *virtuale e potenzialmente infinita*.

L'inferenza statistica studia l'analisi dei dati che costituiscono un **campione casuale**, cioè selezionato mediante un *esperimento casuale* (*aleatorio*).

Nel seguito si considereranno principalmente **campioni casuali semplici** di dimensione $n \geq 1$, che possono venire interpretati come $n$ realizzazioni *indipendenti* di un esperimento di base, nelle *medesime condizioni*.

Dal momento che si considera un esperimento casuale, si coinvolge il **Calcolo delle Probabilità**.

Nell'**inferenza statistica** c'è un *rovesciamento di punto di vista*. Il processo di generazione dei dati sperimentali (modello probabilistico) *non è noto in modo completo*. Il processo in questione è, in definitiva, la popolazione oggetto di indagine.

Con l'inferenza statistica si vogliono ricavare, dai dati campionari, informazioni sulla popolazione (processo di generazione dei dati) e quantificare la attendibilità di tali conclusioni.

I **dati** osservati $x^{oss} = (x_1^{oss}, \ldots, x_n^{oss})$, $n \geq 1$, sono riferiti a caratteristiche di interesse rilevate sulle $n$ unità statistiche, che costituiscono il campione selezionato; in particolare, $x_i^{oss}$, $i = 1, \ldots, n$, indica l'osservazione effettuata sulla $i$-esima unità statistica.

L'*idealizzazione fondamentale* su cui poggia l'inferenza statistica è che $x^{oss}$ è una realizzazione di un **vettore casuale** (variabile casuale multivariata) $X = (X_1, \ldots, X_n)$.

La distribuzione di probabilità di $X$ è, almeno in parte, ignota e si utilizza l'informazione ricavabile dai dati per ottenerne una ricostruzione.

Spesso si assumerà che $x^{oss} = (x_1^{oss}, \ldots, x_n^{oss})$ sia un **campione casuale semplice** (c.c.s.), ossia che il vettore $X = (X_1, \ldots, X_n)$ sia costituito da $n$ *componenti* $X_i$, $i = 1, \ldots, n$, *indipendenti e identicamente distribuite* (i.i.d.).

Nel seguito, si considererà principalmente l'**inferenza statistitica parametrica**: si suppone che la distribuzione di probabilità di $X = (X_1, \ldots, X_n)$ sia nota a meno di una quantità $\theta = (\theta_1, \ldots, \theta_d)$, che è un vettore numerico non noto di dimensione $d \geq 1$.

$\theta$ è detto **parametro** e corrisponde tipicamente a quelli che sono gli *aspetti di interesse*, con riferimento alla popolazione da cui i dati sono tratti come campione casuale.

Un **modello statistico** è una collezione di distribuzioni di probabilità per $X = (X_1, \ldots, X_n)$, che siano compatibili con i dati osservati $x^{oss}$.

Un **modello statistico parametrico** è

- una *collezione di funzioni di probabilità congiunte* $\{f_X(x_1, \ldots, x_n; \theta), \theta \in \Theta\}$ per $X$ indicizzate dal parametro $\theta$, nel **caso discreto**;

- una *collezione di funzioni di densità di probabilità congiunte* $\{f_X(x_1, \ldots, x_n; \theta), \theta \in \Theta\}$ per $X$ indicizzate dal parametro $\theta$, nel **caso continuo**.

Il supporto $S_X$ di $X$ è detto **spazio campionario**; è l'insieme dei possibili campioni $x = (x_1, \ldots, x_n)$, cioè delle possibili relizzazioni di $X = (X_1, \ldots, X_n)$.

L'insieme $\Theta \subseteq \mathbf{R}^d$, $d \geq 1$, è detto **spazio parametrico** e contiene i possibili valori per il parametro $\theta$.

Si suppone che il modello sia correttamente specificato e che esista uno e un solo valore $\theta^0$, detto **vero valore del parametro**.

Nel caso di c.c.s., come conseguenza dell'indipendenza e dell'identica distribuzione delle componenti $X_i$, $i = 1, \ldots, n$,

$$f_X(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta),$$

con $f(\cdot; \theta)$ la funzione di (densità) di probabilità comune a tutte le componenti $X_i$, $i = 1, \ldots, n$.

Utilizzando il campione osservato $x^{oss}$, alla luce del modello statistico parametrico prescelto, si vogliono ricavare informazioni sul parametro ignoto $\theta$, che individua alcuni aspetti di interesse sulla popolazione (processo di generazione dei dati sperimentali o modello probabilistico) di riferimento.

In sostanza, si vuole ricavare informazioni su $\theta_0$, utilizzando i dati $x^{oss}$.

Una buona procedura statistica deve fornire buoni risultati qualsiasi sia il vero valore del parametro $\theta_0$ e deve essere utilizzabile con riferimento ad ogni possibile campione osservato $x^{oss}$.

Per questo motivo, al posto di $\theta_0$ e $x^{oss}$, si adotterà la scrittura $\theta$ e $x = (x_1, \ldots, x_n)$, per indicare un qualsiasi vero valore e un generico campione osservato.

**Esempio.** Per effettuare un controllo di qualità si analizzano $n$ oggetti, scelti a caso tra quelli prodotti da un certo macchinario. Il campione osservato $x = (x_1, \ldots, x_n)$ sarà costituito da una sequenza di valori 0 o 1, che indicano, rispettivamente, se l'oggetto è o non è conforme agli standard di qualità.

Se le $n$ osservazioni sono state effettuate in modo indipendente e nelle medesime condizioni, $X_1, \ldots, X_n$ costituisce un c.c.s.

È ragionevole ipotizzare che le variabili casuali $X_i$, $i = 1, \ldots, n$, seguano una distribuzione bernoulliana o binomiale elementare, con funzione di probabilità $f(x_i; p) = p^{x_i}(1-p)^{1-x_i}$, se $x_i = 0, 1$, e nulla altrove; $p \in (0, 1)$.

Il modello statistico parametrico è dato dalla famiglia delle funzioni di probabilità congiunte

$$f_X(x_1, \ldots, x_n; p) = \prod_{i=1}^{n} f(x_i; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i},$$

con $p \in (0, 1)$.

In questo caso, $\theta = p$ e corrisponde alla probabilità che un singolo oggetto sia difettoso, cioè alla porzione di oggetti difettosi prodotti dal macchinario.

Inoltre, $\Theta = (0, 1)$ e $S_X = \{0, 1\} \times \ldots \times \{0, 1\} = \{0, 1\}^n$.

*Per sintetizzare*, si dice che $X_1, \ldots, X_n$ è un c.c.s. tratto da una popolazione bernoulliana, con parametro $p$ ignoto,

oppure

che il campione $X_1, \ldots, X_n$ è costituito da copie indipendenti di una variabile casuale $Y \sim Ber(p)$, con parametro $p$ ignoto. $\diamond$

**Esempio.** Si misura un determinato oggetto con uno strumento affetto da errore non sistematico. Il campione osservato $x = (x_1, \ldots, x_n)$ sarà costituito da una sequenza di numeri, che corrispondono alle varie misurazioni.

Se le $n$ osservazioni sono state effettuate in modo indipendente e nelle medesime condizioni, $X_1, \ldots, X_n$ costituisce un c.c.s.

È ragionevole ipotizzare che le variabili casuali $X_i$, $i = 1, \ldots, n$, seguano una distribuzione normale di media $\mu$ e varianza $\sigma^2$.

Il modello statistico parametrico è dato dalla famiglia delle funzioni di densità di probabilità congiunte

$$\begin{aligned} f_X(x_1, \ldots, x_n; \mu, \sigma) &= \prod_{i=1}^{n} f(x_i; \mu, \sigma) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{ -\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2} \right\}, \end{aligned}$$

con $\mu \in \mathbf{R}$, $\sigma^2 > 0$.

In questo caso, $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$, dove $\mu$ è la misura vera dell'oggetto in esame e $\sigma^2$ è ricondicibile alla precisione dello strumento di misura.

Inoltre, $\Theta = \mathbf{R} \times \mathbf{R}^+$ e $S_X = \mathbf{R} \times \ldots \times \mathbf{R} = \mathbf{R}^n$.

*Per sintetizzare*, si dice che $X_1, \ldots, X_n$ è un c.c.s. tratto da una popolazione normale con media e/o varianza non note,

oppure

che il campione $X_1, \ldots, X_n$ è costituito da copie indipendenti di una variabile casuale $Y \sim N(\mu, \sigma^2)$, con $\mu$ e/o $\sigma^2$ non noti. $\diamond$

Nell'ambito dell'**inferenza statistica parametrica** si possono individuare *tre classi generali di procedure* che affrontano i seguenti problemi inferenziali, con riferimento al parametro di interesse $\theta$:

- la **stima puntuale**: si vuole ottenere, sulla base dell'osservazione $x$, una congettura puntuale su $\theta$;

- la **stima intervallare** o **regione di confidenza**: si vuole ottenere, sulla base dell'osservazione $x$, un sottoinsieme di $\Theta$ in cui è plausibilmente incluso $\theta$;

- **verifica di ipotesi**: data una congettura o un'ipotesi su $\theta$, si vuole verificare, sulla base dell'osservazione $x$, se essa è accettabile (cioè in accordo con i dati $x$).

## La funzione di verosimiglianza

**Esempio.** Con riferimento all'esempio di pag. 6, si supponga che $n = 40$ e che

$$x^{oss} = (0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,$$
$$0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0);$$

quindi, si rilevano 3 oggetti non conformi agli standard di qualità, su 40 analizzati.

Si può considerare la funzione di probabilità congiunta di $X$ valutata in $x^{oss}$

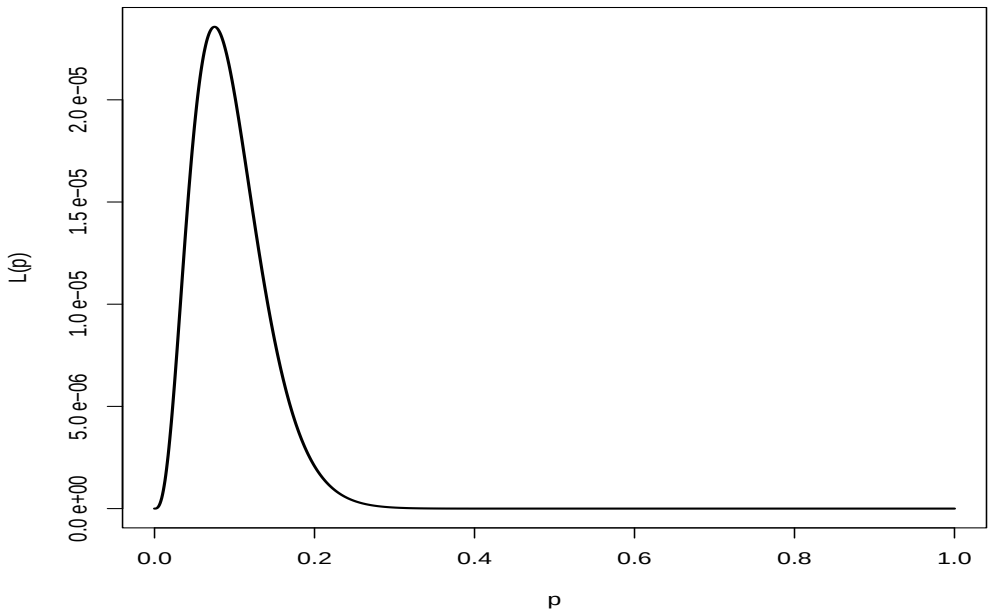
$$f_X(x_1^{oss}, \ldots, x_{40}^{oss}; p) = p^3(1-p)^{37}, \quad p \in (0,1),$$

che corrisponde alla probabilità dell'evento $X = x^{oss}$, sotto $p$, cioè sotto l'ipotesi che $p \in (0,1)$ sia il vero valore del parametro.

Al variare di $p \in (0,1)$ risulta definita la funzione

$$L(p) = L(p; x^{oss}) = p^3(1-p)^{37},$$

che viene chiamata **funzione di verosimiglianza**, il cui grafico è riportato nella figura seguente.




Quindi, la funzione $L(p)$ determina la probabilità dell'evento $X = x^{oss}$, che si è verificato, al variare dei possibili valori per il parametro $p$, nello spazio parametrico $(0, 1)$.

Per valori di $p$ in $(3/10, 1)$, viene assegnata probabilità trascurabile all'evento $X = x^{oss}$, mentre per valori di $p$ vicini a 3/40 (punto di massimo di $L(p)$) la probabilità è elevata.

La funzione $L(p)$ segnala che, alla luce del modello statistico parametrico considerato e del campione osservato $x^{oss}$, i valori più credibili (verosimili) per $p$ sono quelli in prossimità del suo punto di massimo $\widehat{p} = 3/40$. $\diamond$

Le medesime considerazioni si possono fare con modelli statistici parametrici continui, coinvolgendo le funzioni di densità. In questo caso si ha, a meno di una costante di proporzionalità, la probabilità di ottenere una relizzazione per $X$ in un intorno di $x^{oss}$, al variare del valore assunto per il parametro.

**Definizione.** Si consideri un modello statistico parametrico per i dati $x = (x_1, \ldots, x_n)$, riferiti a un campione casuale $X = (X_1, \ldots, X_n)$ con funzione (di densità) di probabilità congiunta $f_X(x_1, \ldots, x_n; \theta)$ e $\theta \in \Theta$ parametro non noto. La funzione $L : \Theta \to \mathbf{R}^+$, nella variabile $\theta$, definita da

$$L(\theta) = L(\theta; x) = f_X(x_1, \ldots, x_n; \theta),$$

è detta **funzione di verosimiglianza** (*likelihood*) di $\theta$ basata sui dati $x$.

$L(\theta)$ **non è** una funzione (di densità) di probabilità.

Alla luce delle osservazioni $x$, $\theta_1$ è più credibile di $\theta_2$, come indicatore del modello generatore dei dati, se

$$L(\theta_1) > L(\theta_2), \quad \text{cioè} \quad \frac{L(\theta_1)}{L(\theta_2)} > 1.$$

Si operano confronti tra coppie di valori per $\theta$ e si valuta la loro adeguatezza relativa.

In pratica, nella definizione di $L(\theta)$, si possono trascurare i fattori moltiplicativi che non dipendono da $\theta$.

Nel caso di c.c.s. costituiti da variabili casuali i.i.d., con funzione (di densità) di probabilità $f(\cdot;\theta)$,

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

Spesso conviene considerare la trasformata logaritmica di $L(\theta)$

$$\ell(\theta) = \ell(\theta; x) = \log L(\theta) = \log f_X(x_1, \ldots, x_n; \theta),$$

chiamata **funzione di log-verosimiglianza**.

L'interpretazione è analoga, con l'unica differenza che i confronti di credibilità tra coppie di valori $\theta_1$ e $\theta_2$ si basano sulle differenze $\ell(\theta_1) - \ell(\theta_2)$.

Nella definizione di $\ell(\theta)$ si possono trascurare le costanti additive che non dipendono da $\theta$.

Nel caso di c.c.s.,

$$\ell(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta).$$

**Esempio.** Con riferimento all'esempio di pag. 8, $\theta = (\mu, \sigma^2)$ e la funzione di log-verosimiglianza è

$$\ell(\theta) = \ell(\mu, \sigma^2; x) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}\right)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2},$$

con dominio $\mathbf{R} \times \mathbf{R}^+$.

Si noti che la costante additiva $-(n/2)\log(2\pi)$ può venire omessa.

Se $\sigma^2$ è nota, si ha la funzione di verosimiglinza per $\mu$; mentre se $\mu$ è noto si ha la funzione di verosimiglinza per $\sigma^2$. $\diamond$

**Esempio.** Se $X \sim Bi(n, p)$ e si osserva l'evento $X = x$, allora, con $p \in (0, 1)$,

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x},$$

$$\ell(p) = \log\left\{\binom{n}{x}\right\} + x\log(p) + (n-x)\log(1-p).$$

Trascurando le costanti moltiplicative e additive, rispettivamente, $L(p)$ e $\ell(p)$ corrispondono alle funzioni che si ottengono nell'esempio di pag. 10. $\diamond$

## Statistiche campionarie e distribuzioni campionarie

Ogni analisi statistica inferenziale è caratterizzata da una componente di incertezza, poiché i dati $x$ sono interpretati come relizzazione di un vettore casuale $X$.

Se si ripete l'esperimento, nelle medesime condizioni, si ottengono dei dati $x'$, tipicamente diversi da $x$.

Ogni inferenza sulla popolazione (sul parametro di interesse) va accompagnata da una valutazione sul suo grado di affidabilità/incertezza.

Nell'effetuare una analisi statistica, i dati $x$ non vengono considerati così come sono ma vengono opportunamente sintetizzati.

**Definizione.** Si chiama **statistica** (**campionaria**) ogni trasformata $T = t(X_1, \ldots, X_n)$, che sintetizza opportunamente il campione casuale $X = (X_1, \ldots, X_n)$.

La scelta della statistica riassuntiva $T$ deve essere fatta in modo accorto, tenedo conto del modello statistico adottato e dello specifico obiettivo dell'inferenza.

Il valore osservato $t = t(x_1, \ldots, x_n)$ di $T$ è un'opportuna sintesi dei dati, utile per l'inferenza su $\theta$.

Se si ripete l'esperimento, nelle medesime condizioni, si ottengono dei dati $x'$, e tipicamente si ha che $t' = t(x') \neq t = t(x)$.

$T$ è una variabile casuale o, in generale, un vettore casuale con una determinata distribuzione di probabilità, chiamata **distribuzione campionaria**.

La bontà di $T$, come statistica riassuntiva per fare inferenza su $\theta$, si può valutare analizzando la sua distribuzione campionaria.

La distribuzione di probabilità di $T$, che è una funzione di $X = (X_1, \ldots, X_n)$, dipende dal parametro ignoto $\theta$.

Quindi, la distribuzione campionaria va intesa *sotto* $\theta$, cioè nell'ipotesi che $\theta$ sia il vero valore del parametro, *qualunque esso sia*.

**Esempio.** Sia $X_1, \ldots, X_n$ un campione casuale. Sono esempi di statistiche campionarie:

- la **media campionaria** $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$;
- le **statistiche ordinate** $X_{(1)} \leq \cdots \leq X_{(n)}$, dove $X_{(i)}$ è la variabile casuale che occupa l'$i$-esima posizione, $X_{(1)} = \min\{X_1, \ldots, X_n\}$ è la **variabile casuale minimo** e $X_{(n)} = \max\{X_1, \ldots, X_n\}$ è la **variabile casuale massimo**;
- $X_{((n+1)/2)}$, se $n$ è dispari, e la coppia $X_{(n/2)}$, $X_{((n/2)+1)}$, se $n$ è pari, che definiscono la **mediana campionaria**;
- $S^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$, chiamata **varianza campionaria**;
- $n^{-1} \sum_{i=1}^n X_i^r$, $n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^r$, $r \in \mathbf{N}^+$;
- $X_{(n)} - X_{(1)}$, $(X_{(1)} + X_{(n)})/2$.

$\diamond$

**Esempio.** Sia $X_1, \ldots, X_n$ un c.c.s. tratto da una popolazione normale.

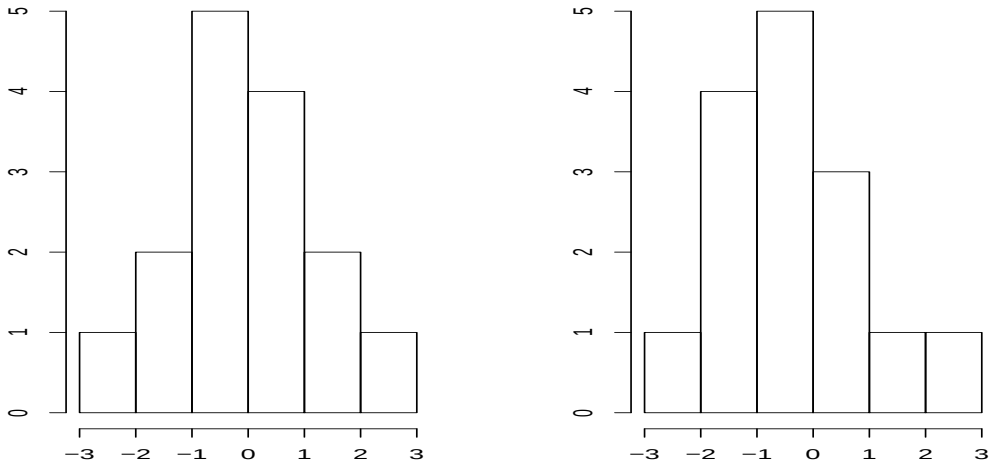
Si ripete per due volte l'esperimento e si osservano i dati

$$x = (-0.89, -0.66, 0.93, 2.42, -2.29, -1.39, -0.86, 0.20,$$
$$1.96, -0.59, -1.36, -0.11, 0.52, 1.17, 0.13),$$
$$x' = (-0.19, -1.52, 2.80, -0.17, -0.30, -0.02, 0.07, 1.69,$$
$$-1.53, -2.74, -1.03, -0.88, 0.21, 0.18, -1.17).$$

Si ha che $x$ e $x'$ sono diversi e si ottengono due realizzazioni diverse per la media campionaria:

$$\frac{1}{15}\sum_{i=1}^{15} x_i \doteq -0.05, \qquad \frac{1}{15}\sum_{i=1}^{15} x'_i \doteq -0.31.$$

Osservando gli istogrammi delle frequenze relative, riferiti ai campioni osservati $x$ e $x'$, si può ragionevolmente confidare che provengono dalla medesima popolazione.




◇

Si considerano in dettaglio alcune statistiche campionarie utilizzate di frequente.

## Media campionaria

Sia $X_1, \ldots, X_n$ un c.c.s. tratto da una popolazione con media $\mu$ e varianza $\sigma^2$.

La variabile casuale media campionaria $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ è tale che

$$E_{\mu,\sigma^2}(\bar{X}_n) = \mu, \quad V_{\mu,\sigma^2}(\bar{X}_n) = \frac{\sigma^2}{n},$$

dove i pedici $\mu, \sigma^2$ indicano che il valor medio e la varianza vengono calcolati nell'ipotesi che $\mu, \sigma^2$ siano i veri valori per i parametri.

Se vengono verificate le ipotesi della Legge debole dei grandi numeri, allora, sotto $\mu, \sigma^2$,

$$\bar{X}_n \xrightarrow{p} \mu.$$

Se vengono verificate le ipotesi del Teorema limite centrale, allora, sotto $\mu, \sigma^2$,

$$\bar{X}_n \dot{\sim} N(\mu, \sigma^2/n),$$

per $n$ sufficientemente elevato.

Nel caso in cui il campione provenga da una popolazione $N(\mu, \sigma^2)$, allora, sotto $\mu, \sigma^2$,

$$\bar{X}_n \sim N(\mu, \sigma^2/n).$$

**Esempio.** Sia $X_1, \ldots, X_n$ un c.c.s. tratto da una popolazione $Ber(p)$. In questo caso, la variabile casuale media campionaria $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ corrisponde alla frequenza relativa di successo ed è tale che

$$E_p(\bar{X}_n) = p, \quad V_p(\bar{X}_n) = \frac{p(1-p)}{n}.$$

Inoltre

$$\bar{X}_n \dot{\sim} N(p, p(1-p)/n),$$

per $n$ sufficientemente elevato. $\diamond$

**Esempio.** Sia $X_1, \ldots, X_n$ un c.c.s. tratto da una popolazione $P(\lambda)$. In questo caso, la variabile casuale media campionaria $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ è tale che

$$E_\lambda(\bar{X}_n) = \lambda, \quad V_\lambda(\bar{X}_n) = \frac{\lambda}{n}.$$

Inoltre

$$\bar{X}_n \dot{\sim} N(\lambda, \lambda/n),$$

per $n$ sufficientemente elevato. $\diamond$

## Varianza campionaria e varianza campionaria corretta

**Definizione** Sia $X_1, \ldots, X_n$ un c.c.s. costituito da variabili casuali con valor medio $\mu$ e varianza $\sigma^2$. La variabile casuale

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

è chiamata **varianza campionaria**.

La varianza campionaria può venire calcolata utilizzando la seguente formula alternativa

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

che ricorda la regola per il calcolo definita per $\sigma^2$.

Si può verificare che

$$E_{\mu,\sigma^2}(S^2) = \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{1}{n}\sigma^2.$$

La statistica campionaria

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

chiamata **varianza campionaria corretta**, è tale che

$$E_{\mu,\sigma^2}(S_c^2) = \frac{n}{n-1} E_{\mu,\sigma^2}(S^2) = \frac{n}{n-1} \frac{n-1}{n} \sigma^2 = \sigma^2.$$

Se vengono verificate le ipotesi della Legge debole dei grandi numeri, si verifica che, sotto $\mu, \sigma^2$,

$$S^2 \xrightarrow{p} \sigma^2, \quad S_c^2 \xrightarrow{p} \sigma^2.$$

Nel caso in cui il c.c.s. $X_1, \ldots, X_n$ sia tratto da una popolazione $N(\mu, \sigma^2)$, si verifica che $\bar{X}_n$ e $S^2$ sono variabili casuali indipendenti ed inoltre

$$\frac{n}{\sigma^2} S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi^2(n-1).$$

In modo analogo,

$$\frac{n-1}{\sigma^2} S_c^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi^2(n-1).$$

**Media campionaria studentizzata e rapporto tra varianze campionarie**

È noto che, se $X_1, \ldots, X_n$ sono copie i.i.d. di una variabile casuale $N(\mu, \sigma^2)$, allora

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

che è la **media campionaria standardizzata**.

Se, invece, al posto di $\sigma$ si considera $S_c = \sqrt{S_c^2}$, allora

$$\frac{\bar{X}_n - \mu}{S_c/\sqrt{n}} \sim t(n-1),$$

dove $t(n-1)$ indica una variabile casuale $t$ di Student con $n-1$ gradi di libertà.

Tale variabile casuale viene chiamata **media campionaria studentizzata**.

Sia $X_1, \ldots, X_n$ un c.c.s., di ampiezza $n$, tratto da una popolazione $N(\mu_X, \sigma_X^2)$ e sia $Y_1, \ldots, Y_m$ un c.c.s., di ampiezza $m$, tratto da una popolazione $N(\mu_Y, \sigma_Y^2)$; i due campioni casuali sono indipendenti.

Si indicano con

$$S_X^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n), \quad S_Y^2 = m^{-1} \sum_{i=1}^{m} (Y_i - \bar{Y}_m)$$

le associate varianze campionarie, che risultano indipendenti.

Poiché $nS_X^2/\sigma_X^2 \sim \chi^2(n-1)$ e $mS_Y^2/\sigma_Y^2 \sim \chi^2(m-1)$, si può verificare che

$$\frac{[nS_X^2/\sigma_X^2]/(n-1)}{[mS_Y^2/\sigma_Y^2]/(m-1)} \sim F(n-1, m-1),$$

dove $F(n-1, m-1)$ indica una variabile casuale $F$ di Fisher con $n-1$ e $m-1$ gradi di libertà.

Si noti che al numeratore si ha una variabile casuale $\chi^2(n-1)$ diviso i suoi gradi di libertà e al denominatore si ha una variabile casuale indipendente $\chi^2(m-1)$ diviso i suoi gradi di libertà.

Se $\sigma_X^2 = \sigma_Y^2$, allora la variabile casuale corrisponde a $[nS_X^2/(n-1)]/[mS_Y^2/(m-1)] = S_{Xc}^2/S_{Yc}^2$.

## La stima puntuale

Le procedure di stima puntuale assegnano, sulla base delle informazioni contenute nel campione osservato, un singolo valore al parametro ignoto della popolazione oggetto di studio.

**Definizione.** Si consideri un modello statistico parametrico per i dati $x = (x_1, \ldots, x_n)$, riferiti a un campione casuale $X = (X_1, \ldots, X_n)$. Il parametro $\theta$ è ignoto. Si consideri un'opportuna applicazione $\widehat{\theta} : S_X \to \Theta$, dallo spazio campionario allo spazio parametrico. Il valore $\widehat{\theta} = \widehat{\theta}(x)$ in $\Theta$, che tale applicazione fa corrispondere ai dati $x$, è detto **stima** di $\theta$. La associata variabile casuale $\widehat{\theta}(X)$ è detta **stimatore** di $\theta$.

Dunque, uno stimatore di $\theta$ è *una opportuna statistica campionaria utilizzata per stimare* $\theta$, mentre la stima di $\theta$ è il suo valore osservato in corrispondenza ai dati $x$.

Si utilizzerà la notazione sintetica $\widehat{\theta}$ sia per lo stimatore che per la stima di $\theta$, poiché il significato appropriato è chiaro dal contesto.

Usualmente, $\theta$ sarà un parametro unidimensionale, e quindi $\widehat{\theta}$ una variabile casuale univariata.

Spesso si utilizzerà la scrittura $\widehat{\theta}_n$ per evidenziare la numerosità $n$ del campione.

Se si ripete l'esperimento, nelle medesime condizioni, si osserva un campione $x'$, usualmente diverso da $x$. La stima basata su $x'$ sarà in genere diversa da quella basata su $x$.

Lo stimatore è una variabile casuale, una statistica campionaria; la *sua distribuzione campionaria sotto $\theta$ è informativa dell'incertezza insita nel procedimento di stima*.

**Esempio.** In un esperimento casuale, si osserva il numero complessivo di successi $x$, in $n$ prove indipendenti con uguale probabilità di successo $p \in (0, 1)$, ignota.

In questo caso, $X \sim Bi(n, p)$, $\theta = p$ e una stima naturale per $p$ è data dalla frequenza relativa di successo $\widehat{p} = x/n$.

Lo stimatore $\widehat{p} = X/n$ è una variabile casuale discreta con supporto $S_{\widehat{p}} = \{0, 1/n, \ldots, (n-1)/n, 1\}$ e funzione di probabilità, sotto $p$, tale che, per $x \in \{0, 1, \ldots, n\}$

$$P_p(\widehat{p} = x/n) = P_p(X/n = x/n) = P_p(X = x)$$
$$= \binom{n}{x} p^x (1-p)^{n-x}.$$

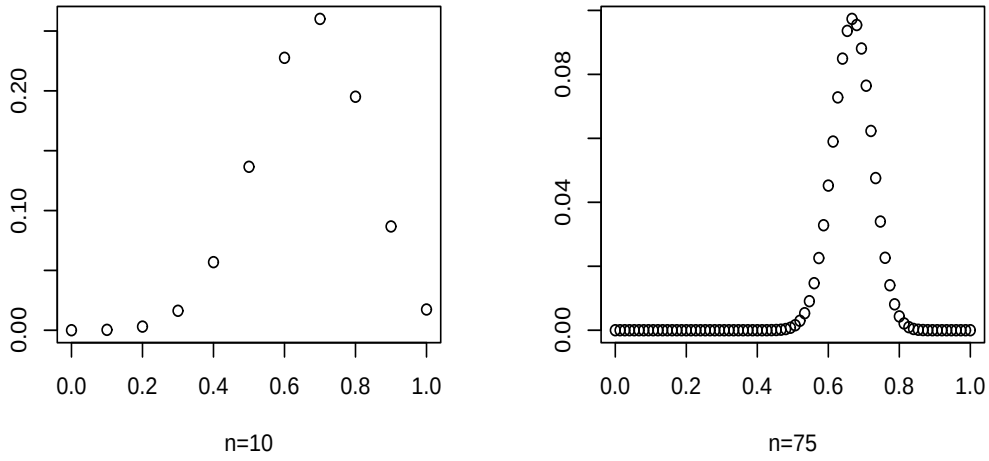
Si verifica facilmente che

$$E_p(\widehat{p}) = p, \quad V_p(\widehat{p}) = \frac{p(1-p)}{n}.$$

Per la Legge debole dei grandi numeri, sotto $p$, $\widehat{p} \xrightarrow{p} p$, per $n \to +\infty$. Per $n$ abbastanza grande, il Teorema limite centrale dà, sotto p,

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

Si riportano i grafici della funzione di probabilità di $\widehat{p}$, per $n = 10$ e $n = 75$, sotto l'ipotesi che $p = 2/3$.




Si noti che le medesime considerazioni si possono fare nel caso in cui si ha un c.c.s. $X = (X_1, \ldots, X_n)$ tratto da una popolazione $Ber(p)$. La frequenza relativa di successo è associata allo stimatore $\widehat{p} = n^{-1} \sum_{i=1}^{n} X_i$, che corrisponde alla media campionaria. $\diamond$

**Esempio.** Si consideri un c.c.s. tratto da una popolazione $N(\mu, \sigma^2)$, con $\mu$ e $\sigma^2$ ignoti.

Si può pensare al caso delle misurazioni ripetute di un determinato oggetto con uno strumento affetto da errore non sistematico.

In questo caso, $\theta = (\mu, \sigma^2)$ e, dato il campione osservato $x = (x_1, \ldots, x_n)$, si può pensare di utilizzare, come ragiovevole stima per $\mu$ e $\sigma^2$, i valori

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

Tali quantità corrispondono, rispettivamente, alle determinazioni osservate delle variabili casuali media campionaria $\bar{X}_n$ e varianza campionaria $S^2$, che sono interpretabili come stimatori per $\mu$ e $\sigma^2$.

Al posto di $S^2$ si può considerare la varianza campionaria corretta $S_c^2$.

Le distribuzioni di probabilità di tali statistiche campionarie sono state analizzate in precedenza. $\diamond$

## Metodi di stima

Nei due esempi considerati in precedenza è adombrato un metodo semplice per ottenere stimatori.

È il **metodo dell'analogia**, in base al quale, per stimare una certa *quantità di popolazione* (ad esempio, un parametro) si utilizza la corrispondente *quantità campionaria* (statistica campionaria).

Ad esempio, un valore atteso si stima con una media campionaria, una varianza con una varianza campionaria (corretta), una covarianza con una covarianza campionaria, ecc.

Se le quantità a cui si applica il metodo dell'analogia sono momenti (ad es. il valore atteso, la varianza, ecc.), si parla di **metodo dei momenti**.

Accanto al metodo dell'analogia è utile considerare il **metodo di sostituzione** (*plug-in*), così specificato.

Sia $\theta$ un parametro per il quale è disponibile una stima (stimatore) $\widehat{\theta}$. Interessa studiare la trasformata $\tau = g(\theta)$, funzione di $\theta$.

La stima (stimatore) per sostituzione di $\tau$ corrisponde a $\widehat{\tau} = g(\widehat{\theta})$, che si ottiene sostituendo in $g(\cdot)$ $\theta$ con $\widehat{\theta}$.

**Esempio.** Si consideri un c.c.s. $X = (X_1, \ldots, X_n)$ tratto da una popolazione $Esp(\lambda)$, con $\lambda$ ignoto.

Dal momento che $\mu = E(X_i) = 1/\lambda$, $i = 1, \ldots, n$, e quindi $\lambda = 1/\mu$, si può pensare di stimare $\lambda$ con $\widehat{\lambda} = 1/\bar{X}_n$, essendo la media campionaria $\bar{X}_n$ uno stimatore per $\mu$.

Si noti che, in generale,

$$E_\lambda(\widehat{\lambda}) = E_\lambda\left(\frac{1}{\bar{X}_n}\right) \neq \frac{1}{E_\lambda(\bar{X}_n)} = \frac{1}{\mu} = \lambda,$$

mentre, per $n \to +\infty$,

$$\widehat{\lambda} = \frac{1}{\bar{X}_n} \xrightarrow{p} \frac{1}{\mu} = \lambda.$$

$\diamond$

## Il metodo della massima verosimiglianza

Ricordando quanto affermato sulla funzione di verosimiglianza, risulta naturale e intuitiva la specificazione del seguente metodo generale per costruire stimatori.

Si consideri un modello statistico parametrico per i dati $x = (x_1, \ldots, x_n)$, con funzione (di densità) di probabilità congiunta $f_X(x_1, \ldots, x_n; \theta)$, $\theta \in \Theta$. Un valore $\widehat{\theta} = \widehat{\theta}(x) \in \Theta$ tale che

$$L(\widehat{\theta}; x) \geq L(\theta; x), \quad \text{per ogni } \theta \in \Theta,$$

è detto **stima di massima verosimiglianza** (s.m.v.) di $\theta$.

Quindi la s.m.v. $\widehat{\theta}$ è punto di massimo della della funzione di verosimiglianza ed è quel valore per $\theta$ in riferimento al quale la probabilità di ottenere una realizzazione per $X$ pari a $x$, o in un intorno di $x$, è massima.

È il valore per $\theta$ che, alla luce dei dati e del modello statistico, risulta più verosimile.

La s.m.v. $\widehat{\theta}$ può essere determinata (spesso in modo più agevole) anche considerando la funzione di log-verosimiglianza $\ell(\theta) = \log L(\theta)$, di cui pure costituisce un massimo.

In generale, non è detto che $\widehat{\theta}$ esista e che sia unico.

Se $\widehat{\theta} = \widehat{\theta}(x)$ esiste ed è unico, la corrispondente variabile casuale $\widehat{\theta} = \widehat{\theta}(X)$ è detta **stimatore di massima verosimiglianza** (s.m.v.) di $\theta$.

Se il modello statistico ha *verosimiglianza regolare*, spesso la s.m.v. $\widehat{\theta} = \widehat{\theta}(x)$ si individua come unica soluzione dell'**equazione di verosimiglianza**, che, con $\theta$ unidimensionale, corrisponde a

$$\frac{d}{d\,\theta}\ell(\theta; x) = 0.$$

La funzione $d\ell(\theta; x)/d\,\theta$ è detta **funzione di punteggio** (score).

**Esempio.** Sia $x$ un'osservazione di $X \sim Bi(n, p)$, con $p \in (0, 1)$ non noto. A meno della costante additiva

$$\ell(p; x) = x \log(p) + (n - x) \log(1 - p)$$

e l'equazione di verosimiglianza corrisponde a

$$\frac{x}{p} - \frac{n - x}{1 - p} = 0.$$

Si ricava facilmente che la s.m.v. è $\widehat{p}(x) = x/n$ e $\widehat{p}(X) = X/n$, in accordo con quanto ottenuto con il metodo dell'analogia.

Se $X = (X_1, \ldots, X_n)$ è un c.c.s. da una popolazione $Ber(p)$, si giunge al medesima conclusione, dove ora $\widehat{p}(X) = n^{-1} \sum_{i=1}^{n} X_i = \bar{X}_n$. $\diamond$

**Esempio.** Sia $X = (X_1, \ldots, X_n)$ un c.c.s. da una popolazione $Geo(p)$. Si ha che

$$\ell(p; x) = n \log(p) + (\sum_{i=1}^{n} x_i - n) \log(1 - p)$$

e l'equazione di verosimiglianza corrsiponde a

$$\frac{n}{p} - \frac{\sum_{i=1}^{n} x_i - n}{1 - p} = 0.$$

Si ricava che lo s.m.v. è $\widehat{p} = n / \sum_{i=1}^{n} X_i = \bar{X}_n^{-1}$. $\diamond$

**Esempio.** Sia $X = (X_1, \ldots, X_n)$ un c.c.s. da una popolazione $N(\mu, \sigma^2)$. Se $\sigma^2$ è nota, si ha che, a meno della costante additiva,

$$\ell(\mu; x) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

e l'equazione di verosimiglianza corrisponde a

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0.$$

Si ricava pertanto che lo s.m.v. è $\widehat{\mu} = n^{-1} \sum_{i=1}^{n} X_i = \bar{X}_n$, in accordo con quanto ottenuto con il metodo dell'analogia.

Se invece $\mu$ è noto, si ha che, a meno della costante additiva,

$$\ell(\sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2},$$

e l'equazione di verosimiglianza corrisponde a

$$-\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^4} = 0.$$

Si ricava pertanto che lo s.m.v. è $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \mu)^2$.

Se sia $\mu$ che $\sigma^2$ sono ignoti, si può verificare che gli s.m.v. sono, rispettivamente, la media campionaria $\bar{X}_n$ e la varianza campionaria $S^2$. $\diamond$

## Proprietà campionarie degli stimatori

Dal momento che uno stimatore è una statistica campionaria (variabile casuale), la sua distribuzione di probabilità sotto $\theta$ sarà informativa sulla bontà del procedimento di stima.

In alcuni casi, come per per la media campionaria e altre statistiche campionarie di uso comune, la distribuzione di probabilità è nota in modo esatto o approssimato.

L'obiettivo della stima *non è l'esattezza*, *ma l'accuratezza*, ossia che l'errore di stima sia usualmente piccolo, al variare del campione osservato.

Come misura della precisione di uno stimatore $\widehat{\theta}$ si può considerare, se $\theta$ è unidimensionale, l'**errore quadratico medio di stima** (*standard error*) sotto $\theta$,

$$se_\theta(\widehat{\theta}) = \sqrt{E_\theta(\widehat{\theta} - \theta)^2}.$$

Non è sensato ricercare lo stimatore tale che $se_\theta(\widehat{\theta}) = 0$ per ogni $\theta \in \Theta$, che esiste solo in casi banali e non interessanti.

Dati due stimatori $\widehat{\theta}_1$ e $\widehat{\theta}_2$ per $\theta$, si preferisce $\widehat{\theta}_1$ se $se_\theta(\widehat{\theta}_1) \leq se_\theta(\widehat{\theta}_2)$ per ogni $\theta \in \Theta$, cioé se ha standard error uniformemente più piccolo.

È raro che esista uno stimatore con standard error uniformemente minimo. È possibile individuarlo in alcune classi particolari di stimatori.

Poiché $se_\theta(\widehat{\theta})$ dipende in genere da $\theta$, che è ignoto, viene usualmente stimato con l'**errore quadratico medio di stima stimato** (*estimated standard error*),

$$\widehat{se}(\widehat{\theta}) = se_{\widehat{\theta}}(\widehat{\theta}) = \sqrt{E_\theta(\widehat{\theta} - \theta)^2}\,|_{\theta=\widehat{\theta}}\,.$$

Si ottiene sostituendo in $se_\theta(\widehat{\theta})$ al posto di $\theta$ la stima $\widehat{\theta}$.

Si verifica facilmente che

$$E_\theta(\widehat{\theta} - \theta)^2 = V_\theta(\widehat{\theta}) + [E_\theta(\widehat{\theta}) - \theta]^2,$$

dove la differenza $[E_\theta(\widehat{\theta}) - \theta]$ corrisponde alla **distorsione** dello stimatore.

Un stimatore $\widehat{\theta}$ è detto **non distorto** se

$$E_\theta(\widehat{\theta}) = \theta, \quad \text{per ogni } \theta \in \Theta.$$

In questo caso, $se_\theta(\widehat{\theta}) = \sqrt{V_\theta(\widehat{\theta})}$, che corrisponde allo scarto quadratico medio di $\widehat{\theta}$.

In alcuni contesti si riesce a individuare uno stimatore **efficiente fra i non distorti**, cioé uno stimatore non distorto che presenta standard error, e quindi varianza, uniformemente minima fra tutti i possibili stimatori non distorti per $\theta$.

Se uno stimatore con forte distorsione è di fatto inutile, perché presenta nella generalità dei campioni un errore sistematico, può bastare, in pratica, la seguente richiesta più tenue della non distorsione.

Uno stimatore $\widehat{\theta}$ è detto **asintoticamente non distorto** se la successione degli stimatori $\widehat{\theta}_n$, $n \in \mathbf{N}^+$, basati sui campioni $X_1, \ldots, X_n$, $n \in \mathbf{N}^+$, è tale che

$$\lim_{n \to \infty} E_\theta(\widehat{\theta}_n) = \theta, \quad \text{per ogni } \theta \in \Theta.$$

Un'ulteriore ragionevole richiesta di natura asintotica è che, per $n$ sufficientemente elevato, lo stimatore $\widehat{\theta}_n$ produca relizzazioni vicine al vero valore del parametro con elevata probabilità.

Più precisamente, uno stimatore $\widehat{\theta}$ è detto **consistente** per $\theta$ se la successione degli stimatori $\widehat{\theta}_n$, $n \in \mathbf{N}^+$, basati sui campioni $X_1, \ldots, X_n$, $n \in \mathbf{N}^+$, è tale che, sotto $\theta$,

$$\widehat{\theta}_n \xrightarrow{p} \theta, \quad \text{per } n \to \infty.$$

Una applicazione della condizione sufficiente per la convergenza in probabilità a una costante assicura che, se uno stimatore è asintoticamente non distorto e tale che $\lim_{n \to \infty} V_\theta(\widehat{\theta}_n) = 0$, allora è consistente.

**Esempio.** Sia $X = (X_1, \ldots, X_n)$, $n > 2$, un c.c.s. da una popolazione con media $\mu$ e varianza $\sigma^2$. Si confrontano i seguenti stimatori per $\mu$

$$\widehat{\mu}_1 = \bar{X}_n, \quad \widehat{\mu}_2 = \frac{X_1 + X_2}{2}.$$

Poiché $E_{\mu,\sigma^2}(\widehat{\mu}_1) = E_{\mu,\sigma^2}(\widehat{\mu}_2) = \mu$, sono entrambi non distorti, ma $\widehat{\mu}_1$ è più efficiente di $\widehat{\mu}_2$ essendo

$$V_{\mu,\sigma^2}(\widehat{\mu}_1) = \frac{\sigma^2}{n} \leq \frac{\sigma^2}{2} = V_{\mu,\sigma^2}(\widehat{\mu}_2).$$

Inoltre, $\bar{X}_n$ è consistente. Lo standard error di $\bar{X}_n$ è $se_{\mu,\sigma^2}(\bar{X}_n) = \sigma/\sqrt{n}$, mentre lo standard error stimato è $\widehat{se}(\bar{X}_n) = \widehat{\sigma}/\sqrt{n}$, con $\widehat{\sigma}$ un'opportuna stima per $\sigma$.

Come stimatori per $\sigma^2$ si possono considerare la varianza campionaria $S^2$ e la varianza campionaria corretta $S_c^2$. Ricordando quanto detto in precedenza, $S^2$ è distorto, ma asintoticamente non distorto, mentre $S_c^2$ è non distorto. Entrambi soddisfano la proprietà della consistenza. $\diamond$

**Esempio.** Sia $X = (X_1, \ldots, X_n)$, $n > 2$, un c.c.s. da una popolazione $U(0, \theta)$, con $\theta$ ignoto.

Poiché $X_i \sim U(0, \theta)$, $i = 1, \ldots, n$, $E_\theta(X_i) = \theta/2 = \mu$ e $V_\theta(X_i) = \theta^2/12 = \sigma^2$.

Applicando il metodo di sostituzione, si specifica lo stimatore $\widehat{\theta} = 2\bar{X}_n$. Tale stimatore è non distorto e consistente, infatti

$$E_\theta(\widehat{\theta}) = 2E_\theta(\bar{X}_n) = 2\,\mu = 2\,\frac{\theta}{2} = \theta,$$

$$V_\theta(\widehat{\theta}) = 4V_\theta(\bar{X}_n) = 4\,\frac{\sigma^2}{n} = 4\,\frac{\theta^2}{12\,n} = \frac{\theta^2}{3n}.$$

Inoltre, per la non distorsione, lo standard error corrisponde a $se_\theta(\widehat{\theta}) = \sqrt{V_\theta(\widehat{\theta})} = \theta/\sqrt{3n}$. $\diamond$

**Esempio.** Sia $X = (X_1, \ldots, X_n)$, $n > 2$, un c.c.s. da una popolazione $Bi(m, p)$, con $p$ ignoto.

Poiché $X_i \sim Bi(m, p)$, $i = 1, \ldots, n$, $E_p(X_i) = mp = \mu$ e $V_p(X_i) = mp(1-p) = \sigma^2$.

Applicando il metodo di sostituzione, si specifica lo stimatore $\widehat{p} = \bar{X}_n/m = \sum_{i=1}^{n} X_i/(nm)$. Tale stimatore è non distorto e consistente, infatti

$$E_p(\widehat{p}) = \frac{1}{m} E_p(\bar{X}_n) = \frac{1}{m}\,\mu = \frac{1}{m}\,mp = p,$$

$$V_p(\widehat{p}) = \frac{1}{m^2} V_p(\bar{X}_n) = \frac{1}{m^2}\,\frac{\sigma^2}{n} = \frac{1}{m^2}\,\frac{mp(1-p)}{n} = \frac{p(1-p)}{mn}.$$

Inoltre, per la non distorsione, lo standard error corrisponde a $se_p(\widehat{p}) = \sqrt{V_p(\widehat{p}} = \sqrt{p(1-p)/(mn)}$.

Se $m = 1$ si ottiene il caso particolare di un c.c.s. da una popolazione $Ber(p)$. $\diamond$

## Intervalli di confidenza

Con le procedure di stima puntuale si ottiene un valore di stima che quasi certamente non coincide con il vero e ignoto valore del parametro $\theta$.

Con la stima intevallare o intervalli di confidenza si cerca di incorporare nel procedimento di stima una misura di accuratezza.

Il parametro ignoto non viene stimato con un punto, ma con un sottoinsieme più ampio dello spazio parametrico, in genere un intervallo.

**Esempio.** Sia $X = (X_1, \ldots, X_5)$ un c.c.s. da una popolazione $N(\mu, 16)$, con $\mu$ ignoto.

Dal momento che si è interessati a fare inferenza su $\mu$, si considera, come statistica campionaria, la media campionaria $\bar{X}_5 \sim N(\mu, 16/5)$.

Poiché la media campionaria standardizzata

$$\frac{\bar{X}_5 - \mu}{4/\sqrt{5}} \sim N(0, 1)$$

ha distribuzione di probabilità che non dipende dal parametro ignoto $\mu$ (**quantità pivotale**), ricordando quanto detto sui quantili di $N(0, 1)$, si ha che, per ogni $\mu \in \mathbf{R}$,

$$P_\mu \left( -1.96 \leq \frac{\bar{X}_5 - \mu}{4/\sqrt{5}} \leq 1.96 \right) = 0.95.$$

Se, nelle due disequazioni che specificano l'evento, si eplicita il parametro $\mu$ si ottiene che, per ogni $\mu \in \mathbf{R}$,

$$P_\mu\left(\bar{X}_5 - 1.96\frac{4}{\sqrt{5}} \leq \mu \leq \bar{X}_5 + 1.96\frac{4}{\sqrt{5}}\right) = 0.95.$$

Risultano individuate due statistiche campionarie che definiscono l'intervallo casuale (aleatorio)

$$\left[\bar{X}_5 - 1.96\frac{4}{\sqrt{5}}, \bar{X}_5 + 1.96\frac{4}{\sqrt{5}}\right],$$

che contiene il vero valore del parametro $\mu$, qualunqhe esso sia, con probabilità 0.95. Tale intervallo è detto **intervallo di confidenza** per $\mu$, con **livello di confidenza** 0.95.

Se si osserva $x^{oss} = (169, 171, 174, 177, 179)$, si ottiene che $\bar{x}_5 = 174$ e l'intervallo di confidenza osservato, con livello 0.95, è $[170.49, 177.51]$.

È sbagliato affermare che $[170.49, 177.51]$ contiene $\mu$ con probabilità 0.95.

Il risultato si interpreta nel seguente modo: poiché l'intervallo casuale contiene $\mu$ con probabilità 0.95, si ha fiducia di aver osservato un campione $x$ a cui che corrisponde un intervallo che contiene $\mu$. Si sa che questo accade, mediamente, 95 volte su 100. $\diamond$

Si considera il caso in cui il parametro ignoto $\theta$ è unidimensionale. Si parlerà di intervalli di confidenza anche se, in generale, si possono definire regioni di confidenza che non corrsipondono a intevalli.

**Definizione.** Si consideri un modello statistico parametrico per i dati $x = (x_1, \ldots, x_n)$, riferiti a un campione casuale $X = (X_1, \ldots, X_n)$. Il parametro $\theta$ è ignoto. Si considerino due opportune statistiche campionarie $T_1 = t_1(X)$ e $T_2 = t_2(X)$. Se, fissato un livello di confidenza $1 - \alpha$, $\alpha \in (0, 1)$, si ha che, per ogni $\theta \in \Theta$,

$$P_\theta \left( T_1 \leq \theta \leq T_2 \right) = 1 - \alpha,$$

$[T_1, T_2]$ definisce un **intervallo di confidenza** per $\theta$, con **livello di confidenza** $1 - \alpha$.

Le statistiche $T_1$ e $T_2$ si chiamano limite inferiore e limite superiore di confidenza, mentre $T_2 - T_1$ specifica la lunghezza dell'intevallo.

In corrispondenza ai dati $x$, si determina l'intervallo di confidenza (osservato) $[t_1(x), t_2(x)]$ per $\theta$, con livello di confidenza $1 - \alpha$. La sua interpretazione è analoga a quella fornita nell'esempio precedente.

Nel seguito, non si svilupperà la teoria delle regioni di confidenza. Si forniranno delle indicazioni pratiche, utili per la costruzione di intervalli di confidenza per particolari problemi statistici.

La procedura che verrà usualmente adottata per definire intervalli di confidenza per $\theta$ richiede

- la definizione di una opportuna statistica campionaria per fare inferenza su $\theta$ (in genere uno stimatore);

- la costruzione, a partire da tale statistica, di una **quantità pivotale**, la cui distribuzione di probabilità non dipende dal parametro ignoto $\theta$.

In alcuni situazioni, si definiranno quantità pivotali *approssimate*, e quindi intervalli di confidenza con livello di confidenza *approssimato* $1 - \alpha$.

**Intervalli di confidenza per la media di una popolazione normale**

Sia $X = (X_1, \ldots, X_n)$ un c.c.s. da una popolazione $N(\mu, \sigma^2)$. Si vuole determinare un intervallo di confidenza per $\mu$ con livello $1 - \alpha$.

Si considerano due casi.

(1) **varianza $\sigma^2$ nota**.

Come statistica campionaria, si considera la **media campionaria** $\bar{X}_n$ e, come quantità pivotale, la **media campionaria standardizzata** $\sqrt{n}(\bar{X}_n - \mu)/\sigma \sim N(0,1)$.

Per ogni $\mu \in \mathbf{R}$,

$$P_\mu\left(-z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right)$$

$$= P_\mu\left(\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

con $z_{\alpha/2}$ tale che $P(Z \geq z_{\alpha/2}) = \alpha/2$, dove $Z \sim N(0,1)$.

Quindi

$$\left[\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

è un intervallo di confidenza per $\mu$ con livello $1 - \alpha$.

(2) **varianza $\sigma^2$ ignota**.

Come statistica campionaria, si considera la **media campionaria** $\bar{X}_n$ e, come quantità pivotale, la **media campionaria studentizzata** $\sqrt{n}(\bar{X}_n - \mu)/S_c \sim t(n-1)$.

Per ogni $\mu \in \mathbf{R}$,

$$P_\mu\left(-t_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{S_c/\sqrt{n}} \leq t_{\alpha/2}\right)$$

$$= P_\mu\left(\bar{X}_n - t_{\alpha/2}\frac{S_c}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{\alpha/2}\frac{S_c}{\sqrt{n}}\right) = 1-\alpha,$$

con $t_{\alpha/2}$ tale che $P(T \geq t_{\alpha/2}) = \alpha/2$, dove $T \sim t(n-1)$.

Quindi

$$\left[\bar{X}_n - t_{\alpha/2}\frac{S_c}{\sqrt{n}}, \bar{X}_n + t_{\alpha/2}\frac{S_c}{\sqrt{n}}\right]$$

è un intervallo di confidenza per $\mu$ con livello $1-\alpha$.

**Esempio.** Sia $X_1, \ldots, X_{50}$ un c.c.s. di dimensione $n = 50$ da una popolazione normale con $\mu$ ignoto e $\sigma^2 = 2$.

Nell'ipotesi che i risultati dell'indagine campionaria siano tali che $\sum_{i=1}^{50} x_i = 94.15$, si fornisca un intervallo di confidenza per $\mu$ con livello $1 - \alpha = 0.9$.

Poiché $\bar{x}_{50} = \sum_{i=1}^{50} x_i/50 = 1.883$, $\alpha/2 = 0.05$ e $z_{0.05} = 1.645$, sulla base dei risultati di pag. 44, si conclude che

$$\left[1.883 - 1.645\,\frac{\sqrt{2}}{\sqrt{50}}, 1.883 + 1.645\,\frac{\sqrt{2}}{\sqrt{50}}\right] = [1.554, 2.212]$$

è l'intervallo di confidenza cercato. $\diamond$

**Esempio.** Sia $X_1, \ldots, X_{30}$ un c.c.s. di dimensione $n = 30$ da una popolazione normale con $\mu$ e $\sigma^2$ ignoti.

Nell'ipotesi che i risultati dell'indagine campionaria siano tali che $\sum_{i=1}^{30} x_i = 36.76$ e $\sum_{i=1}^{30} x_i^2 = 102.27$, si fornisca un intervallo di confidenza per $\mu$ con livello $1-\alpha = 0.95$.

In questo caso, $\bar{x}_{30} = \sum_{i=1}^{30} x_i/30 = 1.225$,

$$s^2 = \frac{1}{30}\sum_{i=1}^{30} x_i^2 - \bar{x}_{30}^2 = 1.908, \quad s_c^2 = \frac{30}{29}s^2 = 1.973.$$

Inoltre, $\alpha/2 = 0.025$ e, considerando un distribuzione $t$ di Student con 29 gradi di libertà, $t_{0.025} = 2.045$. Sulla base dei risultati di pag. 44, si conclude che

$$\left[1.225 - 2.045\,\frac{\sqrt{1.973}}{\sqrt{30}}, 1.225 + 2.045\,\frac{\sqrt{1.973}}{\sqrt{30}}\right]$$

$$= [0.701, 1.749]$$

è l'intervallo di confidenza cercato. $\diamond$

## Intervalli di confidenza per la media di una popolazione non normale

Sia $X = (X_1, \ldots, X_n)$ un c.c.s. da una popolazione non normale con media $\mu$ e varianza $\sigma^2$ ignote. Si vuole determinare un intervallo di confidenza per $\mu$ con livello $1 - \alpha$

Se la numerosità campionaria $n$ è sufficientemente elevata, si riesce a determinare, con relativa facilità, intervallo di confidenza per $\mu$ con **livello** $1 - \alpha$ **approssimato**.

Come statistica campionaria, si considera la **media campionaria** $\bar{X}_n$ e, come *quantità pivotale approssimata*, la **media campionaria standardizzata** $\sqrt{n}(\bar{X}_n - \mu)/\sigma \dot{\sim} N(0, 1)$.

Dal momento che $n$ è elevato, si può sostituire a $\sigma$, che in genere non è noto, un'opportuno stimatore consistente $\widehat{\sigma}$.

Per ogni $\mu \in \mathbf{R}$,

$$P_\mu \left( -z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\widehat{\sigma}/\sqrt{n}} \leq z_{\alpha/2} \right)$$

$$= P_\mu \left( \bar{X}_n - z_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}} \right) \doteq 1 - \alpha,$$

con $z_{\alpha/2}$ tale che $P(Z \geq z_{\alpha/2}) = \alpha/2$, dove $Z \sim N(0, 1)$.

Quindi

$$\left[\bar{X}_n - z_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}\right]$$

è un intervallo di confidenza per $\mu$ con livello di confidenza approssimato $1-\alpha$.

Si evidenziano i seguenti casi.

(1) Sia $X = (X_1, \ldots, X_n)$ un c.c.s. da una popolazione $Ber(p)$, con $p$ ignoto. Si suppone che $n$ sia sufficientemente elevato.

Poiché, $\mu = p$, $\sigma^2 = p(1-p)$ e $\widehat{p} = \bar{X}_n$,

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

è un intervallo di confidenza per $p$ con livello di confidenza approssimato $1-\alpha$.

(2) Sia $X = (X_1, \ldots, X_n)$ un c.c.s. da una popolazione $P(\lambda)$, con $\lambda$ ignoto. Si suppone che $n$ sia sufficientemente elevato.

Poiché, $\mu = \lambda$, $\sigma^2 = \lambda$ e $\widehat{\lambda} = \bar{X}_n$,

$$\left[\widehat{\lambda} - z_{\alpha/2}\frac{\sqrt{\widehat{\lambda}}}{\sqrt{n}}, \widehat{\lambda} + z_{\alpha/2}\frac{\sqrt{\widehat{\lambda}}}{\sqrt{n}}\right]$$

è un intervallo di confidenza per $\lambda$ con livello di confidenza approssimato $1-\alpha$.

**Esempio.** Si vuole studiare l'efficacia di un farmaco per curare una determinata malattia. Si effettua una sperimentazione su 550 pazienti affetti dalla patologia e si riscontra che in farmaco è efficace in 393 casi.

Si vuole determinare un intervallo di confidenza con livello 0.95 per la frequenza relativa $p$ dei casi in cui il farmaco è efficace, con riferimento all'intera popolazione.

Si può ragionevolmente pensare che il campione osservato, di dimensione $n = 550$, provenga da una popolazione $Ber(p)$, con $p$ ignoto.

Quindi, poiché $\widehat{p} = 393/550 = 0.715$, $\alpha/2 = 0.025$ e $z_{0.025} = 1.96$,

$$\left[0.715 - 1.96\frac{\sqrt{0.715 \cdot 0.285}}{\sqrt{550}}, 0.715 + 1.96\frac{\sqrt{0.715 \cdot 0.285}}{\sqrt{550}}\right]$$

$$= [0.677, 0.753]$$

è un intervallo di confidenza per $p$ con livello di confidenza approssimato 0.95. $\diamond$

**Intervalli di confidenza per la varianza di una popolazione normale**

Sia $X = (X_1, \ldots, X_n)$ un c.c.s. da una popolazione $N(\mu, \sigma^2)$. Si vuole determinare un intervallo di confidenza per $\sigma^2$ con livello $1 - \alpha$.

Come statistica campionaria, si considera la **varianza campionaria** $S^2$ e, come quantità pivotale, la trasformata $nS^2/\sigma^2 \sim \chi^2(n-1)$.

Per ogni $\sigma^2 \in \mathbf{R}^+$,

$$P_{\mu,\sigma^2}\left(\chi^2_{1-\alpha/2} \leq \frac{nS^2}{\sigma^2} \leq \chi^2_{\alpha/2}\right) =$$

$$P_{\mu,\sigma^2}\left(\frac{nS^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{nS^2}{\chi^2_{1-\alpha/2}}\right) = 1 - \alpha,$$

con $\chi^2_{1-\alpha/2}$ tale che $P(Y \geq \chi^2_{1-\alpha/2}) = 1 - \alpha/2$ e $\chi^2_{\alpha/2}$ tale che $P(Y \geq \chi^2_{\alpha/2}) = \alpha/2$, dove $Y \sim \chi^2(n-1)$.

Quindi

$$\left[\frac{nS^2}{\chi^2_{\alpha/2}}, \frac{nS^2}{\chi^2_{1-\alpha/2}}\right]$$

è un intervallo di confidenza per $\sigma^2$ con livello $1 - \alpha$.

Il medesimo intervallo si può ottenere considerando la **varianza campionaria corretta** $S_c^2$ e, come quantità pivotale, la trasformata $(n-1)S_c^2/\sigma^2 \sim \chi^2(n-1)$.

**Esempio.** Sia $X_1, \ldots, X_{30}$ un c.c.s. di dimensione $n = 30$ da una popolazione normale con $\mu$ e $\sigma^2$ ignoti.

Nell'ipotesi che i risultati dell'indagine campionaria siano tali che $\sum_{i=1}^{30} x_i = 36.76$ e $\sum_{i=1}^{30} x_i^2 = 102.27$, si fornisca un intervallo di confidenza per $\sigma^2$ con livello $1-\alpha = 0.95$.

In questo caso, $\bar{x}_{30} = \sum_{i=1}^{30} x_i/30 = 1.225$ e

$$s^2 = \frac{1}{30}\sum_{i=1}^{30} x_i^2 - \bar{x}_{30}^2 = 1.908.$$

Inoltre, $\alpha/2 = 0.025$ e, considerando un distribuzione $\chi^2$ con 29 gradi di libertà, $\chi^2_{1-0.025} = \chi^2_{0.975} = 16.047$ e $\chi^2_{0.025} = 45.722$. Sulla base dei risultati di pag. 51, si conclude che

$$\left[\frac{30 \cdot 1.908}{45.722}, \frac{30 \cdot 1.908}{16.047}\right] = [1.252, 3.567]$$

è l'intervallo di confidenza cercato. $\diamond$

## Verifica di ipotesi

Con le procedure di verifica di ipotesi si vuole verificare, sulla base dell'osservazione $x$ e del modello statistico, se una data congettura o ipotesi sulla popolazione oggetto di indagine sia accettabile o meno.

Si è propensi ad accettare l'ipotesi se essa è in accordo con i dati $x$.

Nell'ambito dell'inferenza statistica parametrica, l'**ipotesi** (statistica) è una affermazione o una congettura sul parametro ignoto $\theta$.

L'ipotesi è **semplice**, se specifica in modo completo la popolazione (il processo generatore dei dati); ad esempio, $\theta = 3$.

L'ipotesi è **composta**, se non specifica in modo completo la popolazione (il processo generatore dei dati); ad esempio, $\theta > 3$, $\theta < 3$, $\theta \neq 3$.

Il **test statistico** è un procedimento che consente di rifuitare o non rifiutare (e quindi accettare) un'ipotesi.

La decisione dipende dalla discrepanza, più o meno accentuata, che si rileva tra quanto ci si attende sulla base dell'ipotesi formulata e quanto si osserva nel campione.

L'ipotesi sottoposta a verifica viene chiamata **ipotesi nulla**; nel seguito si considereranno ipotesi nulle semplici del tipo

$$H_0 : \theta = \theta_0$$

dove $\theta_0$ è un valore fissato dello spazio parametrico.

L'ipotesi nulla è, in genere, un'assunzione semplificatrice sul modello statistico o che descrive le conoscenze attuali sulla popolazione (fenomeno) oggetto di indagine.

Congetture non contemplate da $H_0$ costituiscono l'**ipotesi alternativa** $H_1$. Nel seguito si considereranno le seguenti tipologie di ipotesi alternative:

- $H_1 : \theta = \theta_1$, con $\theta_1 \neq \theta_0$, detta **alternativa semplice**;
- $H_1 : \theta > \theta_0$, detta **alternativa unilaterale destra**;
- $H_1 : \theta < \theta_0$, detta **alternativa unilaterale sinistra**;
- $H_1 : \theta \neq \theta_0$, detta **alternativa bilaterale**.

**Esempio.** Si vuole indagare se un nuovo antipiretico sia più efficace del migliore attualmente in commercio.

In questo caso, $H_0$ traduce l'ipotesi che l'efficacia dei due prodotti sia equivalente, mentre $H_1$ attribuisce maggiore efficacia al nuovo prodotto.

Si è interessati a verificare se, alla luce del campione osservato, può essere ragionevole abbandonare l'ipotesi $H_0$ a favore di $H_1$. $\diamond$

Tipicamente interessa provare l'inaccettabilità di $H_0$ a favore di $H_1$, ammettendo una certa quota di errore.

La conclusione di un test statistico, che porta ad accettare l'ipotesi nulla o a rifiutarla a favore dell'alternativa (in tal caso si dice che il *test* è *significativo*), contempla la possibilità di errore.

Non si dice nulla circa la verità o falsità di $H_0$, si afferma unicamente che l'ipotesi risulta più o meno ragionevole, alla luce dei dati e del modello statistico.

**Esempio.** Una industria produce batterie elettriche con durata media dichiarata di 36 mesi. Un acquirente desidera comprare delle batterie, ma vuole accertarsi che la durata media non sia più bassa di quella dichiarata.

Si osserva la durata di $n = 40$ batterie. Di fatto si considera un c.c.s. $X_1, \ldots, X_{40}$ tratto da una popolazione $N(\mu, \sigma^2)$, con media $\mu$ ignota e varianza $\sigma^2 = 9$ supposta nota.

Il compratore vuole verificare l'ipotesi $H_0 : \mu = 36$ contro l'alternativa, per lui sfavorevole, $H_1 : \mu < 36$.

Dal momento che si vuole fare inferenza su $\mu$, si considera, come statistica campionaria, la media campionaria $\bar{X}_{40}$.

*Se si osservano valori piccoli per $\bar{X}_{40}$, si conclude che c'è una discrepanza tra l'ipotesi nulla e il campione osservato.*

*Cosa vuol dire piccoli?*

Si può affrontare il problema da due punti di vista.

(1) Quale è la soglia $c$ sotto la quale rifiuto $H_0$ a favore di $H_1$?

Tipicamente, la soglia $c$ viene determinata di modo che risulti fissata ad un valore sufficientemente piccolo, ad esempio $\alpha = 0.05$, la probabilità di rifiutare $H_0$, nonostante sia vera.

Formalmente, $c$ è tale che

$$P_0(\bar{X}_{40} \leq c) = 0.05,$$

dove $P_0(\cdot)$ indica la probabilità sotto $H_0$, cioè nell'ipotesi che $\mu = 36$.

Poiché la media campionaria standardizzata ha legge $N(0, 1)$,

$$P_0(\bar{X}_{40} \leq c) = P_0\left(\frac{\bar{X}_{40} - 36}{3/\sqrt{40}} \leq \frac{c - 36}{3/\sqrt{40}}\right) = 0.05$$

e $z_{0.05} = 1.645$, si conclude che $(c - 36)/(3/\sqrt{40}) = 1.645$, da cui si ricava che la soglia è

$$c = 36 - 1.645 \cdot (3/\sqrt{40}) = 35.22.$$

Quindi, se si osserva un campione $x$ tale che $\bar{x}_{40} \leq 35.22$, si rifiuterà $H_0$ a favore di $H_1$.

(2) Se si osserva un campione $x$ tale che $\bar{x}_{40} = 35$ (valore che sembra vicino a 36), come considero l'ipotesi $H_0$?

Poiché valori piccoli per $\bar{X}_{40}$ sono sinonimo di disaccordo tra il campione osservato e $H_0$, si considera la probabilità, sotto $H_0$, dell'evento $\bar{X}_{40} \leq 35$; più precisamente

$$P_0(\bar{X}_{40} \leq 35) = P_0\left(\frac{\bar{X}_{40} - 36}{3/\sqrt{40}} \leq \frac{35 - 36}{3/\sqrt{40}}\right)$$
$$= P_0(Z \leq -2.11) = 0.017,$$

essendo $Z \sim N(0, 1)$.

Quindi, nonostante l'apparenza, il valore $\bar{x}_{40} = 35$ è un valore anomalmente piccolo per $\bar{X}_{40}$, se è vera $H_0$.

*C'è una scarsa conformità tra il campione osservato e l'ipotesi nulla, nella direzione dell'ipotesi alternativa.* $\diamond$

Nel seguito, non si svilupperà la teoria della verifica di ipotesi. Si forniranno delle indicazioni pratiche, utili per la costruzione di test di ipotesi per particolari problemi statistici.

**Definizione.** Dato un campione $X = (X_1, \ldots, X_n)$ e le ipotesi $H_0$ e $H_1$, si chiama **statistica test** una statistica campionaria $T = t(X)$ che permette di evidenziare se sia più ragionevole accettare $H_0$ o rifiutare $H_0$ in favore di $H_1$.

In genere, si sceglie come statistica test uno stimatore per $\theta$ (o una sua trasformata) per il quale risulta nota, in forma esatta o approssimata, la distribuzione di probabilità (per lo meno sotto $H_0$).

Un test statistico non è una procedura libera da errori. Si individuano le seguenti tipologie d'errore:

- **errore di prima specie**, se si rifiuta $H_0$ quando è vera; la associata *probabilità* viene *indicata con* $\alpha$;

- **errore di seconda specie**, se si accetta $H_0$ quando è falsa; la associata *probabilità* viene *indicata con* $\beta$.

La probabilità $1-\alpha$ di accettare $H_0$ quando è vera è detta **livello di protezione** del test, mentre la probabilità $1-\beta$ di rifiutare $H_0$ quando è falsa è detta **potenza** del test.

Le probabilità $\alpha$ e $\beta$ sono interdipendenti. *Non è possibile minimizzare entrambe le probabilità d'errore*.

In genere, si fissa $\alpha$ detta **livello di significatività** (ad esempio, $\alpha = 0.1, 0.05, 0.01$) e si cerca il test che, a parità di $\alpha$, presenta potenza più elevata.

La procedura che verrà usualmente adottata per definire test statistici è la seguente.

Se $H_0 : \theta = \theta_0$ e $H_1 : \theta > \theta_0$ (**alternativa unilaterale destra**), avendo *fissato* il livello di significatività $\alpha$, si considera come statistica test un opportuno stimatore $T = t(X)$ per $\theta$.

Valori grandi per $T$ non sono conformi con $H_0$ e quindi, in tal caso, si è propensi a rifiutare l'ipotesi nulla a favore di $H_1$.

Avendo fissato $\alpha$, il valore soglia $c_\alpha$ è tale che

$$P_0(T \geq c_\alpha) = \alpha,$$

e si suppone calcolabile in forma esatta o approssimata.

Risulta quindi individuata la regione

$$R_\alpha = \{x : t(x) \geq c_\alpha\},$$

dello spazio campionario, detta **regione di rifiuto** o regione critica di livello $\alpha$. In modo speculare si individua la **regione di accettazione** $A_\alpha$.

Quindi,

- se il campione osservato $x$ è tale che $x \in R_\alpha$, si rifiuta $H_0$ a favore di $H_1$ e si dice che il test è significativo al $100\alpha\%$;

- se $x$ è tale che $x \in A_\alpha$, si accetta (si mantiene) l'ipotesi nulla $H_0$.

Il test statistico così definito è detto **test unilaterale destro di livello** $\alpha$.

Nel caso in cui $H_1 : \theta < \theta_0$ (**alternativa unilaterale sinistra**),

$$R_\alpha = \{x : t(x) \leq c_\alpha\},$$

con $c_\alpha$ tale che $P_0(T \leq c_\alpha) = \alpha$ e si ha un **test unilaterale sinistro di livello** $\alpha$.

Nel caso in cui $H_1 : \theta \neq \theta_0$ (**alternativa bilaterale**),

$$R_\alpha = \{x : t(x) \leq c_{\alpha/2} \text{ oppure } t(x) \geq d_{\alpha/2}\},$$

con $c_{\alpha/2}$, $d_{\alpha/2}$ tali che

$$P_0(T \leq c_{\alpha/2}) = \alpha/2, \quad P_0(T \geq d_{\alpha/2}) = \alpha/2$$

e si ha un **test bilaterale di livello** $\alpha$.

Sia $t^{oss} = t(x^{oss})$ il valore osservato di una statistica test con regione di rifiuto unilaterale destra. Poiché valori grandi di $t^{oss}$ sono sinonimo di disaccordo tra $H_0$ e $x^{oss}$, la probabilità

$$\alpha^{oss} = P_0(T \geq t^{oss}),$$

detta **livello di significatività osservato** (*P-value*) del test, fornisce una misura sintetica del grado di conformità tra dati e ipotesi nulla.

Nel caso di test con regione di rifiuto unilaterale sinistra,

$$\alpha^{oss} = P_0(T \leq t^{oss}).$$

Nel caso di test con regione di rifiuto bilaterale,

$$\alpha^{oss} = 2\min\{P_0(T \leq t^{oss}), P_0(T \geq t^{oss})\}.$$

Ovviamente, $\alpha^{oss} \in [0, 1]$ e un valore $\alpha^{oss}$ vicino a zero corrisponde ad un valore osservato $t^{oss}$ che risulta anomalo per $T$ sotto $H_0$.

$\alpha^{oss}$ vicino a zero indica scarsa conformità tra $x^{oss}$ e $H_0$, nella direzione di $H_1$.

**Verifica di ipotesi sulla media di una popolazione normale**

Sia $X = (X_1, \ldots, X_n)$ un c.c.s. da una popolazione $N(\mu, \sigma^2)$. Si vuole definire un test sulla media $\mu$ con livello di significatività $\alpha$.

Come statistica campionaria si considera la **media campionaria** $\bar{X}_n$. Si individuano due casi.

(1) **varianza $\sigma^2$ nota**.

A) $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$ (**alternativa unilaterale destra**)

Poiché, avendo fissato $\alpha$,

$$P_0(\bar{X}_n \geq c_\alpha) = P_0\left(\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \geq \frac{c_\alpha - \mu_0}{\sigma/\sqrt{n}}\right) = \alpha$$

e $z_\alpha$ è tale che $P(Z \geq z_\alpha) = \alpha$, con $Z \sim N(0,1)$, si conclude che $(c_\alpha - \mu_0)/(\sigma/\sqrt{n}) = z_\alpha$ e

$$R_\alpha = \{x : \bar{x}_n \geq \mu_0 + z_\alpha \sigma/\sqrt{n}\}.$$

Inoltre

$$\alpha^{oss} = P_0\left(\bar{X}_n \geq \bar{x}_n\right) = 1 - \Phi\left(\frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}\right).$$

B) $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$ (**alternativa unilaterale sinistra**)

In modo analogo si ottiene che

$$R_\alpha = \{x : \bar{x}_n \leq \mu_0 - z_\alpha \sigma/\sqrt{n}\},$$

$$\alpha^{oss} = P_0\left(\bar{X}_n \leq \bar{x}_n\right) = \Phi\left(\frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}\right).$$

C) $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ (**alternativa bilaterale**)

In modo analogo si ottiene che

$$R_\alpha = \{x : \bar{x}_n \leq \mu_0 - z_{\alpha/2}\sigma/\sqrt{n} \text{ oppure } \bar{x}_n \geq \mu_0 + z_{\alpha/2}\sigma/\sqrt{n}\},$$

$$\alpha^{oss} = 2\min\left\{\Phi\left(\frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}\right), 1 - \Phi\left(\frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}\right)\right\}.$$

(2) **varianza $\sigma^2$ ignota**.

Si considera la **varianza campionaria corretta** $S_c^2$ e la **media campionaria studentizzata** $\sqrt{n}(\bar{X}_n - \mu)/S_c \sim t(n-1)$.

Con una procedura simile alla precedente, si ottengono le seguenti conclusioni (*test t*).

A) $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$ (**alternativa unilaterale destra**)

$$R_\alpha = \{x : \bar{x}_n \geq \mu_0 + t_\alpha s_c/\sqrt{n}\},$$

dove $t_\alpha$ è tale che $P(Y \geq t_\alpha) = \alpha$ con $Y \sim t(n-1)$. Inoltre

$$\alpha^{oss} = P_0\left(\bar{X}_n \geq \bar{x}_n\right) = 1 - F_Y\left(\frac{\bar{x}_n - \mu_0}{s_c/\sqrt{n}}\right),$$

con $F_Y(\cdot)$ la funzione di ripartizione di $Y$.

B) $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$ (**alternativa unilaterale sinistra**)

In modo analogo si ottiene che

$$R_\alpha = \{x : \bar{x}_n \leq \mu_0 - t_\alpha s_c/\sqrt{n}\},$$

$$\alpha^{oss} = P_0\left(\bar{X}_n \leq \bar{x}_n\right) = F_Y\left(\frac{\bar{x}_n - \mu_0}{s_c/\sqrt{n}}\right).$$

C) $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ (**alternativa bilaterale**)

In modo analogo si ottiene che

$$R_\alpha = \{x : \bar{x}_n \leq \mu_0 - t_{\alpha/2} s_c/\sqrt{n} \text{ oppure } \bar{x}_n \geq \mu_0 + t_{\alpha/2} s_c/\sqrt{n}\},$$

$$\alpha^{oss} = 2\min\left\{F_Y\left(\frac{\bar{x}_n - \mu_0}{s_c/\sqrt{n}}\right), 1 - F_Y\left(\frac{\bar{x}_n - \mu_0}{s_c/\sqrt{n}}\right)\right\}.$$

**Esempio.** Si vuole studiare la concentrazione di urea nel sangue di una certa popolazione di bovini. Dall'analisi della letteratura si ricava che la concentrazione di urea, misurata in mg per 100 ml, è descritta da una variabile casuale $N(25, 45)$.

Si misura la concentrazione di urea in 10 bovini e si osserva una concentrazione media di 22 mg per 100 ml. È ragionevole pensare che il campione di bovini provenga dalla medesima popolazione studiata in letteratura?

In questo caso si ha un c.c.s. $X_1, \ldots, X_{10}$ da una popolazione $N(\mu, 45)$ e si vuole verificare l'ipotesi $H_0 : \mu = 25$ contro l'alternativa $H_1 : \mu \neq 25$ con un livello di significatività $\alpha = 0.05$.

Dal campione osservato $x$ si ricava che $\bar{x}_{10} = 22$ e, poiché $n = 10$, $\mu_0 = 25$, $\sigma = 6.708$, $z_{0.025} = 1.96$,

$$x \notin R_{0.05} = \{x : \bar{x}_n \leq 20.842 \text{ oppure } \bar{x}_n \geq 29.158\}.$$

L'ipotesi $H_0$ non viene rifiutata. Tuttavia il grado di conformità tra $H_0$ e i dati non è molto alto poiché

$$\alpha^{oss} = 2 \min \left\{ \Phi\left(\frac{22-25}{6.708/\sqrt{10}}\right), 1 - \Phi\left(\frac{22-25}{6.708/\sqrt{10}}\right) \right\}$$
$$= 0.157.$$

$\diamond$

**Esempio.** Si consideri un problema analogo al precedente. L'unica differenza è che, dall'analisi della letteratura, si ricava che la concentrazione di urea, misurata in mg per 100 ml, è descritta da una variabile casuale $N(25, \sigma^2)$, con $\sigma^2$ non nota.

In questo caso si ha un c.c.s. $X_1, \ldots, X_{10}$ da una popolazione $N(\mu, \sigma^2)$ e si vuole verificare l'ipotesi $H_0 : \mu = 25$ contro l'alternativa $H_1 : \mu \neq 25$ con un livello di significatività $\alpha = 0.05$.

Dal campione osservato $x$ si ricava che $\bar{x}_{10} = 22$ e che $(1/10)\sum_{i=1}^{10} x_i^2 = 526.4$. Quindi,

$$s^2 = \frac{1}{10}\sum_{i=1}^{10} x_i - \bar{x}_{10}^2 = 42.4, \quad s_c^2 = \frac{10}{9}\, s^2 = 47.111.$$

Poiché $n = 10$, $\mu_0 = 25$, $s_c = 6.864$ e, con riferimento a una distribuzione $t$ di Student con 9 gradi di libertà, $t_{0.025} = 2.262$,

$$x \notin R_{0.05} = \{x : \bar{x}_n \leq 20.09 \text{ oppure } \bar{x}_n \geq 29.91\}.$$

L'ipotesi $H_0$ non viene rifiutata. Tuttavia il grado di conformità tra $H_0$ e i dati non è molto alto poiché

$$\alpha^{oss} = 2 \min \left\{ F_Y \left( \frac{22 - 25}{6.864/\sqrt{10}} \right), 1 - F_Y \left( \frac{22 - 25}{6.864/\sqrt{10}} \right) \right\}$$

$$\doteq 0.2,$$

essendo $Y \sim t(9)$. $\diamond$

**Verifica di ipotesi sulla media di una popolazione non normale**

Sia $X = (X_1, \ldots, X_n)$ un c.c.s. da una popolazione non normale con media $\mu$ e varianza $\sigma^2$ ignote. Si vuole definire un test sulla media $\mu$ con livello di significatività $\alpha$.

Se la numerosità campionaria $n$ è *sufficientemente elevata*, si riesce a determinare, con relativa facilità, un test con **livello di significatività $\alpha$ approssimato**.

Come statistica test si considera la **media campionaria** $\bar{X}_n$ e, dal momento che $n$ è elevato, nella procedura di standardizzazione si può sostituire a $\sigma$, se non risulta specificato da $H_0$, un'opportuno stimatore consistente $\widehat{\sigma}$.

Quindi, posto $z_\alpha$ tale che $P(Z \geq z_\alpha)$ con $Z \sim N(0,1)$, si conclude che

A) se $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$ (**alternativa unilaterale destra**)

$$R_\alpha = \{x : \bar{x}_n \geq \mu_0 + z_\alpha \widehat{\sigma}/\sqrt{n}\}$$

$$\alpha^{oss} = P_0\left(\bar{X}_n \geq \bar{x}_n\right) \doteq 1 - \Phi\left(\frac{\bar{x}_n - \mu_0}{\widehat{\sigma}/\sqrt{n}}\right);$$

B) se $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$ (**alternativa unilaterale sinistra**)

$$R_\alpha = \{x \, : \, \bar{x}_n \leq \mu_0 - z_\alpha \widehat{\sigma}/\sqrt{n}\},$$

$$\alpha^{oss} = P_0\left(\bar{X}_n \leq \bar{x}_n\right) \doteq \Phi\left(\frac{\bar{x}_n - \mu_0}{\widehat{\sigma}/\sqrt{n}}\right);$$

C) se $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ (**alternativa bilaterale**)

$R_\alpha = \{x \, : \, \bar{x}_n \leq \mu_0 - z_{\alpha/2}\widehat{\sigma}/\sqrt{n}$ oppure $\bar{x}_n \geq \mu_0 + z_{\alpha/2}\widehat{\sigma}/\sqrt{n}\}$,

$$\alpha^{oss} \doteq 2 \min\left\{\Phi\left(\frac{\bar{x}_n - \mu_0}{\widehat{\sigma}/\sqrt{n}}\right), 1 - \Phi\left(\frac{\bar{x}_n - \mu_0}{\widehat{\sigma}/\sqrt{n}}\right)\right\}.$$

*Se, sotto $H_0$, si può concludere che $\sigma^2 = \sigma_0^2$, si sostituisce $\widehat{\sigma}$ con $\sigma_0$.*

Si evidenziano i seguenti casi.

(1) Sia $X = (X_1, \ldots, X_n)$ un c.c.s. da una popolazione $Ber(p)$, con $p$ ignoto. Si suppone che $n$ sia sufficientemente elevato.

Si consideri $H_0 : p = p_0$, $H_1 : p \neq p_0$ (si procede in modo analogo nel caso di alternative unilaterali).

Poiché $\mu = p$ e $\sigma^2 = p(1-p)$, la regione di rifiuto del test di livello $\alpha$ approssimato, si ottiene a partire da $\widehat{p} = \bar{X}_n$.

In questo caso, sotto $H_0$, $p = p_0$ e $\sigma_0^2 = p_0(1 - p_0)$ e quindi

$$R_\alpha = \{x \, : \, \widehat{p} \leq p_0 - z_{\alpha/2}\sqrt{p_0(1-p_0)/n} \text{ oppure}$$

$$\widehat{p} \geq p_0 + z_{\alpha/2}\sqrt{p_0(1-p_0)/n}\}.$$

(2) Sia $X = (X_1, \ldots, X_n)$ un c.c.s. da una popolazione $P(\lambda)$, con $\lambda$ ignoto. Si suppone che $n$ sia sufficientemente elevato.

Si consideri $H_0 : \lambda = \lambda_0$, $H_1 : \lambda \neq \lambda_0$ (si procede in modo analogo nel caso di alternative unilaterali).

Poiché $\mu = \lambda$ e $\sigma^2 = \lambda$, la regione di rifiuto del test di livello $\alpha$ approssimato, si ottiene a partire da $\widehat{\lambda} = \bar{X}_n$.

In questo caso, sotto $H_0$, $\lambda = \lambda_0$ e $\sigma_0^2 = \lambda_0$ e quindi

$$R_\alpha = \{x \, : \, \widehat{\lambda} \leq \lambda_0 - z_{\alpha/2}\sqrt{\lambda_0/n} \text{ oppure } \widehat{\lambda} \geq \lambda_0 + z_{\alpha/2}\sqrt{\lambda_0/n}\}.$$

**Esempio.** In un campione di 120 individui, 37 sono risultati talassemici. Sapendo che l'incidenza della talassemia nella regione da cui provengono gli individui è del 20%, ci si chiede se la proporzione osservata può essere considerata in accordo con tale dato.

Si può ragionevolmente pensare che il campione osservato, di dimensione $n = 120$, provenga da una popolazione $Ber(p)$, con $p$ ignoto.

Si vuole verificare l'ipotesi nulla $H_0 : p = 0.2$ contro un'alternativa bilaterale $H_1 : p \neq 0.2$, con un livello di significatività (approssimato) $\alpha = 0.05$.

Quindi, essendo $n$ elevato, poiché $\widehat{p} = 36/120 = 0.3$, $\alpha/2 = 0.025$, $z_{0.025} = 1.96$ e $\sigma_0 = \sqrt{0.2 \cdot (1 - 0.2)} = 0.4$, sotto $H_0$, si conclude che

$$x \in R_{0.05} = \{x : \bar{x}_n \leq 0.128 \text{ oppure } \bar{x}_n \geq 0.272\}.$$

L'ipotesi $H_0$ viene rifiutata e il test risulta significativo al livello (approssimato) $\alpha = 0.05$. Il grado di conformità tra $H_0$ e i dati corrisponde a

$$\alpha^{oss} \doteq 2 \min \left\{ \Phi\left(\frac{0.3 - 0.2}{0.4/\sqrt{120}}\right), 1 - \Phi\left(\frac{0.3 - 0.2}{0.4/\sqrt{120}}\right) \right\}$$

$$\doteq 0.006$$

$\diamond$

**Verifica di ipotesi sulla varianza di una popolazione normale**

Sia $X = (X_1, \ldots, X_n)$ un c.c.s. da una popolazione $N(\mu, \sigma^2)$. Si vuole definire un test sulla varianza $\sigma^2$ con livello di significatività $\alpha$.

Come statistica campionaria si considera la **varianza campionaria** $S^2$, da cui si ha che $n\,S^2/\sigma^2 \sim \chi^2(n-1)$. Si individuano i seguenti casi.

A) $H_0 : \sigma^2 = \sigma_0^2$, $H_1 : \sigma^2 > \sigma_0^2$ (**alternativa unilaterale destra**)

Poiché, avendo fissato $\alpha$,

$$P_0(S^2 \geq c_\alpha) = P_0\left(\frac{n\,S^2}{\sigma_0^2} \geq \frac{n\,c_\alpha}{\sigma_0^2}\right) = \alpha$$

e $\chi^2_\alpha$ è tale che $P(Y \geq \chi^2_\alpha) = \alpha$, con $Y \sim \chi^2(n-1)$, si conclude che $n\,c_\alpha/\sigma_0^2 = \chi^2_\alpha$ e

$$R_\alpha = \{x \,:\, s^2 \geq \sigma_0^2\,\chi^2_\alpha/n\}.$$

Inoltre

$$\alpha^{oss} = P_0\left(S^2 \geq s^2\right) = 1 - F_Y\left(\frac{n\,s^2}{\sigma_0^2}\right),$$

dove $F_Y(\cdot)$ è la funzione di ripartizione di $Y \sim \chi^2(n-1)$.

B) $H_0 : \sigma^2 = \sigma_0^2$, $H_1 : \sigma^2 < \sigma_0^2$ (**alternativa unilaterale sinistra**)

In modo analogo si ottiene che

$$R_\alpha = \{x \, : \, s^2 \leq \sigma_0^2 \, \chi^2_{1-\alpha}/n\},$$

con $\chi^2_{1-\alpha}$ tale che $P(Y \geq \chi^2_{1-\alpha}) = 1 - \alpha$, e

$$\alpha^{oss} = P_0\left(S^2 \leq s^2\right) = F_Y\left(\frac{n\, s^2}{\sigma_0^2}\right).$$

C) $H_0 : \sigma^2 = \sigma_0^2$, $H_1 : \sigma^2 \neq \sigma_0^2$ (**alternativa bilaterale**)

In modo analogo si ottiene che

$$R_\alpha = \{x \, : \, s^2 \leq \sigma_0^2 \, \chi^2_{1-\alpha/2}/n \text{ oppure } s^2 \geq \sigma_0^2 \, \chi^2_{\alpha/2}/n\},$$

con $\chi^2_{1-\alpha/2}$ tale che $P(Y \geq \chi^2_{1-\alpha/2}) = 1 - \alpha/2$ e $\chi^2_{\alpha/2}$ tale che $P(Y \geq \chi^2_{\alpha/2}) = \alpha/2$, e

$$\alpha^{oss} = 2 \min\left\{F_Y\left(\frac{n\, s^2}{\sigma_0^2}\right), 1 - F_Y\left(\frac{n\, s^2}{\sigma_0^2}\right)\right\}.$$

**Esempio.** Si vuole studiare la variabilità dei tempi di crescita di un gruppo di 13 piante ottenute con una varietà di semi modificati geneticamente. È noto che i tempi di crescita per tale tipologia di piante sono descritti da un modello $N(\mu, \sigma_0^2)$, con $\mu$ ignoto e $\sigma_0^2 = 3.83$.

In questo caso si ha un c.c.s. $X_1, \ldots, X_{13}$ da una popolazione $N(\mu, \sigma^2)$ e si vuole verificare l'ipotesi $H_0 : \sigma^2 = 3.83$ contro l'alternativa $H_1 : \sigma^2 \neq 3.83$ con un livello di significatività $\alpha = 0.05$.

Dal campione osservato $x$ si ricava che $\bar{x}_{13} = 12.4$ e che $(1/13)\sum_{i=1}^{13} x_i^2 = 161.02$. Quindi,

$$s^2 = \frac{1}{13}\sum_{i=1}^{13} x_i - \bar{x}_{13}^2 = 7.26.$$

Poiché $n = 13$, $\sigma_0^2 = 3.83$, $\alpha/2 = 0.025$ e, con riferimento a una distribuzione $\chi^2$ con 12 gradi di libertà, $\chi^2_{1-0.025} = 4.404$ e $\chi^2_{0.025} = 23.337$,

$$x \in R_{0.05} = \{x \,:\, s^2 \leq 1.297 \text{ oppure } s^2 \geq 6.875\}.$$

L'ipotesi $H_0$ viene rifiutata e il test è significativo al livello $\alpha = 0.05$. Il grado di conformità tra $H_0$ e i dati è descritto da

$$\alpha^{oss} = 2\min\left\{F_Y\left(\frac{13 \cdot 7.26}{3.83}\right), 1 - F_Y\left(\frac{13 \cdot 7.26}{3.83}\right)\right\}$$

$$\doteq 0.033,$$

essendo $Y \sim \chi^2(12)$. $\diamond$

## Test di verosimiglianza

Si considerino le ipotesi semplici

$$H_0 : \theta = \theta_0, \quad H_1 : \theta = \theta_1$$

e, con riferimento al campione casuale $X = (X_1, \ldots, X_n)$, si calcoli la funzione di verosimiglianza $L(\theta) = L(\theta; x)$ per $\theta$, basata sui dati $x$.

Per il **Lemma di Neyman-Pearson**, il **test rapporto di verosimiglianza**, basato sulla statistica

$$\frac{L(\theta_0; X)}{L(\theta_1; X)},$$

con regione di rifiuto

$$R_\alpha = \{x : L(\theta_0; x)/L(\theta_1; x) \leq c_\alpha\},$$

dove $c_\alpha$ è tale che $P_0(L(\theta_0; X)/L(\theta_1; X) \leq c_\alpha) = \alpha$, è il più potente tra quelli di livello di significatività $\alpha$.

L'ipotesi nulla viene rifiutata se si osserva per $L(\theta_0; x)$ un valore significativamente più basso di $L(\theta_1; x)$; in caso contrario l'ipotesi nulla viene mantenuta.

In alcune situazioni la statistica test rapporto di verosimiglianza è equivalente a statistiche test di uso comune.

Ad esempio, per la verifica di ipotesi sulla media o sulla varianza di una popolazione normale, tale statistica test è equivalente a $\bar{X}_n$ e $S^2$, rispettivamente (definiscono le medesime regioni di rifiuto).

In alcuni casi, la proprietà di ottimalità dei test basati sul rapporto di verosimiglianza si mantiene anche quando si hanno ipotesi composte.

In particolare, con riferimento a ipotesi sulla media e sulla varianza di una popolazione normale, il test rapporto di verosimiglianza è ottimale ed è equivalente a quelli visti in precedenza.

Se si hanno le ipotesi composte

$$H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_1,$$

con $\Theta_0$, $\Theta_1$ una partizione dello spazio parametrico $\Theta$, il **test rapporto di verosimiglianza**, si basa sulla statistica

$$\frac{\max_{\theta \in \Theta_0} L(\theta; X)}{\max_{\theta \in \Theta_1} L(\theta; X)}.$$

Più di frequente si considera la forma equivalente basata sulla statistica

$$\lambda = \frac{\max_{\theta \in \Theta_0} L(\theta; X)}{\max_{\theta} L(\theta; X)} = \frac{\max_{\theta \in \Theta_0} L(\theta; X)}{L(\widehat{\theta}; X)},$$

con $\widehat{\theta}$ lo s.m.v., e quindi si rifiuterà $H_0$ se si osservano valori prossimi a zero, mentre si mantiene $H_0$ per valori osservati prossimi a 1.

Nel caso in cui $H_0 : \theta = \theta_0$ e $H_1 : \theta \neq \theta_0$,

$$\lambda(\theta_0; X) = \frac{L(\theta_0; X)}{L(\widehat{\theta}; X)}.$$

In alternativa a $\lambda$ si può considerare la statistica **log-rapporto di verosimiglianza**

$$W(\theta_0; X) = -2 \log \lambda(\theta_0; X) = 2\{\ell(\widehat{\theta}; X) - \ell(\theta_0; X)\}.$$

Valori grandi di $W(\theta_0; x)$ sono critici per l'ipotesi nulla.

In molti casi la soglia $c_\alpha$ si può determinare con relativa facilità se $n$ *è elevato*, poiché, sotto $H_0$,

$$W(\theta_0; X) \dot{\sim} \chi^2(p),$$

dove $p$ è la dimensione del parametro ignoto $\theta$.

**Esempio.** Si consideri l'esempio di pag. 10, dove si è descritta la funzione di verosimiglianza

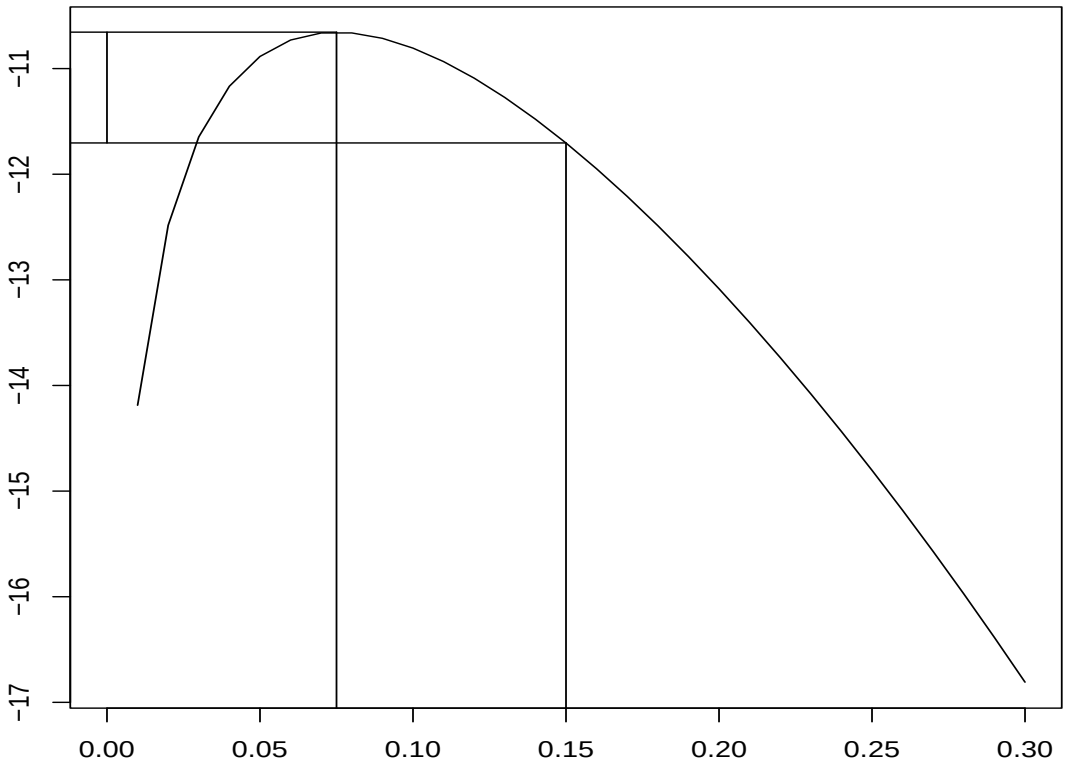
$$L(p) = p^3(1-p)^{37}, \quad p \in (0,1),$$

riferita a un campione osservato di dimensione $n = 40$ tratto da una popolazione $Ber(p)$.

Nella figura sottostante si riporta il grafico della associata funzione di log-verosimiglianza per $p \in (0, 0.3)$, con s.m.v. $\widehat{\theta} = 3/40 = 0.075$.

Si consideri $H_0 : \theta = 0.15$ e $H_1 : \theta \neq 0.15$.

La statistica test $W(0.15; x) = 2\{\ell(0.075; x) - \ell(0.15; x)\}$ è proporzionale all'ampiezza dell'intervallo evidenziato sull'asse delle ordinate.




◇