

Statistica e Laboratorio

7. Inferenza statistica: statistiche campionarie

Paolo Vidoni

Dipartimento di Scienze Economiche e Statistiche
Università di Udine
via Tomadini 30/a - Udine
paolo.vidoni@uniud.it

<https://elearning.uniud.it/>

Sommario

1 **Sommario e introduzione**

2 Campionamento

3 Modelli statistici

4 Statistiche campionarie

Sommario

- **Introduzione**
- **Campionamento**
- **Modelli statistici**
- **Statistiche campionarie**

Introduzione

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

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

La **statistica inferenziale** fornisce strumenti e metodi per:

- ricavare dai dati campionari informazioni sulla popolazione (fenomeno) di interesse;
- quantificare (utilizzando il linguaggio del *calcolo delle probabilità*) la fiducia da assegnare a tali informazioni.

Quindi, con le informazioni rilevate sul solo campione, l'obiettivo è arrivare a conclusioni che possano essere valide per tutta la popolazione o per il fenomeno generale di interesse.

Sommario

- 1 Sommario e introduzione
- 2 Campionamento**
- 3 Modelli statistici
- 4 Statistiche campionarie

Campionamento

I dati che si hanno a disposizione, quando si effettua una **indagine campionaria**, sono stati ottenuti mediante misurazioni o rilevazioni di certe caratteristiche di interesse su una parte delle unità statistiche di una popolazione.

Si effettuano indagini campionarie per le seguenti *motivazioni*:

- presenza di vincoli di tempo e/o problemi di costo;
- la popolazione di interesse può essere infinita e virtuale (si è interessati allo studio di un fenomeno aleatorio per cui non è definita una popolazione reale di riferimento);
- la rilevazione potrebbe distruggere le unità statistiche o essere potenzialmente dannosa (ad esempio nel controllo di qualità oppure nelle sperimentazioni mediche);
- la precisione dei risultati nelle rilevazioni censuarie potrebbe non essere adeguata.

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

Esempio. Farmaco. Per studiare l'efficacia di un determinato farmaco si studiano i suoi effetti su un gruppo di pazienti, che costituiscono il *campione*. La *popolazione* di riferimento è *virtuale* (potenzialmente infinita) ed è costituita da tutti i pazienti a cui si potrebbe somministrare il farmaco. L'interesse sulla popolazione, in questo caso, è sinonimo di interesse sull'efficacia del farmaco. ◇

Esempio. Controllo di qualità. 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. ◇

Esempio. Misurazioni. Si effettuano misurazioni ripetute, nelle medesime condizioni, di un determinato oggetto con uno strumento affetto da errore non sistematico. La *popolazione virtuale* di riferimento è costituita dall'insieme di tutte le infinite, potenziali misurazioni. Il *campione* è costituito dal numero finito di misurazioni che sono state effettuate. L'interesse sulla popolazione è, in questo caso, sinonimo di interesse sulla vera dimensione dell'oggetto. ◇

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

È essenziale che i dati campionari possano essere interpretati come risultato di un esperimento aleatorio, perché altrimenti verrebbe meno la *rappresentatività* del campione e la possibilità di ricavare informazioni utili sulla popolazione (fenomeno) di interesse (*inferenza*).

L'utilizzazione di un **campione di convenienza** (basato su metodi di tipo non probabilistico) rende improponibile ogni analisi statistica inferenziale.

Campioni casuali semplici

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**.

In un campione casuale semplice le unità vengono selezionate dalla popolazione di riferimento di modo che ognuna abbia la stessa possibilità di far parte del campione (estrazioni, con reinserimento, da un'urna).

Per popolazioni virtuali (infinite) o reali (finite) molto numerose, campionare con reinserimento o senza reinserimento (in blocco) non porta a differenze sostanziali.

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

Lo schema che caratterizza il campionamento casuale semplice verrà esteso considerando anche il caso di n realizzazioni **indipendenti** di un esperimento di base, **non necessariamente nelle medesime condizioni**.

Come evidenziato in precedenza, se non specificato diversamente, si assumerà di disporre di un campione casuale semplice, ottenuto selezionando le unità mediante estrazioni con reinserimento o con procedure che sono di fatto ad essa equivalenti.

Infatti, poiché solitamente si prenderanno in esame popolazioni virtuali (infinite) o reali (finite) molto numerose, considerare campioni con reinserimento o senza reinserimento (in blocco) non porta a differenze sostanziali nella trattazione.

Alla luce di quanto detto, i **dati osservati** (che costituiscono il **campione osservato**) $x = (x_1, \dots, x_n)$, $n \geq 1$, sono riferiti ad una caratteristica di interesse, rilevata sulle n unità statistiche che costituiscono il campione; in particolare, x_i , $i = 1, \dots, n$, indica l'osservazione effettuata sulla i -esima unità statistica.

L'**ipotesi fondamentale** su cui poggia l'inferenza statistica è che i dati campionari x sono una realizzazione di un **vettore di variabili casuali** $X = (X_1, \dots, X_n)$.

Questa ipotesi tiene conto del fatto che si è estratto uno tra i molti possibili campioni e che quindi siamo in presenza di un *esperimento aleatorio*.

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

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

Nel **campionamento casuale semplice**, le variabili casuali X_1, \dots, X_n sono **indipendenti e identicamente distribuite**, cioè con lo stesso modello probabilistico e tali da non influenzarsi a vicenda.

Ne seguito si considererà anche il caso più generale di variabili casuali X_1, \dots, X_n **indipendenti** ma **non necessariamente identicamente distribuite**, cioè con modello probabilistico che può risultare influenzato dalle eventuali diverse condizioni sperimentali.

Sommario

- 1 Sommario e introduzione
- 2 Campionamento
- 3 Modelli statistici**
- 4 Statistiche campionarie

Modelli statistici parametrici

Dato un campione casuale semplice X_1, \dots, X_n , la **distribuzione di probabilità** delle singole variabili casuali dipende dalla natura dei dati e del fenomeno oggetto di indagine.

Ad esempio, per dati binari sarà naturale ipotizzare $X_i \sim Ber(p)$, per misurazioni $X_i \sim N(\mu, \sigma^2)$, per conteggi $X_i \sim P(\lambda)$, per tempi di funzionamento $X_i \sim Esp(\lambda)$, ecc. Queste distribuzioni di probabilità possono rappresentare una buona approssimazione della realtà.

La distribuzione assunta per le variabili casuali del campione dipende da costanti ignote dette **parametri**; ad esempio, le quantità $p, \mu, \sigma^2, \lambda$, ecc.

Nell'*inferenza statistica parametrica* si assume che la distribuzione delle variabili casuali del campione sia nota a meno dei valori dei *parametri*, che corrispondono tipicamente agli *aspetti di interesse dell'analisi*.

Le assunzioni viste, ed elencate nel seguito, definiscono un **modello statistico parametrico** per i dati di un campione casuale semplice:

- le variabili casuali X_1, \dots, X_n sono indipendenti;
- tutte le X_i hanno la stessa distribuzione di probabilità (come detto in precedenza, questa ipotesi può anche essere rilassata);
- tale distribuzione è nota a meno dei valori di uno o più parametri, indicati genericamente come $\theta = (\theta_1, \dots, \theta_d)$, $d \geq 1$.

Scopo dell'inferenza statistica parametrica è utilizzare i dati osservati x_1, \dots, x_n per ottenere informazioni su θ , i cui possibili valori appartengono ad un certo insieme Θ , chiamato **spazio parametrico**.

Il supporto congiunto di X_1, \dots, X_n è detto **spazio campionario** e corrisponde all'insieme dei possibili campioni x_1, \dots, x_n che si possono osservare.

Le tre assunzioni citate in precedenza non è detto che siano soddisfatte nella pratica, ma possono rappresentare una *descrizione semplice e operativamente utile di una realtà complessa*.

La **scelta del modello** è molto importante, poiché le conclusioni inferenziali dipendono fortemente dalle assunzioni fatte.

Nella specificazione del modello statistico parametrico è importante considerare:

- la natura dei dati (qualitativi o quantitativi, discreti o continui, ecc.);
- gli aspetti notevoli presenti nei dati come, ad esempio, posizione, variabilità, simmetria, curtosi, ecc.;
- tutte le informazioni sul meccanismo generatore dei dati.

Esistono anche **modelli per dati dipendenti e/o non identicamente distribuiti** (ad esempio per serie storiche e spaziali) e modelli che prescindono dalla forma della distribuzione di probabilità delle variabili casuali del campione (**modelli non parametrici**).

Esempio. *Controllo di qualità* (continua). Per effettuare un controllo di qualità si analizzano n oggetti, scelti a caso, tra quelli prodotti da un certo macchinario.

Il campione osservato $x = (x_1, \dots, x_n)$ 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 osservazioni sono state effettuate in modo indipendente e nelle medesime condizioni, è ragionevole ipotizzare che le X_i , $i = 1, \dots, n$, siano indipendenti con distribuzione $Ber(p)$.

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, lo spazio parametrico è $\Theta = (0, 1)$ e lo spazio campionario è $S_X = \{0, 1\} \times \dots \times \{0, 1\} = \{0, 1\}^n$, cioè l'insieme di tutti i possibili vettori n -dimensionali costituiti da 0 e 1.



Esempio. *Misurazioni* (continua). Si misura un determinato oggetto con uno strumento affetto da errore non sistematico. Il campione osservato $x = (x_1, \dots, x_n)$ sarà costituito da una sequenza di numeri, che corrispondono alle $n \geq 1$ misurazioni.

Se le n osservazioni sono state effettuate in modo indipendente e nelle medesime condizioni, è ragionevole ipotizzare che le X_i , $i = 1, \dots, n$, siano indipendenti con distribuzione $N(\mu, \sigma^2)$.

Si può verificare graficamente l'ipotesi di normalità considerando, ad esempio, istogrammi e q-q plot.

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

Inoltre, lo spazio parametrico è $\Theta = \mathbf{R} \times \mathbf{R}^+$ e lo spazio campionario $S_X = \mathbf{R} \times \dots \times \mathbf{R} = \mathbf{R}^n$ corrisponde all'insieme dei vettori reali n -dimensionali.



Verifica del modello

In alcuni casi, soprattutto per dati continui, la specificazione del modello statistico parametrico può non essere banale, ed è allora necessario procedere ad un **controllo empirico del modello**.

A tal fine, come visto nel caso del modello normale, si possono utilizzare alcuni strumenti già considerati in precedenza, come ad esempio:

- l'**istogramma delle frequenze relative** e la **stima della densità**;
- i grafici dei quantili (**q-q plot**).

L'*istogramma* e la *stima della densità* basate sui dati campionari possono essere interpretate, in ambito inferenziale, come **stime della funzione di densità**.

Tali stime sono valide a prescindere da quale sia la vera distribuzione dei dati e sono tanto più precise quanto più l'informazione campionaria aumenta.

Procedure inferenziali

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 θ :

- la **stima puntuale**: si vuole ottenere, sulla base dell'osservazione x , una *congettura puntuale* su θ ;
- la **stima intervallare** o **regione di confidenza**: si vuole ottenere, sulla base dell'osservazione x , un *sottoinsieme (intervallo)* di Θ in cui è plausibilmente incluso θ ;
- **verifica di ipotesi**: data una congettura o un'*ipotesi* su θ , si vuole verificare, sulla base dell'osservazione x , se essa è accettabile (cioè in accordo con i dati x).

Una procedura statistica deve fornire buoni risultati qualsiasi sia il vero valore del parametro θ e deve essere utilizzabile con riferimento ad ogni possibile campione osservato x .

Sommario

- 1 Sommario e introduzione
- 2 Campionamento
- 3 Modelli statistici
- 4 Statistiche campionarie**

Statistiche campionarie

Ogni analisi statistica inferenziale è caratterizzata da una componente di incertezza, poiché i dati campionari x sono interpretati come realizzazione 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, in termini di probabilità, sul suo grado di affidabilità/incertezza.

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

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

La scelta della statistica riassuntiva T deve essere fatta tenendo conto del modello statistico e dell'obiettivo dell'inferenza.

Il **valore osservato** $t = t(x_1, \dots, x_n)$ di T è un'opportuna *sintesi dei dati campionari osservati* x_1, \dots, x_n , utile per l'inferenza su θ .

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 con una determinata distribuzione di probabilità, chiamata **distribuzione campionaria**.

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

La distribuzione di probabilità di T , che è una funzione di $X = (X_1, \dots, X_n)$, dipende dal parametro ignoto θ . Quindi, va intesa **sotto** θ , cioè nell'ipotesi che θ sia il vero valore del parametro, *qualunque esso sia*.

Dato un campione casuale X_1, \dots, X_n , sono esempi di statistiche campionarie:

- la **somma campionaria** $S_n = \sum_{i=1}^n X_i$, la **media campionaria** $\bar{X}_n = n^{-1} S_n$, la **varianza campionaria** $S^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ e la sua versione **corretta** S_c^2 ;
- le **statistiche ordinate** $X_{(1)} \leq \dots \leq X_{(n)}$, dove $X_{(i)}$ è la variabile casuale che occupa l' i -esima posizione nel campione;
- la **variabile casuale minimo** $X_{(1)} = \min\{X_1, \dots, X_n\}$ e la **variabile casuale massimo** $X_{(n)} = \max\{X_1, \dots, X_n\}$;
- la **mediana campionaria**, definita da $X_{((n+1)/2)}$, se n è dispari, e da $(X_{(n/2)} + X_{((n/2)+1)})/2$, se n è pari;
- i **momenti campionari**, centrati e non centrati, $n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^r$, $n^{-1} \sum_{i=1}^n X_i^r$, $r \in \mathbf{N}^+$.

Nel seguito, oltre che definire le sintesi campionarie S_n , \bar{X}_n , S^2 e S_c^2 , si presenteranno le loro proprietà, considerando opportuni risultati di calcolo delle probabilità.

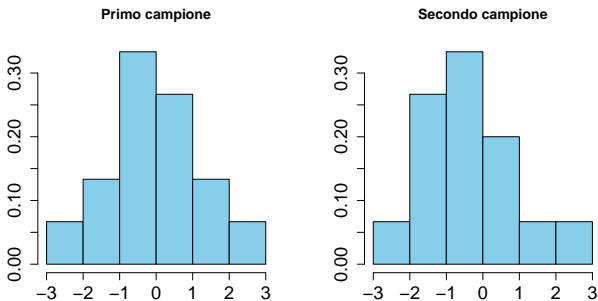
Esempio. Campioni gaussiani. Sia X_1, \dots, X_n un campione casuale semplice tratto da una popolazione normale, cioè $X_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$. Si è interessati al parametro $\theta = \mu$ (*media della popolazione*) ed è ragionevole considerare, come sintesi del campione, la media campionaria $\bar{X}_n = (1/n) \sum_{i=1}^n X_i$ (*media del campione*).

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

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

I due campioni osservati x e x' sono diversi e danno origine a due realizzazioni diverse per la media campionaria: rispettivamente, $\sum_{i=1}^{15} x_i/15 = -0.05$ e $\sum_{i=1}^{15} x'_i/15 = -0.31$.

Si determinano gli istogrammi delle frequenze relative riferiti ai campioni osservati x e x' .



Dalla analisi degli istogrammi si può ragionevolmente confermare che x e x' provengono dalla medesima popolazione.

Quindi, nei due casi, l'inferenza su μ porterà a conclusioni simili, anche se non uguali per effetto della variabilità campionaria. \diamond

Somma e media campionaria

Sia X_1, \dots, X_n un campione casuale semplice tratto da una determinata popolazione, si definiscono, rispettivamente, **somma campionaria** (somma del campione) e **media campionaria** (media del campione) le variabili casuali

$$S_n = \sum_{i=1}^n X_i, \quad \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i = \frac{S_n}{n}.$$

Poiché le X_1, \dots, X_n sono **indipendenti** e **identicamente distribuite** (quindi anche con la stessa media μ e la stessa varianza σ^2), allora

$$E(S_n) = \sum_{i=1}^n E(X_i) = n\mu, \quad V(S_n) = \sum_{i=1}^n V(X_i) = n\sigma^2,$$

$$E(\bar{X}_n) = \frac{E(S_n)}{n} = \mu, \quad V(\bar{X}_n) = \frac{V(S_n)}{n^2} = \frac{\sigma^2}{n}.$$

Se le variabili casuali X_1, \dots, X_n sono **gaussiane** $N(\mu, \sigma^2)$, allora anche *somma e media campionaria* sono variabili casuali *gaussiane*; più precisamente,

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

Valgono, inoltre, i seguenti risultati con riferimento a variabili casuali X_1, \dots, X_n indipendenti:

- se $X_i \sim Bi(k_i, p)$, $i = 1, \dots, n$, allora $S_n \sim Bi(\sum_{i=1}^n k_i, p)$;
- se $X_i \sim P(\lambda_i)$, $i = 1, \dots, n$, allora $S_n \sim P(\sum_{i=1}^n \lambda_i)$;
- se $X_i \sim \chi^2(r_i)$, $i = 1, \dots, n$, allora $S_n \sim \chi^2(\sum_{i=1}^n r_i)$.

In generale, non è detto che risultati simili valgano per modelli diversi da quelli sopra considerati..

La *media campionaria* \bar{X}_n (media del campione) è utile in ambito inferenziale quando si vuole fare *inferenza su μ* (media della popolazione).

Quindi, la media campionaria \bar{X}_n definisce uno **stimatore** per μ e il suo valore osservato \bar{x}_n , che corrisponde alla media calcolata sul campione, viene utilizzato come **stima** per μ .

La bontà della conclusione inferenziale ottenuta si può valutare indirettamente, considerando le proprietà dello stimatore \bar{X}_n .

Per quanto detto sul valore atteso e sulla varianza di \bar{X}_n , si conclude che, in questo caso, lo stimatore si distribuisce attorno a μ e, all'aumentare della dimensione n del campione, la sua varianza, che è pari a σ^2/n , diminuisce.

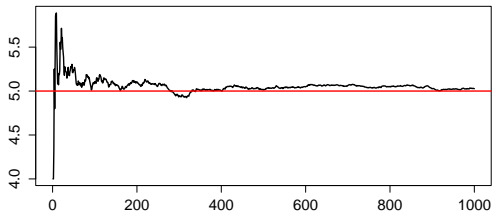
Quindi, al crescere di n , la variabile casuale media campionaria ha una distribuzione di probabilità sempre più concentrata attorno alla media della popolazione μ .

Formalmente, si afferma che vale la **legge debole dei grandi numeri**, cioè che, nelle condizioni poste in precedenza, se $n \rightarrow +\infty$,

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

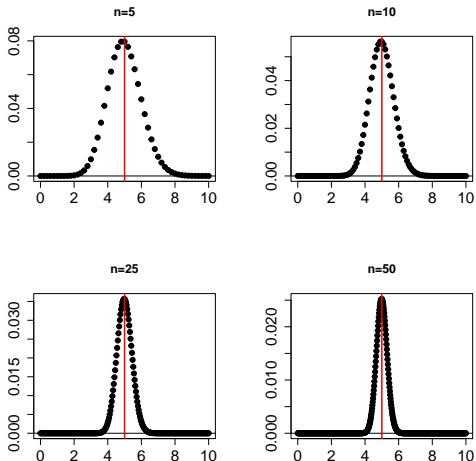
Con la scrittura \xrightarrow{P} si indica la **convergenza in probabilità**, una nozione di convergenza probabilistica illustrata nel seguente esempio.

Esempio. *Media di conteggi.* Si consideri la sequenza dei valori osservati delle medie campionarie di $n = 1, \dots, 1000$ variabili casuali indipendenti con distribuzione $P(5)$.



Si noti che al crescere di n i valori osservati della media campionaria tendono ad essere sempre più concentrati attorno alla media della popolazione $\mu = \lambda = 5$ (in rosso).

Si considerano le funzioni di probabilità di \bar{X}_n per $n = 5, 10, 25, 50$.



Si vede che, al crescere di n , la distribuzione di probabilità della media campionaria è sempre più concentrata attorno a 5, la media della popolazione.



Per la le variabili casuali **somma** e **media campionaria** vale un importante risultato: il **teorema limite centrale**.

Data una *successione di variabili casuali* X_i , $i \geq 1$, *indipendenti e identicamente distribuite*, con media μ e varianza $\sigma^2 \neq 0$ finite, allora la **somma standardizzata** e la **media campionaria standardizzata** coincidono e sono tali che, per $n \rightarrow +\infty$,

$$\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} Z \sim N(0, 1).$$

La scrittura \xrightarrow{d} indica la **convergenza in distribuzione**: al crescere di n la distribuzione di probabilità è sempre più simile a quella di Z .

Per n fissato sufficientemente elevato (almeno $n > 30$), valgono le seguenti utili *approssimazioni*:

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

dove \sim indica la distribuzione approssimata.

È noto che, se le variabili casuali X_1, \dots, X_n hanno distribuzione $N(\mu, \sigma^2)$, allora la somma campionaria S_n e la media campionaria \bar{X}_n sono gaussiane.

Per il teorema limite centrale, se n è sufficientemente elevato, si possono ancora utilizzare tali **distribuzioni gaussiane (approssimate)** per S_n e \bar{X}_n , anche se le variabili casuali X_1, \dots, X_n non hanno distribuzione gaussiana.

In particolare, per n fissato sufficientemente elevato, valgono le seguenti relazioni approssimate: per ogni $a, b \in \mathbf{R}$, $a < b$,

$$P(a < \bar{X}_n \leq b) \doteq \Phi\left(\frac{b - \mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a - \mu}{\sigma/\sqrt{n}}\right),$$

$$P(a < S_n \leq b) \doteq \Phi\left(\frac{b - n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{a - n\mu}{\sigma\sqrt{n}}\right).$$

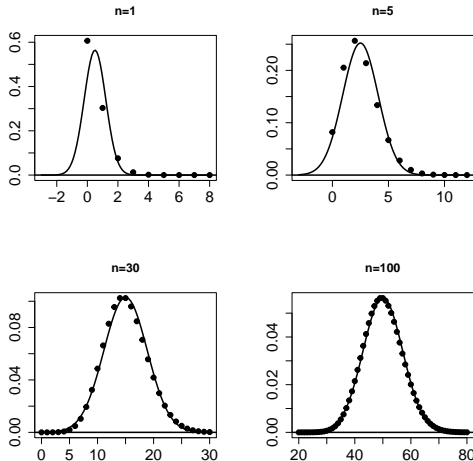
Si possono evidenziare, a questo proposito, i seguenti casi di interesse per le applicazioni:

- se $X_i \sim \text{Ber}(p)$, $i = 1, \dots, n$, allora $S_n \dot{\sim} N(np, np(1-p))$ e $\bar{X}_n \dot{\sim} N(p, p(1-p)/n)$ (si ottengono buone approssimazioni per n tale che $np \geq 5$ e $n(1-p) \geq 5$);
- se $X_i \sim P(\lambda)$, $i = 1, \dots, n$, allora $S_n \dot{\sim} N(n\lambda, n\lambda)$ e $\bar{X}_n \dot{\sim} N(\lambda, \lambda/n)$ (si ottengono buone approssimazioni per n tale che $n\lambda > 10$).

Esempio. Somma di conteggi. Si consideri una successione $\{X_n\}_{n \geq 1}$ di variabili casuali X_n , $n \geq 1$, indipendenti con distribuzione $P(\lambda)$.

È noto che $S_n \sim P(n\lambda)$; inoltre, dal teorema limite centrale, si conclude che, se n è elevato, $S_n \dot{\sim} N(n\lambda, n\lambda)$.

Si confrontano le distribuzioni di probabilità esatte e gaussiane approssimate per S_n , con $n = 1, 5, 30, 100$ e $\lambda = 0.5$.



Al crescere di n le funzioni di probabilità esatte sono sempre più simili a funzioni di densità di una opportuna distribuzione gaussiana (linea continua). L'approssimazione è già accettabile per $n = 30$.




Esempio. *Procedura di controllo.* Si è verificato un inconveniente su una linea di produzione che determina la presenza di $1/10$ di pezzi difettosi.

La procedura di controllo della qualità prevede che, se si individuano almeno 5 pezzi difettosi su $n \geq 1$ scelti a caso, il processo viene posto in revisione. Sia S_n la somma di $n \geq 1$ variabili casuali $Ber(1/10)$ indipendenti.

Si cerca il valore per n tale che ci sia una probabilità pari a 0.9 di porre il processo in revisione. Quindi, $n \geq 1$ deve essere tale che

$$P(S_n \geq 5) = P\left(\frac{S_n - (n/10)}{\sqrt{n9/100}} \geq \frac{5 - (n/10)}{\sqrt{n9/100}}\right) \doteq P\left(Z \geq \frac{5 - (n/10)}{\sqrt{n9/100}}\right)$$

sia 0.9, con $Z \sim N(0, 1)$. Poiché il valore critico $z_{0.9} = -1.282$, si cerca n tale che $[5 - (n/10)]/\sqrt{n9/100} \doteq -1.282$, con $n \geq 50$.

Si ottiene come soluzione il valore 85.58, quindi $n = 86$ può essere una scelta ragionevole. 

Varianza campionaria

Sia X_1, \dots, X_n un campione casuale semplice tratto da una determinata popolazione, si definisce **varianza campionaria** (varianza del campione) la variabile casuale

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

La varianza campionaria può venire calcolata utilizzando la seguente *regola di calcolo* (analoga a quella vista in precedenza)

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

Poiché le X_i , $i = 1, \dots, n$, sono **indipendenti** e **identicamente distribuite** (quindi con la stessa media μ e la stessa varianza σ^2 finita), si ha che

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

La **varianza campionaria** S^2 (varianza del campione) è utile in ambito inferenziale quando, utilizzando i dati campionari, si vuole fare *inferenza su* σ^2 (varianza della popolazione).

Il valore osservato di S^2 viene indicato con s^2 ; esso corrisponde alla varianza calcolata sul campione effettivamente osservato e può essere utilizzato come **stima** per σ^2 .

La bontà della conclusione inferenziale ottenuta si può valutare indirettamente, considerando le proprietà dello strumento utilizzato; in questo caso, le proprietà della varianza campionaria S^2 , che definisce uno **stimatore** per σ^2 .

Per quanto detto sul valore atteso di S^2 , si conclude che, in questo caso, lo stimatore non si distribuisce attorno a σ^2 poiché presenta un valore atteso inferiore a σ^2 . Si noti che al crescere di n tale differenza diventa trascurabile.

Per ovviare a questo problema, che per n piccolo può essere rilevante, si considera un'opportuna modificazione di S^2 , chiamata **varianza campionaria corretta** e definita come

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

Si vede facilmente che

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

quindi S_c^2 definisce uno stimatore alternativo a S^2 , con valore atteso pari a σ^2 .

Sia per la varianza campionaria che per la sua versione corretta vale la **legge debole dei grandi numeri**, dal momento che, per $n \rightarrow +\infty$,

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

Quindi, al crescere della dimensione n del campione, le associate distribuzioni di probabilità saranno sempre più concentrate attorno alla varianza σ^2 della popolazione .

Se le variabili casuali X_1, \dots, X_n sono gaussiane $N(\mu, \sigma^2)$, allora la **varianza campionaria** e la **varianza campionaria corretta** hanno una distribuzione di probabilità legata al modello χ^2 ; più precisamente,

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

dove $\chi^2(n-1)$ indica un modello *chi-quadrato* di parametro (gradi di libertà) $n-1$.

Inoltre, le variabili casuali media campionaria e varianza campionaria (corretta) sono **indipendenti**.

Sempre nel caso di variabili casuali indipendenti X_1, \dots, X_n con distribuzione $N(\mu, \sigma^2)$, valgono i seguenti risultati che risultano utili per la statistica inferenziale.

Se si standardizza la variabile casuale media campionaria si ha che

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

che è chiamata **media campionaria standardizzata**.

Se al posto di σ si considera $S_c = \sqrt{S_c^2}$, si ha la variabile casuale chiamata **media campionaria studentizzata**. Poiché \bar{X}_n e S_c^2 sono indipendenti e $(n-1)S_c^2/\sigma^2 \sim \chi^2(n-1)$, si ha che

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

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

Il risultato è una conseguenza del fatto che la t di Student si ottiene come rapporto tra una variabile casuale $N(0, 1)$ e una variabile casuale χ^2 indipendente, diviso i suoi gradi di libertà.

Siano X_1, \dots, X_n variabili casuali con distribuzione $N(\mu_X, \sigma_X^2)$ e Y_1, \dots, Y_m variabili casuali con distribuzione $N(\mu_Y, \sigma_Y^2)$; tutte le variabili casuali sono indipendenti.

Indicate 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, si ha che $nS_X^2/\sigma_X^2 \sim \chi^2(n-1)$ e $mS_Y^2/\sigma_Y^2 \sim \chi^2(m-1)$.

Quindi, 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.

Esempio. *Bottiglie.* Una azienda produce giornalmente migliaia di bottiglie di soluzione di glucosio con contenuto nominale 500 ml.

Per tenere sotto controllo il processo di imbottigliamento si estrae casualmente un campione di $n = 25$ bottiglie e si misura il loro contenuto effettivo.

I dati campionari x_1, \dots, x_{25} possono essere ragionevolmente interpretati come osservazioni indipendenti da una popolazione $N(\mu, \sigma^2)$.

Se si è interessati al parametro μ , che definisce il contenuto medio di una generica bottiglia, si può considerare come statistica riassuntiva la media campionaria \bar{X}_n , che con riferimento al campione osservato ha assunto valore 498 ml.

Per poter valutare l'utilità di \bar{X}_n per fare inferenza su μ , e quindi interpretare il valore osservato 498, bisognerebbe conoscere la sua distribuzione di probabilità (distribuzione campionaria).

Dal calcolo delle probabilità è noto che $\bar{X}_n \sim N(\mu, \sigma^2/n)$ e che, se $n \rightarrow +\infty$, $\bar{X}_n \xrightarrow{p} \mu$.

Quindi, le fluttuazioni, da campione a campione, attorno a μ della media campionaria sono sintetizzate dal suo scarto quadratico medio σ/\sqrt{n} .

Al crescere di n , tale valore si riduce di un fattore pari a \sqrt{n} ; questo aspetto è legato alla legge debole dei grandi numeri.

Se si è interessati al parametro σ^2 , che specifica la precisione del processo di imbottigliamento, si può considerare come statistica riassuntiva la varianza campionaria S^2 o la sua versione corretta S_c^2 .

A questo proposito si possono fare valutazioni analoghe a quelle fatte per \bar{X}_n , considerando la distribuzione $\chi^2(n-1)$. \diamond

Esempio. Ricoveri. Il direttore di una clinica vuole capire con che percentuale i suoi pazienti attuali sono stati ricoverati più di una volta nella sua struttura.

Si estrae un campione casuale semplice di $n = 100$ pazienti attualmente ricoverati e si verifica se si tratta del primo ricovero oppure di un ricovero successivo al primo.

I dati campionari x_1, \dots, x_{100} , che costituiscono una sequenza di 0 (primo ricovero) e 1 (ricovero successivo al primo), possono essere ragionevolmente interpretati come osservazioni indipendenti da una popolazione $Ber(p)$.

Nel caso di una popolazione bernoulliana, il valore atteso $\mu = p$ e la varianza $\sigma^2 = p(1 - p)$.

Si è interessati al parametro p , che indica la *proporzione nella popolazione* di pazienti con più di un ricovero, e si considera come statistica riassuntiva la media campionaria \bar{X}_n , che corrisponde alla *proporzione campionaria* di pazienti con più di un ricovero.

Con riferimento al campione osservato, \bar{X}_n ha assunto valore 0.48.

Per poter valutare l'utilità di \bar{X}_n per fare inferenza su p , e quindi interpretare il valore osservato 0.48, bisognerebbe conoscere la sua distribuzione di probabilità, almeno in forma approssimata.

Dal calcolo delle probabilità è noto che, con n elevato,

$\bar{X}_n \sim N(p, p(1-p)/n)$ e che, se $n \rightarrow +\infty$, $\bar{X}_n \xrightarrow{p} p$.

Quindi, le fluttuazioni, da campione a campione, attorno a p della media campionaria sono sintetizzate dal suo scarto quadratico medio

$\sqrt{p(1-p)/n}$.

Al crescere di n , tale valore si riduce di un fattore pari a \sqrt{n} ; questo aspetto è legato alla legge debole dei grandi numeri.

