Appunti di Probabilità e Statistica

a.a. 2014/2015 C.d.L. Informatica – Bioinformatica

I. Oliva

Lezione 6

1 Statistica Inferenziale

Obiettivo: dato un fenomeno casuale reale, in genere non si conosce la forma completa della legge di probabilità cui obbedisce. Si deve, allora, determinare tale legge, sulla base di informazioni relative ad osservazioni empiriche, ricavate da una porzione della popolazione, dove

- **Popolazione:** insieme molto grande di oggetti a cui sono associate delle quantità misurabili.
- **Campione:** s.i. ridotto (porzione) della popolazione, rappresentativo della stessa popolazione di riferimento e libero da qualsiasi elemento soggettivo.

I modelli probabilistici maggiormente adottati sono i seguenti:

- modello uniforme, quando si studia un carattere che può assumere qualunque modalità in un intervallo finito, ma incognito

$$f(x;\theta_1,\theta_2) = \frac{1}{\theta_2-\theta_1},\ x \in [\theta_1,\theta_2]\ ;$$

- modello normale, con media incognita:

$$f(x;\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2}\ ;$$

- modello normale, con varianza incognita:

$$f(x;\theta) = \frac{1}{\theta\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\theta}\right)^2};$$

- modello normale, con media e varianza incognite:

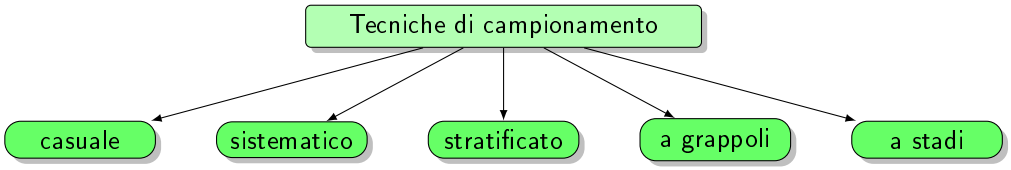
$$f(x;\theta_1,\theta_2) = \frac{1}{\theta_2\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\theta_1}{\theta_2}\right)^2};$$

- modello binomiale, i.e., sequenza di n prove ripetute ed indipendenti, con probabilità di successo incognita:

$$f(x;\theta) = \sum_{k=0}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k};$$

- modello di Poisson:

$$f(x;\theta) = \sum_{k=0}^{+\infty} e^{-\theta} \frac{\theta^k}{k!}.$$



Noi faremo riferimento **solo** al *campionamento casuale semplice (ccs):*

- Campionamento $\Rightarrow$ dare informazioni su grandi popolazioni al *minimo costo,* con la *massima rapiditá* e *maggiore accuratezza* possibili e usando *strumenti raffinati.*

- Casuale $\Rightarrow$ le unità statistiche che entrano a far parte del campione sono estratte in modo *casuale* dalla popolazione di riferimento.

- Semplice $\Rightarrow$ estrazioni indipendenti (*con* reinserimento, se la popolazione è finita, *senza* reinserimento, se è infinita).

NOTA BENE: X_i, $i = 1, \ldots, n$ v.a. con distribuzione che descrive quali sono i valori che possono essere assunti dalla caratteristica su un'unità estratta a caso e con quale probabilità (caso discreto) o densità (caso continuo) verranno osservati. x_i : realizzazione della v.a. X_i, $i = 1, \ldots, n$.

Dunque, un c.c.s. si rappresenta come una n-pla $X_1, \ldots, X_n$ di v.a.i.i.d. con distribuzione F.

In generale, la F è non nota, ma si conosce la famiglia di distribuzioni a cui F appartiene, dunque si utilizza il campione per fare inferenza sui parametri che identificano F (inferenza parametrica). In altri casi, non si ha alcuna informazione su F (inferenza non parametrica).

Noi ci occuperemo solo del primo caso.

Esempio 1.1. *Vogliamo studiare l'altezza della popolazione italiana. Si estrae il seguente c.c.s.* $\{X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8\}$, *le cui realizzazioni sono* $\{x_1 = 170, x_2 = 181, x_3 = 189, x_4 = 160, x_5 = 182, x_6 = 171, x_7 = 158, x_8 = 186\}$.

È ragionevole supporre che l'altezza si distribuisca come una v.a. gaussiana, dunque $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \ldots, 8$.

Per poter fare inferenza sulla popolazione italiana, dovremmo conoscere i parametri della distribuzione. Una soluzione consiste nello stimare tali parametri.

Sorgono però alcune domande:

1. *quanto sono accurate le stime della media e della varianza dell'altezza della popolazione italiana?*

2. *possiamo individuare un intervallo di valori ragionevoli per* μ *e per* σ?

3. *sulla base dei risultati ottenuti, è ragionevole affermare, ad esempio, che l'altezza media della popolazione italiana è superiore a* $170\,cm$?

Alle tre domande precedenti si risponde utilizzando le tecniche di statistica inferenziale, in particolare la stima puntuale, *la* stima intervallare *e la* verifica d'ipotesi.

1.1 Stima puntuale

Quando si conosce la famiglia di appartenenza delle v.a. che formano il c.c.s., non si conoscono i parametri caratterizzanti la distribuzione stessa, ma una sua funzione.

Tale funzione, $T_n = T(X_1, \ldots, X_n)$, si chiama *stimatore (puntuale)* del parametro θ della distribuzione.

La quantità $T(x_1, \ldots, x_n) = \hat{\theta}$ si chiama *stima* (o *statistica*) del parametro.

È bene osservare che, a seconda del campione estratto, si avrà una stima diversa, dunque $T(X_1, \ldots, X_n) = T_n$ avrà una propria distribuzione, detta *distribuzione campionaria.*

Proprietà degli stimatori:

Correttezza. Lo stimatore T_n si dice *corretto* o *non distorto* se il suo valore atteso coincide con il valore teorico del parametro θ da stimare:

$$E(T_n) = \theta \ .$$

Se lo stimatore non è corretto, si dice anche essere *distorto.* La quantità $b(T_n) = E(T_n) - \theta \neq 0$ si chiama *distorsione* o *bias* dello stimatore.

Consistenza. Lo stimatore T_n si dice *consistente (in media quadratica)* se, al crescere della taglia n del c.c.s., la distribuzione campionaria di T_n si concentra sempre di più attorno a θ, cioè

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

La quantità $E[(T_n - \theta)^2]$ si chiama *errore quadratico medio (EQM)* e misura di quanto, in media, le realizzazioni di T_n, su tutti i c.c.s. di taglia n, distano da θ.

Una formula operativa per il calcolo di EQM è la seguente:

$$EQM(T_n) = Var(T_n) + b^2(T_n) \ .$$

Infatti:

$$\begin{aligned} EQM(T_n) &= E[(T_n - \theta)^2] = E[T_n^2 + \theta^2 - 2\theta T_n] \\ &= E[T_n^2] + \theta^2 - 2\theta E[T_n] \\ &= E[T_n]^2 - E[T_n]^2 + E[T_n^2] + \theta^2 - 2\theta E[T_n] \\ &= Var(T_n) + b^2(T_n) \ . \end{aligned}$$

Efficienza. Lo stimatore T_n si dice *efficiente* se rende minimo il grado di dispersione della sua distribuzione, rispetto al parametro θ oggetto di stima.

Dati due stimatori $T_n^{(1)}$ e $T_n^{(2)}$, si dice che $T_n^{(1)}$ è *più efficiente* di $T_n^{(2)}$ se

$$EQM(T_n^{(1)}) \leq EQM(T_n^{(2)}) \ .$$

Se i due stimatori sono corretti, per quanto detto finora, $EQM(T_n) = Var(T_n)$, dunque $T_n^{(1)}$ è più efficiente di $T_n^{(2)}$ se

$$Var(T_n^{(1)}) \leq Var(T_n^{(2)}) \ .$$

Una statistica $T_n = T_n(X_1, \ldots, X_n)$ risulta il miglior stimatore del parametro θ se è il più efficiente tra gli stimatori corretti e consistenti.

Nella pratica, si usa il metodo *BLUE* (Best Linear Unbiased Estimator), che consiste nello scegliere lo stimatore nella classe degli stimatori *lineari, corretti, con varianza minima.* Tale risultato discende dal teorema di Cramer-Rao, che fornisce un valore che minimizza la varianza, dato da

$$-\frac{1}{n \cdot E\left[\frac{\partial^2}{\partial \theta^2}\left(\ln f(x,\theta)\right)\right]} \ .$$

Un metodo per determinare la stima puntuale dei parametri va sotto il nome di *metodo della massima verosimiglianza.* Siano $X_1, \ldots, X_n$ n v.a.i.i.d., con densità $f(x_i, \theta),\ i = 1, \ldots, n$.

Consideriamo la quantità (nota come *funzione di verosimiglianza*)

$$L(X_1, \ldots, X_n; \theta) := \Pi_{i=1}^{n} f(x_i; \theta) \ .$$

Cerchiamo il valore da assegnare al parametro θ affinchè la funzione di verosimiglianza sia massima (i.e., assuma il valore più grande possibile.) Per far questo, se $L(X, \theta)$ è derivabile rispetto a θ, allora esistono le derivate di L. Si procede ponendo uguale a 0 tali derivate e si risolve il sistema così ottenuto. Per alleggerire i calcoli, invece di considerare la funzione di verosimiglianza L, si preferisce studiare la funzione di *log-verosimiglianza,* $\ln(L)$.

Esercizio 1.1. *Sia X una v.a. di Poisson di parametro ignoto λ. Calcolare lo stimatore di massima verosimiglianza di λ.*

Distribuzioni campionarie.

Media campionaria. Sia $\{X_1, \ldots, X_n\}$ un c.c.s. estratto dalla popolazione.

Si definisce *media campionaria* la v.a.

$$\bar{X} := \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i \ .$$

Valore atteso di $\bar{X}$: se $E(X_i) = \mu$, per ogni $i = 1, \ldots, n$, allora

$$E(\bar{X}) = E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{E(X_1) + \cdots + E(X_n)}{n} = \mu \ .$$

Quindi, la distribuzione della media campionaria è centrata sulla media delle singole v.a. che formano il campione. Questo garantisce che la media campionaria è uno stimatore corretto della media di popolazione.

Varianza di $\bar{X}$: se $Var(X_i) = \sigma^2$, per ogni $i = 1, \ldots, n$, allora

$$Var(\bar{X}) = Var\left(\frac{X_1 + \cdots + X_n}{n}\right) = Var\left(\frac{X_1}{n}\right) + \cdots + Var\left(\frac{X_1}{n}\right) = \frac{\sigma^2}{n} .$$

Quindi, la varianza della media campionaria è inversamente proporzionale alla taglia del c.c.s. Inoltre,

$$EQM(\bar{X}) = Var(\bar{X}) \to 0, \text{ se } n \to \infty ,$$

ossia, la media campionaria è anche uno stimatore consistente della media di popolazione.

Se $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$, allora $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. (senza dimostrazione).

Se, invece, il c.c.s. $\{X_1, \ldots, X_n\}$ non proviene da una popolazione gaussiana, ma la taglia del campione è abbastanza grande, e supponendo $m = E(X_i)$, $v^2 = Var(X_i)$, $i = 1, \ldots, n$ si ha $\bar{X} \sim \mathcal{N}(m, v^2/n)$ (la dimostrazione segue immediatamente dall'applicazione del Teorema del limite centrale).

Varianza campionaria. Sia $\{X_1, \ldots, X_n\}$ un c.c.s. estratto dalla popolazione. Si definisce *varianza campionaria* la v.a.

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

Valore atteso di S^2 : innanzitutto, osserviamo che

$$\begin{aligned} S^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 + \bar{X}^2 - 2\bar{X}X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \frac{1}{n}\sum_{i=1}^{n}\bar{X}^2 - 2\frac{1}{n}\sum_{i=1}^{n}\bar{X}X_i \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \bar{X}^2 - 2\bar{X}\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{n\bar{X}} \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \bar{X}^2 - 2\bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 \end{aligned}$$

Dunque, si ha:

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

D'altro canto, posto $m = E(X_i)$ e $v^2 := Var(X_i)$, $i = 1, \ldots, n$ e ricordando la formula operativa $E(X_i^2) = Var(X_i) + E(X_i)^2 = v^2 + m^2$ della varianza di v.a., avremo quindi

$$\begin{aligned} E(S^2) &= \frac{1}{n}\sum_{i=1}^{n}(Var(X_i) + E(X_i)^2) - (Var(\bar{X}) + E(\bar{X})^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}(v^2 + m^2) - \left(\frac{v^2}{n} + m^2\right) \\ &= v^2 + m^2 - \frac{v^2}{n} - m^2 = (n-1)\frac{v^2}{n}. \end{aligned}$$

Questo ci dice che la varianza campionaria non è uno stimatore corretto, in quanto $E(S^2) \neq \frac{v^2}{n}$.

Si può correggere la varianza campionaria nel seguente modo:

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

$\bar{S}^2$ si chiama *varianza campionaria corretta.* Inoltre, si può dimostrare (ma noi non lo faremo) che

$$EQM(\bar{S}^2) = Var(\bar{S}^2)\ \to\ 0,\ \text{se}\ n\ \to\ \infty\ ,$$

ossia, che la varianza campionaria corretta è anche uno stimatore consistente della varianza di popolazione.

Se il c.c.s. proviene da una popolazione gaussiana, $\bar{S}^2$ è uno stimatore efficiente (senza dimostrazione).

Si dimostra che

$$S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}\ ,$$

da cui

$$\bar{S}^2 \sim \chi^2_{n-1}\ .$$

Frazione campionaria di successi. Sia $\{X_1, \ldots, X_n\}$ un c.c.s. estratto da una popolazione bernoulliana, i.e. $X_i \sim Bin(1, p)$. Si definisce *frazione campionaria di successi* la v.a.

$$\pi := \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i \, .$$

Valore atteso di π : essendo $E(X_i) = p$, per ogni $i = 1, \ldots, n$, allora

$$E(\pi) = E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{E(X_1) + \cdots + E(X_n)}{n} = p \, .$$

Quindi, la distribuzione della frazione campionaria di successi è uno stimatore corretto di una probabilità di successo p.

Varianza di π : essendo $Var(X_i) = p(1-p)$, per ogni $i = 1, \ldots, n$, allora

$$Var(\pi) = Var\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n^2} \sum_{i=0}^{n} Var(X_i) = \frac{p(1-p)}{n} \, .$$

Quindi, la varianza della frazione campionaria di successi è inversamente proporzionale alla taglia del c.c.s. Inoltre,

$$EQM(\pi) = Var(\pi) \to 0, \text{ se } n \to \infty \, ,$$

ossia, la frazione campionaria di successi è anche uno stimatore consistente di una probabilità di successo p.

Se la taglia del campione è abbastanza grande, si ha $\pi \sim \mathcal{N}(p, p(1-p)/n)$ (la dimostrazione segue immediatamente dall'applicazione del Teorema del limite centrale).

1.2 Stima intervallare

Con la stima puntuale, si ottiene un valore empirico *approssimato* del parametro da stimare. Quindi, se si conoscesse la distribuzione campionaria della statistica usata per la stima, si potrebbe valutare il grado di errore commesso e determinare, con una probabilità molto prossima a 1, l'intervallo in cui si trova il valore vero del parametro che stiamo stimando.

Si chiama *intervallo di confidenza* l'intervallo $[\theta_{min}, \theta_{max}]$ tale che

$$P(\theta_{min} \leq \theta \leq \theta_{max}) > \alpha \, ,$$

dove α rappresenta il *livello di significatività* misurato in percentuale, della stima effettuata.

Costruzione di un intervallo di confidenza.

1. Si sceglie il livello di confidenza α (oppure $1-\alpha$)
2. Si identifica uno stimatore T di θ
3. Si cerca una funzione di T e di θ, detta *quantità pivotale* e indicata con $Q(T,\theta)$, la cui distribuzione di probabilità sia completamente nota
4. Indichiamo con $q_{\alpha/2}$ e $q_{1-\alpha/2}$ i quantili della quantità pivotale di ordine $\alpha/2$ e $1-\alpha/2$, rispettivamente, e si scrive l'intervallo di confidenza come segue:
$$P(q_{\alpha/2} \leq \theta \leq q_{1-\alpha/2}) > \alpha$$
5. Si ricava l'intervallo di confidenza, risolvendo la disequazione rispetto al parametro θ

IC per la media di una popolazione normale– Varianza σ^2 nota Una volta fissato il livello di significatività, osserviamo che, da quanto detto a proposito della stima puntuale, uno stimatore per la media μ corretto, consistente ed efficiente è la media campionaria $\bar{X}$.

Poichè tale stimatore ha una distribuzione normale, occorre prima standardizzarlo:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow Z = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0,1)\ ,$$

dunque, l'intervallo di confidenza che cerchiamo ha la seguente forma:

$$\begin{aligned} & P\left(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}\right) = 1-\alpha \\ \Rightarrow & P\left(\bar{X} - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) = 1-\alpha \end{aligned}$$

Il valore cercato è il quantile di ordine $1-\alpha/2$, ossia $z = z_{1-\alpha/2}$, ottenuto dalle tavole della funzione di ripartizione. Dunque, l'intervallo di confidenza è

$$\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]\ .$$

Esistono anche dei problemi inversi:

1. Determinare il livello di significatività α tale che $\mu = \bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$.
   Si ha:

$$
\begin{aligned}
|\mu - \bar{X}| &= \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \\
\Rightarrow\ z_{1-\alpha/2} &= \frac{|\mu - \bar{X}|\sqrt{n}}{\sigma} \\
\Rightarrow\ 1 - \alpha/2 &= \Phi\left(\frac{|\mu - \bar{X}|\sqrt{n}}{\sigma}\right) \\
\Rightarrow\ \alpha &= 2\left[1 - \Phi\left(\frac{|\mu - \bar{X}|\sqrt{n}}{\sigma}\right)\right]
\end{aligned}
$$

2. Determinare la taglia del c.c.s. a partire dall'intervallo di confidenza.
   Se $IC = \left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$, allora l'ampiezza dell'intervallo è

$$
\begin{aligned}
A &= z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \\
\Rightarrow n &= \left(\frac{2\sigma z_{1-\alpha/2}}{A}\right)^2
\end{aligned}
$$

Esercizio 1.2. *Si supponga che il peso (in kg) di certe confezioni abbia distribuzione normale di media μ ignota e deviazione standard pari a $2,5\,kg$. Su un c.c.s. di 100 confezioni è stato calcolato un peso medio di $11,5\,kg$. Costruire un intervallo di confidenza di livello 95% per μ.*

IC per la media di una popolazione normale – Varianza σ^2 ignota

Se anche la varianza è ignota, occorre prima stimarla in qualche modo per poter costruire l'intervallo di confidenza richiesto. La stima puntuale garantisce che possiamo sostituire σ^2 con $\bar{S}^2$, quindi possiamo considerare la v.a. media campionaria

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

da cui si ha

$$
\begin{aligned}
\frac{\bar{X} - \mu}{\sqrt{\frac{\bar{S}^2}{n}}} &= \frac{\bar{X} - \mu}{\sqrt{\frac{\bar{S}^2(n-1)\sigma^2}{n(n-1)\sigma^2}}} \\
&= \frac{(\bar{X} - \mu)\sqrt{\frac{\sigma^2}{n}}}{\sqrt{\frac{\bar{S}^2(n-1)}{(n-1)\sigma^2}}} \sim t_{n-1}
\end{aligned}
$$

L'intervallo di confidenza in questo caso è

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

Dunque, l'intervallo di confidenza è

$$\left[\bar{X} - t_{n-1;1-\alpha/2}\sqrt{\frac{\bar{S}^2}{n}}; \bar{X} + t_{n-1;1-\alpha/2}\sqrt{\frac{\bar{S}^2}{n}}\right]$$

oppure

$$\left[\bar{X} - t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n-1}}; \bar{X} + t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n-1}}\right] .$$

IC per la varianza di una popolazione normale Vogliamo costruire un intervallo di confidenza di livello $1-\alpha$ per σ^2, quando anche μ è ignoto. Una volta fissato α, scegliamo lo stimatore più adatto, che, in questo caso, è la varianza campionaria corretta.

Poichè, per una popolazione normale,

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

avremo

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

Dunque, l'intervallo di confidenza per σ^2 è

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

oppure

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

se invece volessimo un intervallo di confidenza per la deviazione standard.

Esercizio 1.3. *Una nuova terapia è stata sperimentata su un campione di* 12 *pazienti e i tempi di guarigione osservati sono stati (in giorni)*

$$\{15,\ 23,\ 32,\ 18,\ 25,\ 16,\ 27,\ 22,\ 30,\ 41,\ 18,\ 29\}\ .$$

1. *Assumendo una distribuzione normale per il tempo di guarigione, trovare un intervallo di confidenza di livello* 95% *per il tempo medio di guarigione dei pazienti sottoposti a terapia.*
2. *Si dia un intervallo di confidenza al* 90% *per la varianza* σ^2.

IC per una probabilità Sia $\{X_1, \ldots, X_n\}$ un c.c.s. estratto da una popolazione bernoulliana (i.e., $X_i \sim Bin(1, p)$, $i = 1, \ldots, n$), di cui non si conosce la probabilità di successo p.

Una volta fissato il livello di significatività α, ricordiamo che uno stimatore corretto, consistente ed efficiente per p è

$$\pi = \frac{1}{n}\sum_{i=0}^{n} X_i \sim stand\left(p, \frac{p(1-p)}{n}\right)\ .$$

Se standardizziamo la v.a. p, possiamo costruire il nostro intervallo di confidenza:

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

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

Dunque, l'intervallo di confidenza per σ^2 è

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

Nell'espressione precedente, si vede che l'intervallo di confidenza dipende dal parametro p che, in realtà, non conosciamo. Allora, possiamo sostituire p con π, in virtù del Teorema del limite centrale , purchè si abbiano almeno 10 osservazioni, di cui almeno 5 successi ed altrettanti insuccessi. In tal caso, l'intervallo di confidenza diventa

$$\left[\pi - z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}; \pi + z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right]\ .$$

Esercizio 1.4. *Il direttore di una banca intende studiare la proporzione di depositanti che vengono pagati mensilmente. Vengono scelti* 200 *depositanti e* 23 *di questi affermano di ricevere il pagamento ogni mese. Stimare la vera proporzione di depositanti della banca pagati mensilmente e costruire un intervallo di confidenza di livello* 90% *per tale proporzione.*

Esercizio 1.5. *Un laboratorio farmaceutico deve calcolare la concentrazione di principio attivo in un dato composto chimico. I risultati dell'analisi non sono certi, ma, ripetuti, mostrano un andamento descrivibile attraverso una distribzuione normale.*

Dato il c.c.s. $\{3.853, 3.588, 3.954\}$ *(misurazioni effettuate in g/l), determinare un intervallo di confidenza per la concentrazione media di principio attivo al* 90%.

Esercizio 1.6. *Tra i pasticcini prodotti artigianalmente in una pasticceria, se ne prelevano* 100. *Risulta che il loro peso medio è* $35\,g$. *Si sa che lo scarto quadratico medio del peso dei pasticcini prodotti è* $4\,g$.

- *Determinare un intervallo di confidenza per il peso medio dei pasticcini prodotti, ad un livello di significatività del* 98%.
- *Di quanto deve aumentare la numerosità campionaria se si vuole che l'ampiezza dell'intervallo si dimezzi?*
- *Determinare quanti pasticcini occorre estrarre se si vuole che lo stimatore del peso medio si discosti dal vero peso medio per meno di un grammo, con probabilità del* 96%.