Inferenza I

Fondamenti della teoria della stima

- Campionamento bernoulliano ed in blocco
- Problema della stima: stima e stimatore
- Proprietà di uno stimatore
- Stima puntuale e per intervallo: valore atteso e varianza

Inferenza statistica

- Inferenza statistica: branca della statistica che cerca di ricavare informazioni relative ad una intera popolazione partendo dall'analisi di un campione.
- *Domande aperte*:
  - Come faccio il campione?
  - Come descrivo una generica popolazione?
  - Che tipo di di informazioni posso ottenere?

Scelta del campione

- Campionamento: processo di formazione del campione.
- Esiste una letteratura infinita sulla scelta del campione.
- Due diverse filosofie di campionamento
  - Estrazione bernoulliana: le n unita statistiche vengono estratte una alla volta e dopo l'estrazione sono nuovamente estraibili.
  - Estrazione in blocco: le n unita statistiche vengono estratte in blocco (non è possibile per una singola unità comparire più volte).
- Tratteremo solo casi di estrazioni bernoulliane.

Popolazione

- Ruolo:
  - essa fornisce le osservazioni.
  - modella uno o più caratteri di un gruppo di unità statistiche.
- Osservazione: il campionamento bernoulliano garantisce che in ogni estrazione un'osservazione ha la stessa probabilità di verificarsi
- Solitamente si descrivere la popolazione come una v.c. P avente d.d.p. (funzione di probabilità) incognita.

Modellazione

- Popolazione: v.c. P con d.d.p. $f(p)$.
- Osservazione della i-sima unità statistica: v.c. X_i
  - $X_1 \sim P$ (la prima estrazione la faccio da P).
  - Nessuna garanzia che la d.d.p. delle estrazioni successive:
    - resti constante ($f(x_i) = f(x_j)$).
    - sia uguale a quella di P ($X_i \sim P$).
- Se si campiona con estrazione bernoulliana si ha che
  - X_i sono i.i.d.
  - $X_i \sim P$ da cui ottengo che $E[X_i] = E[P], Var[X_i] = Var[P]$.

Informazioni ottenibili

Le informazioni si dividono in due diverse tipologie

1. Cerco di ottenere una stima numerica di una caratteristica (spesso un indice) della popolazione.
   - Esempi:
     - Stimare il valore atteso della popolazione.
     - Stimare la varianza della popolazione.
   - Strumento teorico: <u>Teoria della stima</u>.
2. Cerco di rispondere ad una domanda dall'esito binario
   - Esempio:
     - La variabile X è normale?
     - (se P è multi-variata) P_1 e P_2 sono indipendenti?
   - Strumento teorico: <u>Test non parametrici</u>.

Teoria della stima

- Esempi:
  - Stimare il valore atteso della popolazione.
  - Stimare la varianza della popolazione.
- "Ingredienti" comuni ai vari problemi di stima:
  - Dati di partenza:
    - n osservazioni $O = \{o_i\}$
  - Obiettivo:
    - stima di un <u>parametro</u> θ della della popolazione.
  - Mezzo:
    - Una funzione $g(.)$ dei dati chiamata <u>stimatore</u>
  - Risultato
    - Una <u>stima</u> del parametro $\hat{\theta} = g(O)$

Problema della stima

Problema: dato un campione O di dimensione n estratto da una popolazione P, avente un parametro incognito θ, determinare una funzione $g(.)$ chiamata stimatore che fornisca una stima $\hat{\theta} = g(O)$ di θ.

- Esempio Stimare il valore atteso della popolazione.
  - Parametro $\theta_1 = E[P]$.
  - Stimatore $g_1(.)$
  - Stima $\hat{\theta}_1 = g_1(O)$
- Esempio: Stimare la varianza della popolazione.
  - Parametro $\theta_2 = Var[P]$.
  - Stimatore $g_2(.)$
  - Stima $\hat{\theta}_2 = g_2(O)$

Stimatore: considerazioni.

- Esempio: Uso giornaliero dei mezzi pubblici
  - Popolazione: cittadini di Vr
  - Campione di $n = 100$ persone
  - v.c. X_i risposta del i-simo intervistato.
  - Esempio di stimatore. (Media campionaria) $$g(.) = \frac{X_1 + X_2 + X_3 + ... + X_n}{n}$$
- Osservazione: Il valore dello stimatore (stima) dipende da n eventi casuali (l'estrazione delle unità statistiche). Quindi si ha che:
  - Lo stimatore è una v.c. Θ
    - Ha una d.d.p. da cui un valore atteso e una varianza
  - La stima $\hat{\theta}$ è una realizzazione dello stimatore

Stimatore: proprietà - I

Quali caratteristiche vorrei avesse uno stimatore?

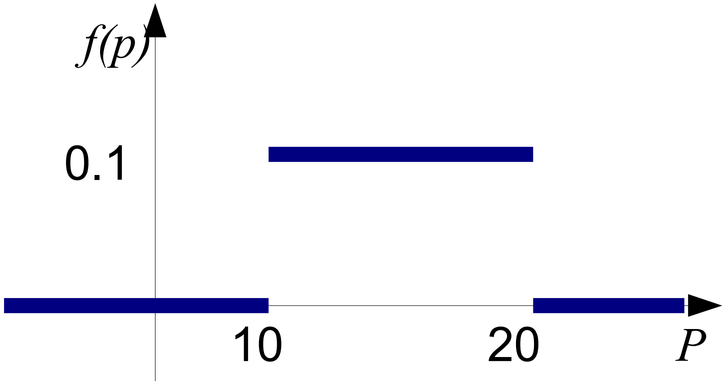
- Correttezza: il valore atteso dello stimatore è il parametro da stimare

$$E\left[\Theta\right]=\theta$$

- Esempio:
  - Popolazione P uniforme
  - $Var[P] = 100/12 = 25/3$
  - Stimatore di $Var[P]$ corretto

    $$E\left[\Theta\right]=\frac{25}{3}$$

  - Possibili d.d.p. di uno stimatore corretto



$$N\left(\frac{25}{3}\ ;\ 2\right) \qquad N\left(\frac{25}{3}\ ;\ 4\right) \qquad \chi^2\left(\frac{25}{3}\right)$$

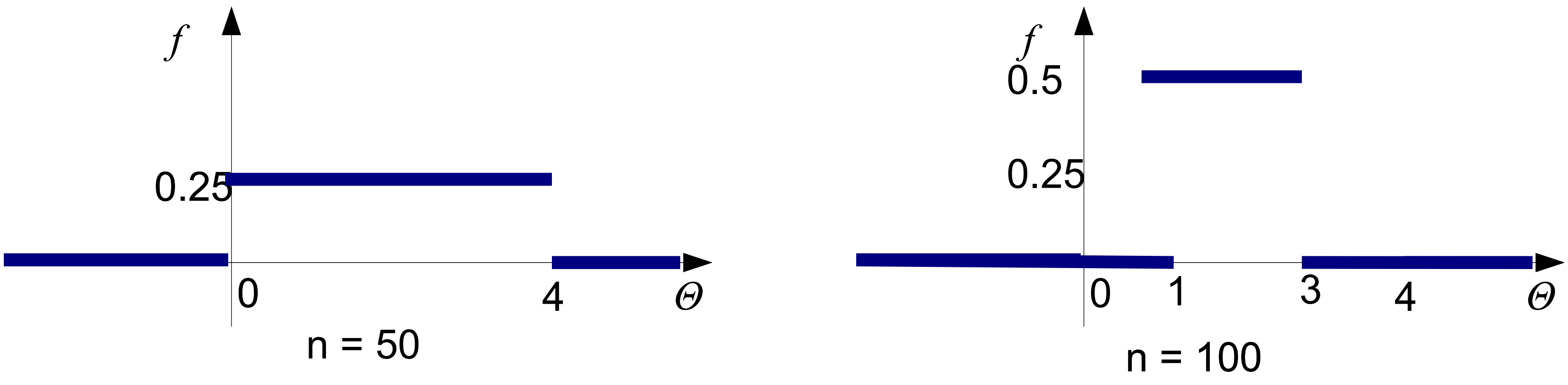
Stimatore: proprietà - II

Quali caratteristiche vorrei avesse uno stimatore?

- Consistenza: al crescere della dimensione del campione le stime son sempre più vicine al parametro

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}-\theta\right|>\varepsilon\right)=0 \qquad \forall\, \varepsilon>0$$

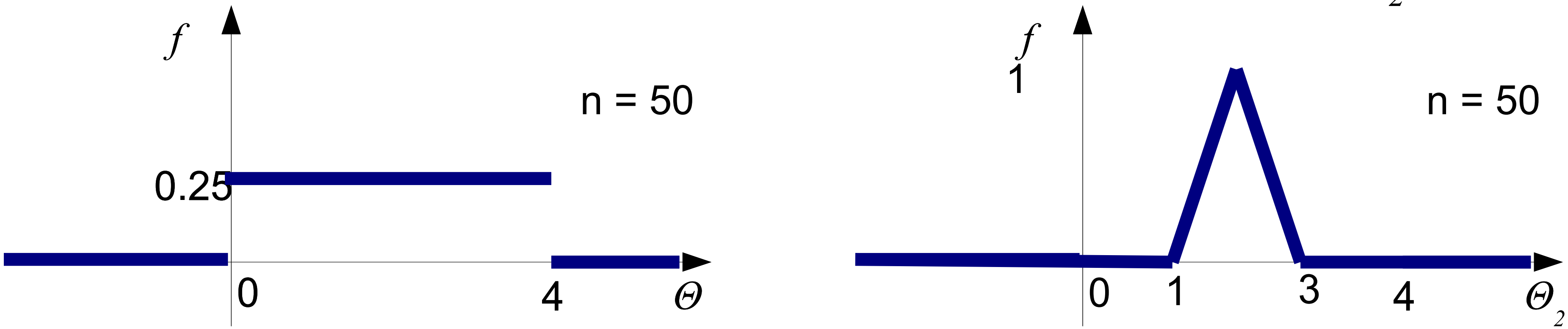
- Esempio
  - Popolazione $P\sim \mathrm{N}(10;2)$
  - Stimatore della Varianza corretto



Stimatore: proprietà - III

Quali caratteristiche vorrei avesse uno stimatore?

- Efficienza: lo stimatore possiede la varianza minima.
  (utile per il confronto fra più stimatori:
  scelgo quello con la varianza minore)
- Esempio
  - Popolazione $P \sim N(10 \,;\, 2)$
  - 2 Stimatori della Varianza corretti $E[\Theta]=E[\Theta_2] = 2$



Migliore perché ha varianza minore

Media campionaria

- Si indica sovra segnando la grandezza mediata

- Definizione:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

- Osservazione: la media campionaria è una combinazione lineare di valori su cui è calcolata

- Diverse interpretazioni
  - Indice di posizione (statistica descrittiva)
  - Variabile casuale (teoria delle probabilità)
  - Stimatore (inferenza statistica)

Media campionaria: variabile casuale

- Ipotesi
  - v.c. X_i risposta del i-simo intervistato.
  - Estrazione Bernoulliana sono X_i i.i.d.

$$E[X_i]=E[P]; Var[X_i]=Var[P] \qquad i=0,1,\dots,n$$

- La media campionaria come v.c.
  - $$E[\bar{X}]=\frac{E[X_1]+....+E[X_n]}{n}=\frac{nE[P]}{n}=E[P]$$
  - $$Var[\bar{X}]=\frac{\sum_{i=1}^{n} Var[X_i]}{n^2}=\frac{nVar[P]}{n^2}=\frac{Var[P]}{n}$$
  - $$\lim_{n\to\infty}\overline{X}=\frac{X_1+X_2+....+X_n}{n}\sim N\left(E[P];\frac{Var[P]}{n}\right)$$

Media campionaria: stimatore.

La media campionaria è uno stimatore del valore atteso

- Lo stimatore è corretto: infatti si ha che

$$E\left[\overline{X}\right]=\theta$$

- Lo stimatore è consistente.
  - Dimostrazione (intuitiva)

  Poiché

$$\lim_{n\to\infty} Var\left[\overline{X}\right]=\lim_{n\to\infty}\frac{Var\left[P\right]}{n}=0$$

  Al crescere di n la media campionaria tende ad essere una costante (ha varianza nulla). Quindi

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}-\theta\right|>\varepsilon\right)=0 \qquad \forall\,\varepsilon>0$$

Varianza campionaria

- Definisco varianza campionaria:

$$s^2=\frac{\sum_i^n\left(o_i-\overline{O}\right)^2}{n-1}=\sigma^2\frac{n}{n-1}=\left(\frac{\sum_i^n o_i^2}{n}-\overline{O}^2\right)\frac{n}{n-1}$$

- S^2 come v.c.
  - v.c. X_i risposta del i-simo intervistato.
  - Estrazione Bernoulliana X_i sono i.i.d.

$$S^2=\frac{\left(\sum_i^n X_i^2\right)-\overline{X}^2}{n-1}=\frac{\left(\sum_i^n P^2\right)-E[P]^2}{n-1}$$

  - Si dimostra che:
    - $E[S^2] = Var[P]$.
    - $P\sim N(\mu,\sigma^2)\Rightarrow S^2\sim\frac{\sigma^2}{n-1}\chi^2(n-1)$

Varianza campionaria: stimatore.

S^2 è uno stimatore della varianza.

- Lo stimatore è corretto: infatti si ha che

$$E[S^2] = Var[P]$$

- Lo stimatore è consistente.
  - Dimostrazione (solo per P normali)

$$P \sim N(\mu, \sigma^2) \Rightarrow S^2 \sim \frac{\sigma^2}{n-1} \chi^2(n-1)$$

$$Var[S^2] = \frac{\sigma^4}{(n-1)^2} Var[\chi^2(n-1)] = \frac{\sigma^4}{(n-1)^2} 2(n-1) = \frac{2\sigma^4}{(n-1)}$$

$$\lim_{n\to\infty} Var\left[S^2\right] = \lim_{n\to\infty} \frac{2\sigma^4}{n-1} = 0$$

La varianza dello stimatore tende a zero al crescere del campione quindi la stima diviene costante.

Esempio - I

- Esempio: Data una v.c. $X \sim N(\mu\ ;\ \sigma^2)$ si sono ottenute le seguenti realizzazioni

  94.07 101.03 102.26 97.98

  Determinare una stima di μ e σ^2.

- Svolgimento:
  - Si stima $E[X] = \mu$:

    $$\bar{x} = \frac{94.07 + 101.03 + 102.26 + 97.98}{4} = 98.83$$

  - Si stima $Var[P] = \sigma^2$:

    $$s^2 = \left(\frac{94.07^2 + 101.03^2 + 102.26^2 + 97.98^2}{4} - 98.83^2\right)\frac{4}{3} = 13.35$$

- Osservazione: dati estratti da $X \sim N(100\ ;\ 25)$.

Esempio - II

- Si vuole stimare la capacità riproduttiva di una tipologia di batteri. Pertanto si sono infettati 16 topi. Dopo 15 gg. si è rilevata la popolazione batterica nelle 16 unità

10 12 11 13 9 10 11 15 12 11 11 15 12 12 9 10

Determinare una stima del valor atteso e della varianza.

- Svolgimento
  - Si ipotizza
    - P:popolazione batterica dopo 15 gg. in un topo sano
    - campionamento sia di tipo bernoulliano
  - Si stima $E[P]$: $\overline{p}=\dfrac{10+12+11+13+...+9+10}{16}=\dfrac{183}{16}$
  - Si stima $Var[P]$: $s^2=\left(\dfrac{100+144+...+100}{16}-\left(\dfrac{183}{16}\right)^2\right)\dfrac{16}{15}$

Stime: considerazioni

- Diverse stime di uno stimatore consistente
  - Caso 1) $n = 10 \rightarrow$ Stima 1
  - Caso 2) $n = 1000 \rightarrow$ Stima 2
  - Quale stima è più affidabile?
- Diverse stime di uno stimatore consistente
  - Caso 1) $n = 100$, $Var[O_1] \rightarrow$ Stima 1
  - Caso 2) $n = 100$, $Var[O_2] > Var[O_1] \rightarrow$ Stima 2
  - Quale stima è più affidabile?
- Osservazione: poiché le stime forniscono un solo valore non è facile discernere.

Stime puntuali e per intervallo

- Per analisi accurate conviene poter essere sicuri della stima fatta.

- Si introducono due tipi di stime

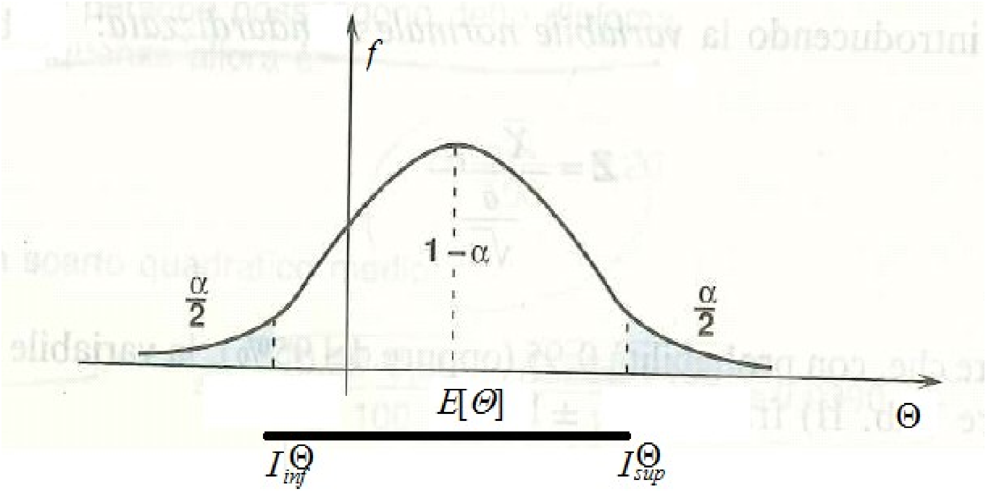
  - Stima puntuale: si stima un solo valore per il parametro ignoto.

  - Stima per intervallo: si stima un intervallo in cui si è fiduciosi ricada il parametro ignoto.

Stime per intervallo: principio base I

- Problema: Come ricavo un intervallo I in cui ci si aspetta ricada il parametro θ che debbo stimare?
- Osservazione: Nota $f(\Theta)$ posso trovare un intervallo I^{Θ} che
  - Abbia una (alta) probabilità $1-\alpha$ di contenere la stima $\hat{\theta}$
  - Bipartisca la probabilità α nelle code.

  Esempio per $f(\Theta)$ gaussiana e stimatore corretto



- Osservazione: la d.d.p. di Θ descrive la probabilità che la mia stima assuma un determinato valore ed è legata a θ.

Stime per intervallo: principio base II

- Metodo:
  - Dati:
    - la probabilità $1\text{-}\alpha$,
    - stimatore $g(.)$ e la sua d.d.p.
  - Cerco

    1) di ottenere un intervallo

    $$I^{\Theta} \ : \ P\left(\Theta \in I^{\Theta}(\theta)\right) = 1 - \alpha$$

    2) esplicito il legame fra il θ e Θ in modo da ottenere

    $$I \ : \ P(\theta \in I) = 1 - \alpha$$

- Definizioni:
  - I: Intervallo di confidenza.
  - $1\text{-}\alpha$: livello di confidenza.

Stime per intervallo:valore atteso - I

- La media campionaria
  - $\bar{x}$ = stima puntuale di $E[P]$.
  - Per n “grande” ho che $\overline{X} \sim N\left(E[P]; \dfrac{Var[P]}{n}\right)$

1) Ricavo l'intervallo con probabilità

$$P(I_{inf}^{\Theta} \leq \overline{X} \leq I_{sup}^{\Theta}) = 1 - \alpha$$

- Standardizzo

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\overline{X} - E[P]}{\sqrt{\left(\frac{Var[P]}{n}\right)}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

Stime per intervallo:valore atteso - II

$$P\left(-z_{\frac{\alpha}{2}} \le \frac{\overline{X}-E[P]}{\sqrt{\frac{Var[P]}{n}}} \le z_{\frac{\alpha}{2}}\right)=1-\alpha$$

2) Ricavo un intervallo per il parametro ($E[P]$)

$$P\left(-z_{\frac{\alpha}{2}}\sqrt{\frac{Var[P]}{n}} \le \overline{X}-E[P] \le z_{\frac{\alpha}{2}}\sqrt{\frac{Var[P]}{n}}\right)=1-\alpha$$

$$P\left(-\overline{X}-z_{\frac{\alpha}{2}}\sqrt{\frac{Var[P]}{n}} \le -E[P] \le -\overline{X}+z_{\frac{\alpha}{2}}\sqrt{\frac{Var[P]}{n}}\right)=1-\alpha$$

$$P\left(\overline{X}-z_{\frac{\alpha}{2}}\sqrt{\frac{Var[P]}{n}} \le E[P] \le \overline{X}+z_{\frac{\alpha}{2}}\sqrt{\frac{Var[P]}{n}}\right)=1-\alpha$$

Ottengo l'intervallo $I=\left[\overline{x}-z_{\frac{\alpha}{2}}\sqrt{\frac{Var[P]}{n}}\ ;\ \overline{x}+z_{\frac{\alpha}{2}}\sqrt{\frac{Var[P]}{n}}\right]$

Stime per intervallo:valore atteso - III

- Stima nel caso di varianza nota

$$I = \left[\overline{x} - z_{\frac{\alpha}{2}} \sqrt{\frac{Var[P]}{n}} \ ; \ \overline{x} + z_{\frac{\alpha}{2}} \sqrt{\frac{Var[P]}{n}} \right]$$

- Problema: $Var[P]$ è spesso ignota.
- Soluzione: la stimo usando s^2.

$$I = \left[\overline{x} - z_{\frac{\alpha}{2}} \sqrt{\frac{s^2}{n}} \ ; \ \overline{x} + z_{\frac{\alpha}{2}} \sqrt{\frac{s^2}{n}} \right]$$

- Stima nel caso di varianza ignota

$$I = \left[\overline{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \ ; \ \overline{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right]$$

Esempio - III

- Esempio: Data una v.c. $X \sim N(\mu\ ;\ \sigma^2)$ si sono ottenute le seguenti realizzazioni

  94.07 101.03 102.26 97.98

  Determinare una stima per intervallo al 95% di μ.

- Svolgimento:
  - Indici campionari $\bar{x} = 98.83$ $s^2 = 13.35 \Rightarrow s = 3.653$
  - Valori standardizzata $z_{0.025} = 1.96$
  - Stima richiesta

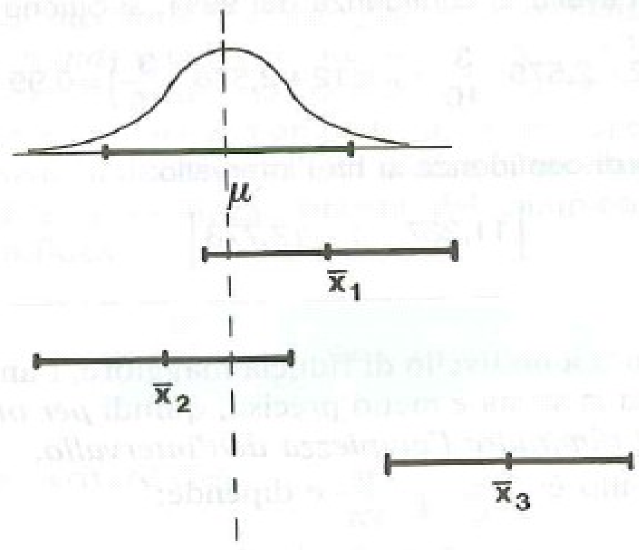
$$I = \left[\bar{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\ ;\ \bar{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right] = [95.25\ ;\ 102.41]$$

- Osservazione: l'approssimazione vale per n molto grande pertanto il risultato non è molto attendibile!

Stime per intervallo:considerazioni

- Cosa vuol dire fare la stima per intervallo ad un livello di confidenza (es. *95%*)? Perché non si usa il termine probabilità?

- Osservazione: il parametro è costante.
- Osservazione: la stima è una v.c.
- Pertanto è
  - Errato: il parametro è contenuto nella stima con una probabilità pari al *95%*.
  - Corretto: estratti tanti campioni ad n elementi, la probabilità che una contenga la stima è del *95%*



Stime per intervallo: varianza - I

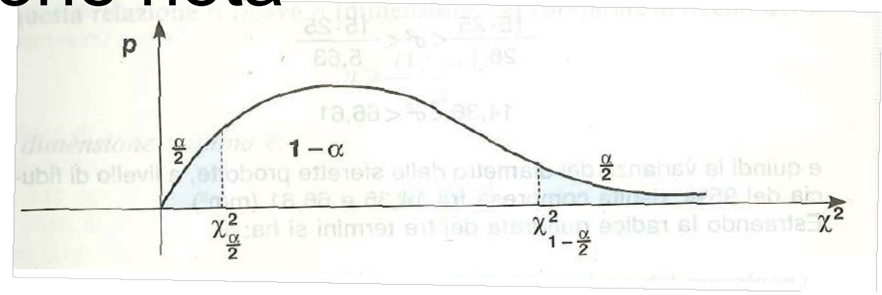
- La varianza campionaria
  - s^2 = stima puntuale di $Var[P]$.
  - Per n "grande" e P gaussiana ho che $S^2 \sim \frac{Var[P]}{n-1}\chi^2(n-1)$

1) Ricavo l'intervallo con probabilità

$$P(I_{inf}^{\Theta} \leq S^2 \leq I_{sup}^{\Theta}) = 1-\alpha$$

- riconduco ad una distribuzione nota

$$S^2 \frac{n-1}{Var[P]} \sim \chi^2(n-1)$$



- da cui ottengo

$$P\left(\chi^2_{\frac{\alpha}{2}}(n-1) \leq \frac{n-1}{Var[P]}S^2 \leq \chi^2_{1-\frac{\alpha}{2}}(n-1)\right) = 1-\alpha$$

Stime per intervallo: varianza - II

$$P\left(\chi^2_{\frac{\alpha}{2}}(n-1) \le \frac{n-1}{Var[P]}S^2 \le \chi^2_{1-\frac{\alpha}{2}}(n-1)\right)=1-\alpha$$

2) Ricavo un intervallo per il parametro ($Var[P]$)

$$P\left(\frac{\chi^2_{\frac{\alpha}{2}}(n-1)}{(n-1)S^2} \le \frac{1}{Var[P]} \le \frac{\chi^2_{1-\frac{\alpha}{2}}(n-1)}{(n-1)S^2}\right)=1-\alpha$$

$$P\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \le Var[P] \le \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right)=1-\alpha$$

- Ottengo la stima

$$I=\left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \; ; \; \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]$$

Esempio - IV

- Esempio: Data una v.c. $X \sim N(\mu\ ;\ \sigma^2)$ si sono ottenute le seguenti realizzazioni

  94.07 101.03 102.26 97.98

  Determinare una stima per intervallo al 95% di σ^2.

- Svolgimento:
  - Indici campionari $\bar{x} = 98.83$ $s^2 = 13.35 \Rightarrow s = 3.653$
  - Valori chi quadrato $\chi^2_{0.025}(3) = 0.216$ $\chi^2_{0.975}(3) = 9.35$
  - Stima $$I = \left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} ; \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \right] = \left[\frac{3 \cdot 13.35}{9.35} ; \frac{3 \cdot 13.35}{0.216} \right] = [4.28 ; 185.4]$$

- Osservazione: l'approssimazione vale
  - per n “grande”
  - per popolazioni gaussiane (evitabile se n è “veramente” grande)

Ricapitolando - I

- Parametro θ: indice di una popolazione (o v.c.) ignoto..
- Stimatore Θ: funzione $g(.)$ di osservazioni campionarie.
- Stima θ: valore assunto da $g(.)$ una volta estratto il campione.
- Proprietà di uno stimatore
  - Correttezza: $E[\Theta]=\hat{\theta}=\theta$
  - Consistenza: $\lim_{n\to\infty} P\left(\left|\hat{\theta}-\theta\right|>\varepsilon\right)=0 \quad \forall \varepsilon>0$
  - Efficienza: $Var[\Theta]$ piccola
- Stime:
  - Puntuali: si stima un solo valore per il parametro ignoto
  - Per intervallo: si stima un intervallo in cui confido possa essere incluso il parametro ignoto.
    - Regolato dal livello di confidenza.

Ricapitolando - II

- Media campionaria
  - Stimatore del valore atteso
  - Stima corretta, consistente e efficiente.
  - Per n “grande” $\overline{X} \sim N\left(E[P]; \frac{Var[P]}{n} \right)$
- Varianza campionaria:

$$s^2 = \frac{\left(\sum_i^n o_i - \overline{O} \right)^2}{n-1} = \left(\frac{\sum_i^n o_i^2}{n-1} - \overline{O}^2 \right) \frac{n}{n-1}$$

  - Stimatore della varianza
  - Stima corretta, consistente
  - Per n “grande” e P gaussiano

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

Ricapitolando - III

- Stima del valore atteso di una popolazione
  - puntuale $E[P] = \overline{x}$
  - intervallo

$$E[P] \in \left[\overline{x} - z_{\frac{\alpha}{2}} \sqrt{\frac{Var[P]}{n}} \;\; ; \;\; \overline{x} + z_{\frac{\alpha}{2}} \sqrt{\frac{Var[P]}{n}} \right]$$

- Stima della varianza di una popolazione
  - puntuale $Var[P] = s^2$
  - intervallo

$$Var[P] \in \left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} ; \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \right]$$