

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

1

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 informazioni posso ottenere?

2

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 unità statistiche vengono estratte una alla volta e dopo l'estrazione sono nuovamente estraibili.
 - **Estrazione in blocco:** le n unità statistiche vengono estratte in blocco (non è possibile per una singola unità comparire più volte).
- Tratteremo solo casi di estrazioni bernoulliane.

3

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 descrive la popolazione come una v.c. P avente d.d.p. (funzione di probabilità) incognita.

4

Modellazione

- Popolazione: v.c. P con d.d.p. $f(p)$.
- Osservazione della i -sima unità statistica: v.c. X_i
 - $X_i \sim P$ (la prima estrazione la faccio da P).
 - Nessuna garanzia che la d.d.p. delle estrazioni successive:
 - resti costante ($f(x_i) = f(x_j)$).
 - 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 $E[X_i] = E[P], \text{Var}[X_i] = \text{Var}[P]$.

5

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: Teoria della stima.

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: Test non parametrici.

6

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 parametro θ della popolazione.
 - Mezzo:
 - Una funzione $g(\cdot)$ dei dati chiamata stimatore
 - Risultato
 - Una stima del parametro $\hat{\theta} = g(O)$

7

Problema della stima

Problema: dato un campione O di dimensione n estratto da una popolazione P , avente un parametro incognito θ , determinare una funzione $g(\cdot)$ 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 $g_1(\cdot)$
 - Stima $\hat{\theta}_1 = g_1(O)$
- Esempio: Stimare la varianza della popolazione.
 - Parametro $\theta_2 = \text{Var}[P]$.
 - Stimatore $g_2(\cdot)$
 - Stima $\hat{\theta}_2 = g_2(O)$

8

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. $g(.) = \frac{X_1 + X_2 + X_3 + \dots + X_n}{n}$
(Media campionaria)
- 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

9

Stimatore: proprietà - I

Quali caratteristiche vorrei avesse uno stimatore?

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

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

- Esempio:

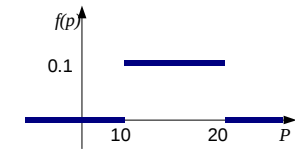
- Popolazione P uniforme
- $Var[P] = 100/12 = 25/3$
- Stimatore di $Var[P]$ corretto

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

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

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

10



Stimatore: proprietà - II

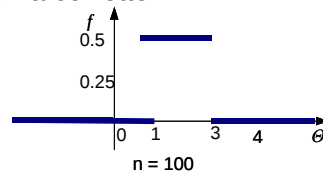
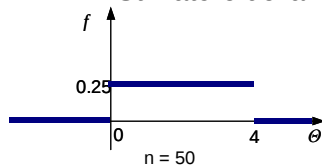
Quali caratteristiche vorrei avesse uno stimatore?

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

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0 \quad \forall \varepsilon > 0$$

- Esempio

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



11

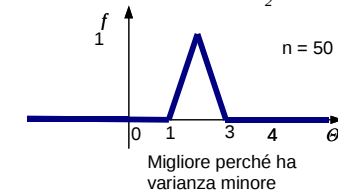
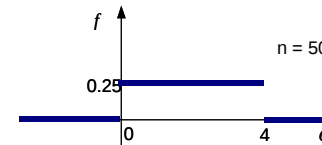
Stimatore: proprietà - III

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

12

Media campionaria

- Si indica sopra segnando la grandezza mediata

- **Definizione:**

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \sum_{i=1}^n \frac{X_i}{n}$$

- **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)

13

Media campionaria: variabile casuale

- Ipotesi

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

$$E[X_i] = E[P]; \text{Var}[X_i] = \text{Var}[P] \quad i=0,1,\dots,n$$

- La media campionaria come v.c.

$$E[\bar{X}] = \frac{E[X_1] + \dots + E[X_n]}{n} = \frac{nE[P]}{n} = E[P]$$

$$\text{Var}[\bar{X}] = \frac{\sum_{i=1}^n \text{Var}[X_i]}{n^2} = \frac{n \text{Var}[P]}{n^2} = \frac{\text{Var}[P]}{n}$$

$$\lim_{n \rightarrow \infty} \bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \sim N\left(E[P]; \frac{\text{Var}[P]}{n}\right)$$

14

Media campionaria: stimatore.

La media campionaria è uno stimatore del valore atteso

- Lo stimatore è **corretto**: infatti si ha che

$$E[\bar{X}] = \theta$$

- Lo stimatore è **consistente**.

- Dimostrazione (intuitiva)

$$\text{Poiché} \quad \lim_{n \rightarrow \infty} \text{Var}[\bar{X}] = \lim_{n \rightarrow \infty} \frac{\text{Var}[P]}{n} = 0$$

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

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0 \quad \forall \varepsilon > 0$$

15

Varianza campionaria

- Definisco varianza campionaria:

$$s^2 = \frac{\sum_{i=1}^n (o_i - \bar{O})^2}{n-1} = \sigma^2 \frac{n}{n-1} = \left(\frac{\sum_{i=1}^n o_i^2}{n} - \bar{O}^2 \right) \frac{n}{n-1}$$

- S^2 come v.c.

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

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

- Si dimostra che:

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

16

Varianza campionaria: stimatore.

S^2 è uno stimatore della varianza.

- Lo stimatore è **corretto**: infatti si ha che

$$E[S^2] = \text{Var}[P]$$

- Lo stimatore è **consistente**.

– Dimostrazione (solo per P normali)

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

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

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

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

17

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 $\text{Var}[P] = \sigma^2$:

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

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

18

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 $E[P]$: $\bar{p} = \frac{10+12+11+13+\dots+9+10}{16} = \frac{183}{16}$

– Si stima $\text{Var}[P]$: $s^2 = \left(\frac{100+144+\dots+100}{16} - \left(\frac{183}{16} \right)^2 \right) \frac{16}{15}$ 19

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) $n = 100, \text{Var}[O_1] \rightarrow$ Stima 1
- Caso 2) $n = 100, \text{Var}[O_2] > \text{Var}[O_1] \rightarrow$ Stima 2
- Quale stima è più affidabile?

- Osservazione:** poiché le stime forniscono un solo valore non è facile discernere.

20

Stime puntuali e per intervallo

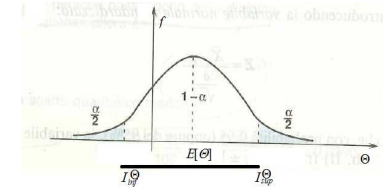
- Per analisi accurate conviene poter essere sicuri della stima fatta.
- Si introducono due tipi di stime
 - **Stima puntuale**: si stima un solo valore per il parametro ignoto.
 - **Stima per intervallo**: si stima un intervallo in cui si è fiduciosi ricada il parametro ignoto.

21

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 .

22

Stime per intervallo: principio base II

- **Metodo**:
 - Dati:
 - la probabilità $1-\alpha$,
 - stimatore $g(\cdot)$ e la sua d.d.p.
 - Cerco
 - 1) di ottenere un intervallo

$$I^\theta : P(\theta \in I^\theta(\theta)) = 1-\alpha$$
 - 2) esplicito il legame fra il θ e θ in modo da ottenere

$$I^\theta(\alpha, g(\cdot))$$
- **Definizioni**:
 - I : Intervallo di confidenza.
 - $1-\alpha$: livello di confidenza.

23

Stime per intervallo: valore atteso - I

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

1) Ricavo l'intervallo con probabilità

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

– Standardizzo

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

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

24

Stime per intervallo: valore atteso - II

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

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

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

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

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

Otengo l'intervallo
$$I = \left[\bar{X} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\text{Var}[P]}{n}} ; \bar{X} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\text{Var}[P]}{n}} \right]$$

Stime per intervallo: valore atteso - III

- Stima nel caso di varianza nota

$$I = \left[\bar{X} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\text{Var}[P]}{n}} ; \bar{X} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\text{Var}[P]}{n}} \right]$$

- Problema:** $\text{Var}[P]$ è spesso ignota.
- Soluzione:** la stimo usando s^2 .

$$I = \left[\bar{X} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{s^2}{n}} ; \bar{X} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{s^2}{n}} \right]$$

- Stima nel caso di varianza ignota

$$I = \left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} ; \bar{X} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right]$$

26

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 standardizzati $z_{0.025} = 1.96$
- 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!

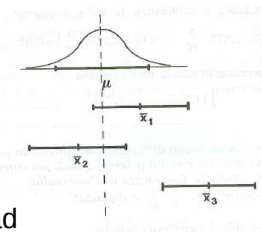
27

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%



28

Stime per intervallo: varianza - I

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

1) Ricavo l'intervallo con probabilità

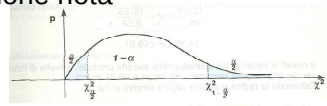
$$P(I_{inf}^{\theta} \leq S^2 \leq I_{sup}^{\theta}) = 1 - \alpha$$

- ricondo ad una distribuzione nota

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

- da cui ottengo

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



29

Stime per intervallo: varianza - II

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

2) Ricavo un intervallo per il parametro ($\text{Var}[P]$)

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

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

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

30

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 $\bar{x} = 98.83 \quad s^2 = 13.35 \Rightarrow s = 3.653$
- Valori chi quadrato $\chi^2_{0.025}(3) = 0.216 \quad \chi^2_{0.975}(3) = 9.35$
- 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)

31

Ricapitolando - I

- Parametro θ : indice di una popolazione (o v.c.) ignoto..
- Stimatore $\hat{\theta}$: funzione $g(\cdot)$ di osservazioni campionarie.
- Stima θ : valore assunto da $g(\cdot)$ una volta estratto il campione.
- Proprietà di uno stimatore
 - Correttezza: $E[\hat{\theta}] = \theta$
 - Consistenza: $\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0 \quad \forall \varepsilon > 0$
 - Efficienza: $\text{Var}[\hat{\theta}]$ 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.

32

Ricapitolando - II

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

- Varianza campionaria:

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

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

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

33

Ricapitolando - III

- Stima del valore atteso di una popolazione

- puntuale $E[P] = \bar{x}$

- intervallo

$$E[P] \in \left[\bar{x} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{Var[P]}{n}} ; \bar{x} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{Var[P]}{n}} \right]$$

- Stima della varianza di una popolazione

- puntuale $Var[P] = s^2$

- intervallo

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

34