Corso di Modelli per l'analisi statistica

Intervalli di Confidenza

Dalla stima puntuale alla stima per intervallo

Anche utilizzando uno stimatore con proprietà ottimali, per effetto del caso la stima puntuale sulla base di un campione può essere molto diversa dal valore vero del parametro

Di solito si preferisce stimare un intervallo di valori per il parametro incognito

Una stima compresa tra due limiti
- riflette meglio l'incertezza legata all'inferenza
- incorpora direttamente l'informazione sul grado di precisione

La stima per intervallo

È un intervallo di valori plausibili a cui associamo un dato livello di confidenza o affidabilità o fiducia (generalmente fissato al 90%, 95% o 99%)

Ci aspettiamo che l'intervallo contenga, con quel livello di fiducia, il valore incognito

Stima per intervallo

Per costruire la stima intervallare del parametro θ al livello di confidenza $1-\alpha$, sulla base delle osservazioni campionarie si stimano due valori, L_1 e L_2 (gli estremi dell'intervallo) in maniera tale che

$$P(L_1 \leq \theta \leq L_2) = 1 - \alpha$$

L_1 e L_2 sono statistiche campionarie, cioè variano al variare dei campioni

$L_1 = L_1(X_1, X_2, \ldots, X_n)$ $\quad$ $L_2 = L_2(X_1, X_2, \ldots, X_n)$

Interpretazione della stima per intervallo

L'affermazione "θ è compreso tra L_1 e L_2 con probabilità pari a $1-\alpha$" va interpretata nello spazio campionario, prima di estrarre il campione effettivo

In questo senso, $1-\alpha$ è la frequenza relativa di campioni per i quali l'intervallo include il valore incognito θ
Si accetta un rischio pari ad α che il campione estratto produca un intervallo che non contenga θ

Interpretazione del livello di confidenza

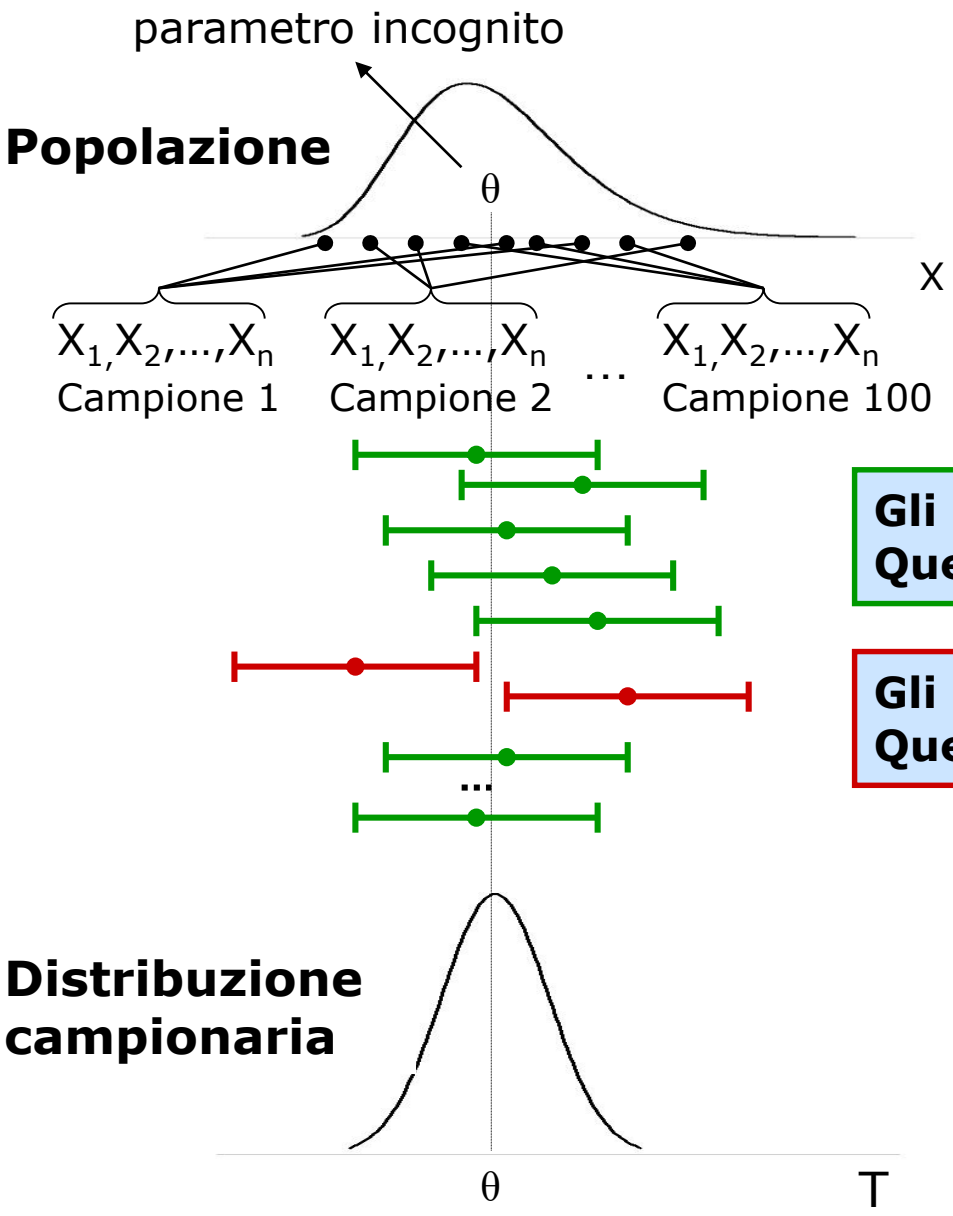
Fissiamo $1-\alpha=0,95$

Ipotizziamo di estrarre successivamente più campioni indipendenti dalla stessa popolazione e costruiamo le corrispondenti stime intervallari

⇨ Per 95 campioni su 100, θ è compreso nell'intervallo stimato

Il campione estratto però potrebbe anche essere uno di quella frazione α (il 5%) per la quale l'intervallo non cattura il valore incognito θ

Intervallo di confidenza per θ



Livello di confidenza $1-\alpha=0,95$

Gli intervalli verdi contengono θ
Questo si verifica per 95 campioni su 100

Gli intervalli rossi NON contengono θ
Questo si verifica per 5 campioni su 100

Non possiamo sapere se il campione che abbiamo effettivamente estratto è uno di quelli per i quali gli intervalli stimati contengono θ oppure no.

Caso di studio

Per programmare meglio il servizio offerto, un'azienda leader nella vendita on line di libri, Cd e DVD vuole conoscere:

- L'importo medio di ogni ordine
- La proporzione di pagamenti fatti con la carta di credito VISA

Estrae un campione casuale di n transazioni delle quali osserva l'importo e il metodo di pagamento

Stima puntuale della media

Ad esempio, la media campionaria dell'importo è pari a

$$\bar{x} = 57{,}24€$$

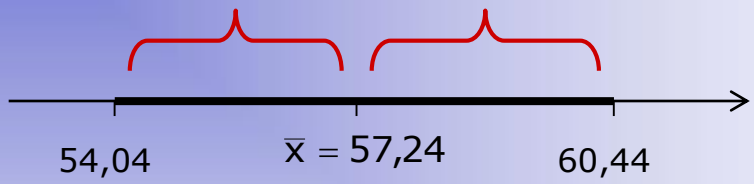
⇨ In media per ogni ordine si spendono 57,24€

La stima puntuale non dà indicazioni sulla accuratezza del risultato
La media campionaria varia da campione a campione a seconda delle unità selezionate

Per effetto del caso, posso essere stato "fortunato" e avere estratto un campione che fornisce un valore medio molto vicino a quello incognito della popolazione
Ma posso anche essere stato particolarmente "sfortunato" e avere estratto un campione di osservazioni che produce una media molto distante da quella incognita

Stima per intervallo della media

Ad esempio, l'intervallo stimato al livello del 95% comprende valori da 54,04 a 60,44



L'intervallo è del tipo $\bar{x} \pm$ margine di errore

⇨ Importi medi compresi tra 54,04€ e 60,44€ li giudichiamo plausibili

Siamo confidenti al 95% che il vero valore dell'importo medio di tutte le transazioni sia compreso tra 54,04 € e 60,44 €

Stima puntuale della proporzione

Ad esempio, la proporzione campionaria di pagamenti fatti con la VISA (stima puntuale) è pari a

$$p = \bar{x} = 0{,}52 \quad (52\%)$$

⇨ Oltre la metà dei pagamenti sono regolati con la carta di credito VISA

Stima per intervallo della proporzione

La stima per intervallo fornisce un insieme di valori plausibili della proporzione incognita cui è associato un dato livello di confidenza

Ad esempio, la stima per intervallo al livello del 95% comprende valori tra 0,422 e 0,618

⇒ Con il prefissato livello di confidenza, per una proporzione di pagamenti compresa tra il 42,2% e il 61,8% i clienti utilizzano la carta VISA

Stima per intervallo della media μ

Tre situazioni

1. Popolazione Normale $X \sim N(\mu;\sigma^2)$
   σ^2 nota

2. Popolazione Normale $X \sim N(\mu;\sigma^2)$
   σ^2 non nota

3. Popolazione non Normale, n grande (n>30)

Pop. Normale, varianza nota

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

Ad un livello di confidenza $1-\alpha=0,95$
Qual è un insieme di valori "plausibili" per Z?

$$P(-z \le Z \le z) = 0,95$$

dalla tavola Z $\Rightarrow$ $P(-1,96 \le Z \le 1,96) = 0,95$

Qual è un insieme di valori "plausibili" per $\bar{X}$?

$$P\left(-1,96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1,96\right) = 0,95$$

L'obiettivo è scrivere un intervallo casuale per μ, i cui estremi dipendono dal campione

Pop. Normale, varianza nota

$$P\left(-\overline{X}-1{,}96\frac{\sigma}{\sqrt{n}} \le -\mu \le -\overline{X}+1{,}96\frac{\sigma}{\sqrt{n}}\right)=0{,}95$$

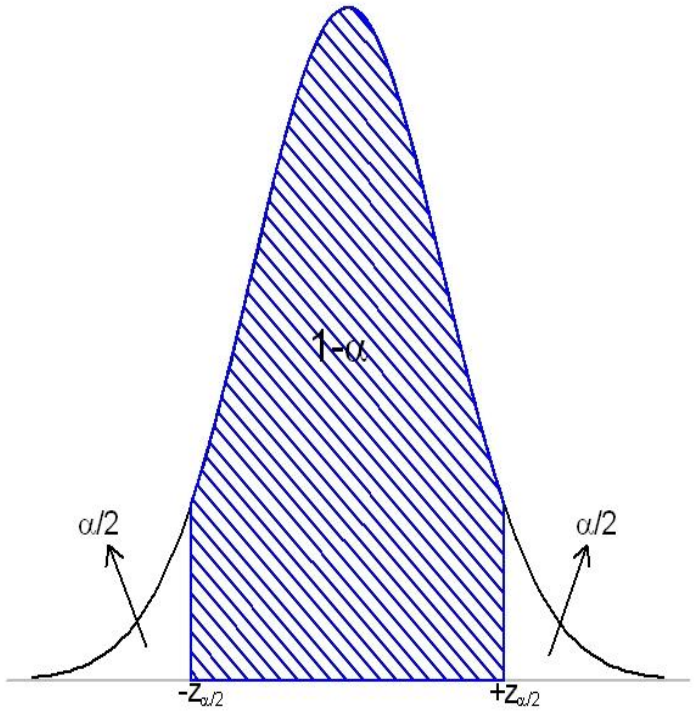
$$P\left(+\overline{X}+1{,}96\frac{\sigma}{\sqrt{n}} \ge +\mu \ge +\overline{X}-1{,}96\frac{\sigma}{\sqrt{n}}\right)=0{,}95$$

$$P\left(\overline{X}-1{,}96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}+1{,}96\frac{\sigma}{\sqrt{n}}\right)=0{,}95$$

$\Rightarrow$ $\bar{x} \pm 1{,}96\frac{\sigma}{\sqrt{n}}$ è la stima per intervallo di μ

$1{,}96\frac{\sigma}{\sqrt{n}}$ → margine di errore

Pop. Normale, varianza nota



Per un generico livello di confidenza 1-α

$$P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1-\alpha$$

da cui

$$P\left(-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$\overline{x} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ è la stima per intervallo di μ

Pop. Normale, varianza nota

Intervallo di confidenza per μ

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

Gli estremi dell'intervallo dipendono da:

1. la media campionaria $\bar{x}$
2. la deviazione standard σ della popolazione
3. il valore $z_{\alpha/2}$
4. la dimensione campionaria n

Lunghezza dell'intervallo e errore della stima intervallare

Lunghezza (ampiezza) dell'intervallo

$$2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

non varia al variare dei campioni

Margine di errore dell'intervallo = semi-lunghezza

$$\text{errore} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Il margine di errore è collegato al concetto di precisione della stima
Minore è l'errore maggiore è la precisione e quindi l'accuratezza della stima per intervallo

Margine di errore

Indica di quanto, al massimo, la stima campionaria si discosta, verosimilmente, dal parametro incognito
Si sottolinea verosimilmente perché esiste una frazione α% di campioni per i quali la stima si discosta dal parametro di una quantità maggiore del margine di errore

Con Excel
L'errore associato alla stima per intervallo della media μ della popolazione con varianza nota si ottiene con la funzione CONFIDENZA(α;σ;n)

Errore e livello di confidenza

Conoscendo σ, per una dimensione campionaria fissata n, l'errore varia direttamente al variare di $z_{\alpha/2}$ che, a sua volta, dipende direttamente dal livello di confidenza $1-\alpha$

La riduzione dell'errore si può realizzare al costo di accettare un livello di confidenza minore

Errore e dimensione campionaria

Conoscendo σ, per un dato livello di confidenza 1-α, l'errore varia inversamente al variare di n

La riduzione dell'errore si può realizzare al costo di aumentare la dimensione del campione

Numerosità campionaria per ottenere una data precisione

Volendo determinare in anticipo la dimensione campionaria che assicura una determinata precisione (un errore massimo ammissibile) si usa la relazione tra n e l'errore
Fissato un errore massimo (o una precisione minima) δ

$$\delta = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad n = \left(z_{\alpha/2}\frac{\sigma}{\delta}\right)^2$$

A parità di livello di confidenza, se voglio dimezzare l'errore devo quadruplicare n

Determinazione della numerosità campionaria

L'intervallo di confidenza stimato per l'importo medio delle transazioni on line è [54,04 ; 60,44]
Ipotizzando n=10; σ^2=26,6; 1-α=0,95
L'errore (semi-larghezza) è 3,2€

Volendo sfruttare queste informazioni per programmare una nuova indagine campionaria, se l'obiettivo è quello di tollerare un errore massimo pari a 2€, qual è la numerosità che devo estrarre?

$$n = \left(z_{\alpha/2} \frac{\sigma}{\delta} \right)^2 = \left(1{,}96 \cdot \frac{5{,}16}{2} \right)^2 = 25{,}57 \quad \Rightarrow \quad n=26$$

Pop. Normale, varianza non nota

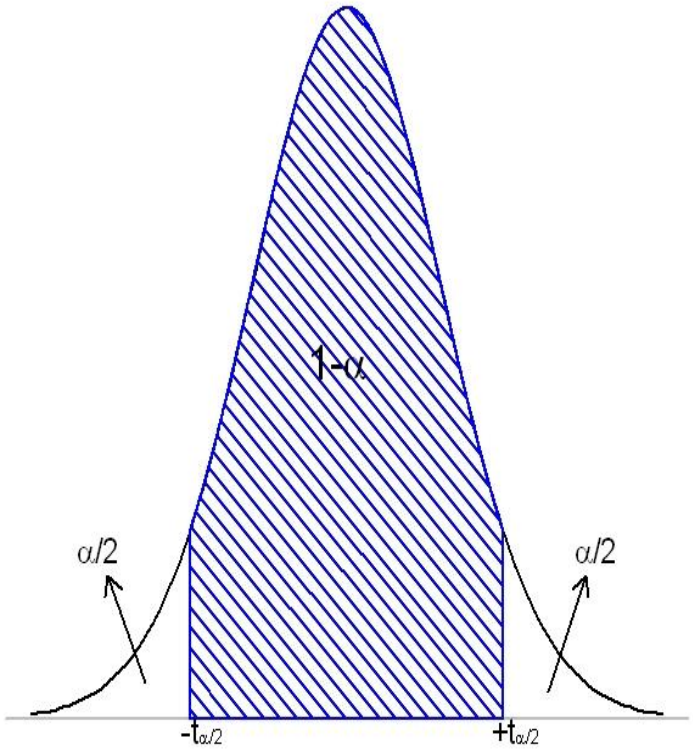
La varianza incognita σ^2 si stima con la varianza campionaria corretta

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

Standardizzando a partire dalla distribuzione di $\overline{X}$ non si ottiene la v.c. $Z \sim N(0,1)$ ma la v.c. t di Student con n-1 gradi di libertà

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

Pop. Normale, varianza non nota



Per un generico livello di confidenza 1-α

$$P\left(-t_{\alpha/2} \le T \le t_{\alpha/2}\right) = 1-\alpha$$

da cui

$$P\left(-t_{\alpha/2} \le \frac{\overline{X}-\mu}{S/\sqrt{n}} \le t_{\alpha/2}\right) = 1-\alpha$$

$$P\left(\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

$\overline{x} \pm t_{\alpha/2}\dfrac{s}{\sqrt{n}}$ è la stima per intervallo di μ

Pop. Normale, varianza non nota

Intervallo di confidenza per μ

$$\left[\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} \;,\; \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

$$\text{errore} = t_{\alpha/2}\frac{s}{\sqrt{n}} \qquad n = \left(t_{\alpha/2}\frac{s}{\delta}\right)^2$$

per un errore massimo fissato pari a δ

Per n grande (n>120) $T \longrightarrow Z$
l'intervallo è approssimato da

$$\left[\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \;,\; \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

Lunghezza dell'intervallo e errore della stima intervallare

Lunghezza (ampiezza) dell'intervallo

$$2t_{\alpha/2}\frac{s}{\sqrt{n}}$$

in questo caso varia al variare dei campioni

Margine di errore dell'intervallo = semi-lunghezza

$$\text{errore} = t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Fissato un errore massimo δ $\Rightarrow$ $n = \left(t_{\alpha/2}\frac{s}{\delta}\right)^2$

s è una stima della varianza della pop. nota prima di estrarre il campione

Pop. non Normale

n grande $\Rightarrow^{TLC}$ $\bar{X}$ tende a $\sim N\left(\mu; \frac{\sigma^2}{n}\right)$

$\Rightarrow$ $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ tende a $\sim N(0;1)$

Intervallo di confidenza per μ

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

con varianza della pop. nota

$$\left[\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \;,\; \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

con varianza della pop. non nota

Stima per intervallo della proporzione π

Popolazione Bernoulliana
Il carattere che si studia assume due sole modalità: Presenza/Assenza di un attributo A
X=1 con prob. π
X=0 con prob. 1-π
In una popolazione finita π è la proporzione di unità che presentano l'attributo A

Come stima puntuale di π si usa la proporzione campionaria p

$$P = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

p è di fatto una media campionaria

Pop. Bernoulliana

n grande $\xrightarrow{\text{TLC}}$ $\overline{X}$ tende a $\sim N\left(\pi; \frac{\pi(1-\pi)}{n}\right)$

$\Rightarrow$ $Z = \frac{\overline{X} - \pi}{\sqrt{\pi(1-\pi)/n}}$ tende a $\sim N(0;1)$

Per un generico livello di confidenza 1-α

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

da cui

$$P\left(\overline{X} - z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \leq \pi \leq \overline{X} + z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right) \cong 1 - \alpha$$

Pop. Bernoulliana

La varianza della proporzione campionaria dipende dalla proporzione incognita π

π si stima con $\bar{x}$ ottenendo l'intervallo

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

$$\text{errore} = z_{\alpha/2} \frac{\sqrt{\bar{x}(1-\bar{x})}}{\sqrt{n}}$$

Fissato un errore massimo $\delta \Rightarrow n = z^2_{\alpha/2} \frac{\hat{\pi}(1-\hat{\pi})}{\delta^2}$

$\hat{\pi}$ è una stima preliminare di π nota prima di estrarre il campione

oppure $\hat{\pi} = 0{,}5$

Sondaggi elettorali. Ballottaggio tra due candidati

Problema statistico: stimare una proporzione incognita π, ossia la proporzione di elettori che intende votare per il candidato XXY alle prossime elezioni

Si estrae un campione casuale di n elettori (n grande)
Supponiamo che la proporzione campionaria a favore di XXY sia pari al 53,2% (stima puntuale)

Il candidato XXY può ritenersi sicuro di vincere?

Sondaggi elettorali. Ballottaggio tra due candidati

Meglio affidarsi ad una stima per intervallo!

Al livello di confidenza del 95% il margine di errore della stima intervallare è stimato da:

$$\text{errore} = 1{,}96\,\frac{\sqrt{0{,}5(1-0{,}5)}}{\sqrt{n}} = \frac{1{,}96}{2\sqrt{n}} = \frac{0{,}98}{\sqrt{n}}$$

utilizzando 0,5 (valore prudenziale) come stima di π nell'espressione della varianza campionaria

Sondaggi elettorali

Intervistando 500 elettori, l'errore è pari al 4,4%
L'intervallo della proporzione incognita è $53{,}2 \pm 4{,}4$
cioè [48,8 ; 57,6]

Se n=2000, l'errore=2,2%
L'intervallo è più accurato
$53{,}2 \pm 2{,}2$ [51,0 ; 55,4]
La forchetta è più stretta

Nel primo caso, verosimilmente, il candidato potrebbe avere una percentuale di preferenze al di sotto del 50%

n	errore
500	**0,044 (4,4%)**
1000	**0,031 (3,1%)**
2000	**0,022 (2,2%)**

$\overline{x} = 53{,}2\%$