Statistica inferenziale

- **Problema diretto o deduttivo:** Da un problema concreto si costruisce un modello teorico o popolazione. A partire da tale modello la teoria della probabilità consentedi prevedere il comportamento potenziale dei dati, se si conosce il parametro θ.
- **Problema inverso o induttivo:** Se non si conosce il parametro θ e si dispone di un campione $(x_1, x_2, \ldots, x_n)$, dai dati osservati si tenta di risalire al parametro incognito.

Inferenza statistica

- **L'inferenza statistica** è l'insieme dei metodi con cui si risolve il problema inverso.
- La **Stima dei parametri** è il procedimento con cui dal campione osservato si traggono informazioni per assegnare al parametro θ un valore (**stima puntuale**) o un insieme di valori (**stima per intervallo**).

Campione causale

- Data una variabile aleatoria X che si vuole analizzare rispetto ad una data popolazione il campione casuale è costituita da una n-pla di variabili aleatorie indipendenti (X_1, X_2,…, X_n) identicamente distribuite.

Distribuzione campionaria

- Dato un campione si chiama statistica una funzione del campione casuale. La distribuzione di probabilità della statistica è chiamata **distribuzione campionaria**. La distribuzione campionaria è determinate da quella della popolazione di riferimento e dall' ampiezza del campione n.
- Per approfondire il tipo di relazione partiamo da una popolazione con media μ e deviazione standard .
- Allora dato un campione estratto da essa (X_1, …, X_n) si ha

Media e varianza

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{1}{n} E(X_1 + X_2 + \ldots + X_n) = \frac{1}{n} n\mu = \mu$$

$$Var(\bar{X}) = Var\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{1}{n} Var(X_1 + X_2 + \ldots + X_n) = \frac{1}{n^2} n\sigma^2 = \sigma^2 / n$$

$$Dev.S\tan d.(\bar{X}) = \sqrt{Var(\bar{X})} = \sigma / \sqrt{n}$$

Teorema del Limite Centrale

Un risultato sosprendente noto come **Teorema del Limite Centrale** stabilisce che se l'ampiezza del campione è sufficientemente grande ($n \geq 50$) allora qualunque sia la distribuzione della popolazione distribuzione della media campionaria può essere ben approssimata dalla distribuzione di Gauss. In particolare se consideriamo la media campionaria standard essa può essere ben approssimata dalla variabile normale standardizzata.

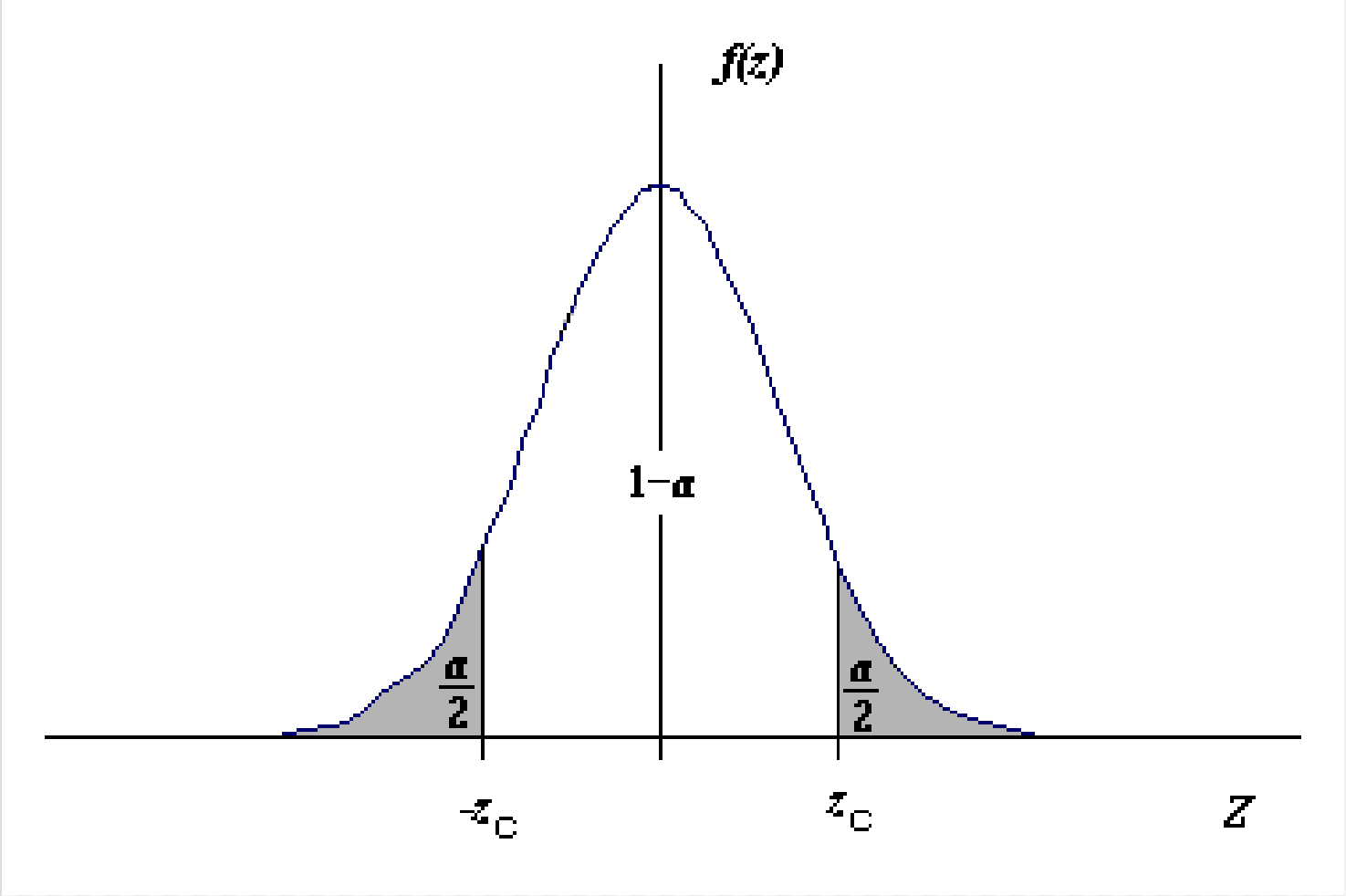
$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

Intervallo di fiducia per una media

- Varianza nota. Sia $X_1, \ldots, X_n$ un campione estratto da una popolazione con media incognita μ e varianza nota . Vogliamo trovare dei valori $(-z_c, z_c)$ tali che Z sia compresa nell' intervallo $(-z_c, z_c)$ con una alta probabilità $1-\alpha$ (grado di fiducia,) cioè

$$P(-z_c \leq Z \leq z_c)=1-\alpha$$

- Dalle tavole di Gauss si vede che tale valore è esattamente $z_c=1,96$ per $1-\alpha = 0,95$


f(z)
1−α
α/2
α/2
−z_C
z_C
Z

Intervallo di fiducia

$$P\left(-z_c \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_c\right) = 1-\alpha$$

$$P\left(-z_c \cdot \sigma/\sqrt{n} \leq \bar{X}-\mu \leq z_c \cdot \sigma/\sqrt{n}\right) = 1-\alpha$$

$$P\left(-\bar{X}-z_c \cdot \sigma/\sqrt{n} \leq -\mu \leq -\bar{X}+z_c \cdot \sigma/\sqrt{n}\right) = 1-\alpha$$

Formula intervallo di fiducia

$$P\left(\overline{X} - z_c \cdot \sigma / \sqrt{n} \leq \mu \leq \overline{X} + z_c \cdot \sigma / \sqrt{n} \right) = 1 - \alpha$$

$$\left(\overline{X} - z_c \cdot \sigma / \sqrt{n}, \overline{X} + z_c \cdot \sigma / \sqrt{n} \right)$$

Errore o livello di fiducia

$$E = z_c \cdot \frac{\sigma}{\sqrt{n}}$$

Esempio

- Un urbanista é interessato alla superficie media μ delle abitazioni della propria città. Uno studio precedente indica che la deviazione standard della popolazione sia circa 8 m^2. In un campione di 50 appartamenti si osserva la media del campione è pari a 120 m^2 . Sulla base di questi dati l' intervallo di confidenza per μ con un grado di fiducia del 95% è

$$\left(120-1{,}96\cdot 8/\sqrt{50}\,,120+1{,}96\cdot 8/\sqrt{50}\right)\approx(117{,}78,122{,}22)$$

Stima dell' ampiezza del campione

- Problema: trovare quanto deve essere grande il campione in modo da avere un errore non più grande di 1.

$$|E| \leq 1 \Rightarrow z_c \cdot \frac{\sigma}{\sqrt{n}} \leq 1 \Rightarrow$$

$$1{,}96 \cdot \frac{8}{\sqrt{n}} \leq 1 \Rightarrow$$

$$\sqrt{n} \geq 8 \cdot 1{,}96 \Rightarrow$$

$$n \geq 15{,}68^2 \Rightarrow$$

$$n \geq [245{,}86]$$

Intervallo di fiducia con varianza non nota

Sia $(X_1, \ldots, X_n)$ un campione estratto da una popolazione con media incognita μ e varianza non nota σ^2
Allora in questo caso il parametro

$$\sigma$$

potrebbe essere approssimato da s'

$$s' = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Si ottiene in questo modo al posto di Z una nuova variabile

$$T = \frac{\bar{X} - \mu}{s'/\sqrt{n}}$$

Distribuzione di Student

Student

$$\frac{s'}{\sqrt{n}} = \frac{s}{\sqrt{n-1}} \quad \Longrightarrow$$

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n-1}}$$

Distribuzione di Student

- Tale distribuzione si chiama di **Student.** E' simile all distribuzione di Gauss e per n molto grande potrebbe essere ben approssimata da essa. Fissando un grado di fiducia a sicercano dei valori critici $(-t_c,t_c)$ dale tavole di Student tali che

$$P(-t_c \leq T \leq t_c)=a$$

Varianza non nota

- A differenza delle tavole di Gauss, quelle di Student dipendono dai **gradi di libertà** del campione cioè n-1. Ripetendo tutti I passaggi precedenti si ottiene come intervallo di fiducia per la media con varianza non nota

$$\left(\bar{X} - t \cdot {s'}/{\sqrt{n}} , \bar{X} + t \cdot {s'}/{\sqrt{n}} \right)$$

Varianza non nota

- O equivalentemente

$$\left(\bar{X} - t \cdot {s}/{\sqrt{n-1}} , \bar{X} + t \cdot {s}/{\sqrt{n-1}} \right)$$

Esempio

- Il produttore di una certa marca di sigarette desidera controllare il quantitativo di catrame in esse contenuto. A questo scopo si osserva un campione di 30 sigarette in cui la media è 10.92 mg e la deviazione standard 0.50 mg . Sulla base di questi dati l' intervallo di fiducia per la media pari al 99%

$$\left(10{,}92 - 2{,}756 \cdot 0{,}51/\sqrt{30}\,,10{,}92 + 2{,}756 \cdot 0{,}51/\sqrt{30}\right) \approx (10{,}66;11{,}18)$$

Intervallo di fiducia con Analisi dei dati di Excel

- L’intervallo di fiducia con Excel si può trovare solo nel caso in cui non si conosce niente della popolazione e vengono utilizzate per questo le tavole di Student.
- Scegliere Statistica descrittiva da Analisi dei dati.
- Scegliere come opzione livello di fiducia per la media.
- Se non si indica viene calcolato come default il livello 0,95.

Analisi dei dati

Come output apparirà il livello di confidenza cioè

$$E = t_c \cdot \frac{s'}{\sqrt{n}}$$

L' intevallo di fiducia è allora

$$(\overline{X} - E, \overline{X} + E)$$