# L’inferenza statistica

- L’inferenza statistica utilizza le informazioni raccolte da un campione (statistiche) al fine di trarre delle conclusioni sulla popolazione (parametri).
- Concetto di distribuzione campionaria.

# La media della popolazione

Esempio: il contenuto delle lattine di birra prodotte deve essere 33cl.

- Il processo produttivo funziona correttamente?
- Il contenuto medio di tutte le lattine prodotte (popolazione) è 33cl ($\mu$=33)?

Analisi su un campione

(una parte delle lattine prodotte).

- Si calcola il contenuto medio di un campione $(\bar{x})$
- Possiamo concludere che il contenuto medio di tutte le lattine prodotte è 33cl ($\mu$=33)?

# La distribuzione campionaria

Quando la media della popolazione (parametro) non è nota, è necessario stimarla attraverso la media campionaria (statistica).

In generale, per stimare un parametro della popolazione, si seleziona di fatto un solo campione casuale di una data ampiezza…anche se, ipoteticamente, sarebbe possibile utilizzare tutti i possibili campioni di quella ampiezza generando una DISTRIBUZIONE (di probabilità) CAMPIONARIA.

Nel caso della media si ha, pertanto, **LA DISTIBUZIONE DELLA MEDIA CAMPIONARIA**!...vedi file excel!

## La distribuzione della media campionaria

- La media campionaria è l'insieme delle medie di tutti i possibili campioni di ampiezza n estraibili dalla popolazione.
- Per distribuzione della media campionaria, si intende la distribuzione di probabilità della media campionaria.
- La media campionaria è corretta (non distorta):

  la media delle medie campionarie è uguale alla media della popolazione.
- La media campionaria è consistente:

  all'aumentare della numerosità del campione (*n*), la media campionaria si avvicina alla media della popolazione.

## Media e varianza della popolazione

Data una popolazione *X* di *N* unità, $x_1, \ldots, x_N$, la media della popolazione è indicata con

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

La varianza della popolazione è indicata con

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

# Media e varianza della media campionaria

Dato un campione di *n* unità, estratto con o senza reinserimento, la media delle medie campionarie è uguale alla media della popolazione

$$E(\overline{X}) = \mu_{\overline{X}} = \mu$$

Dato un campione di *n* unità, estratto con reinserimento da una popolazione finita o senza reinserimento da una popolazione infinita, la varianza della media campionaria è

$$V(\overline{X}) = \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

Varianza della popolazione

Varianza della v.a. media campionaria

Dimensione del campione

# Distribuzione della media campionaria

La media campionaria è una variabile aleatoria e può assumere diversi valori al cambiare del campione selezionato.

È una variabile aleatoria nota ? (ad es. normale?)

Bisogna distinguere i seguenti casi.

# La popolazione ha una forma normale

Se la popolazione *X* è una variabile aleatoria normale con media $\mu$ e varianza $\sigma^2$, anche la media campionaria $\bar{X}$ è una variabile aleatoria normale.

Possiamo dire che nel caso di un campione estratto con reinserimento da una popolazione finita (o senza reinserimento da una popolazione infinita)

$$X \sim N(\mu, \sigma^2) \quad \rightarrow \quad \bar{X} \sim N(\mu, \sigma^2/n)$$

Sta per: X si distribuisce normalmente!

# Teorema del limite centrale (TLC)

Date $X_1$, $X_2$,…,$X_n$ v.c. indipendenti e identicamente distribuite (iid) con media μ e varianza $\sigma^2$ la v.c. $Somma = \sum_{i=1}^{n} X_i$ e di n tende a distribuirsi come una Normale con media nμ e varianza $n\sigma^2$

Se una v.c. risulta dalla somma di un grande numero di v.c. iid, la sua distribuzione può essere approssimata da una curva normale

$$Somma \sim N\left(n\mu, n\sigma^2\right)$$

## Il Teorema del Limite Centrale

Cosa accade in assenza dell'ipotesi $X \sim N$ ?

Interviene il Teorema del Limite Centrale:

quando la numerosità $n$ del campione è sufficientemente grande (pari almeno a 30), la distribuzione della media campionaria può sempre essere approssimata dalla distribuzione normale, indipendentemente dalla distribuzione di $X$.

## Il Teorema del Limite Centrale

Riassumendo:

$$\left.\begin{array}{l} X \sim N(\mu,\sigma^2) \\ \text{oppure} \\ n \geq 30 \end{array}\right\} \quad \bar{X} \sim N(\mu,\ \sigma^2/n)$$

# Se la popolazione segue una distribuzione normale

conviene standardizzare la variabile aleatoria media campionaria $X$, nel modo seguente:

$$Z_{\bar{X}} = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

…Ma… perché è importante considerarla una variabile aleatoria?

Dal momento che è possibile formare tanti campioni differenti con la stessa numerosità (n) che danno luogo a medie campionarie diverse, è possibile calcolare anche la probabilità di estrarre da un campione una particolare media.

# Esempio

**Se la popolazione ha una distribuzione normale con media μ=368 e s.q.m σ=15, quale è la probabilità che un campione casuale di 25 unità abbia una media inferiore a 365?**

Possiamo agevolare i calcoli ricorrendo alla standardizzazione:

$$Z_{\bar{X}} = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{365 - 368}{15 / \sqrt{25}} = \frac{-3}{3} = -1$$

Dalla tavola della normale standardizzata ricaviamo che l'area cumulata fino a Z=-1 è 0.1587. Quindi il 15.87% di tutti i possibili campioni di ampiezza 25 ha una media campionaria inferiore a 365.

# Stima del parametro incognito

La media campionaria $\overline{X}$ è utilizzata per stimare il parametro incognito della popolazione $\mu$.

In tal senso, è possibile pervenire a due tipologie di stima:

- Stima puntuale: il valore del parametro nella popolazione (ad es. la media) è ottenuto come valore numerico di sintesi delle osservazioni rilevate su un singolo campione casuale. Ad esempio:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \quad \Longrightarrow \quad \mu = E(\overline{X})$$

- Stima intervallare: considerando la variabilità della distribuzione campionaria, viene ottenuto un intervallo di stime campionarie, centrato sulla stima puntuale, che contiene il valore del parametro nella popolazione e al quale è associato un livello di affidabilità (INTERVALLO DI CONFIDENZA) pari a 1-α.

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

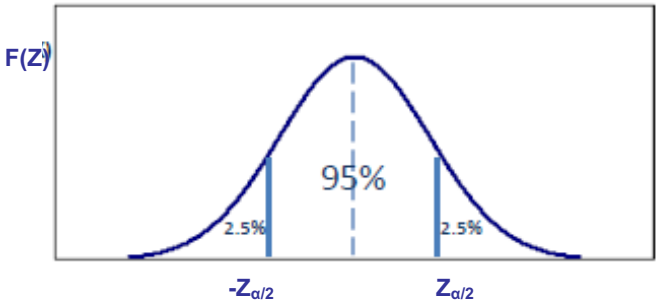
# Generico intervallo di confidenza per la media $\mu$

Se si considerano n= 36 lattine di coca-cola e si osserva il contenuto in cl di queste, sapendo che la varianza della popolazione ($\sigma^2$) è 0,5 …si può costruire un intervallo di stime plausibili per $\mu$ (media della popolazione)?

Ipotizzando che la v.a. X (contenuto lattine) si distribuisca normalmente (con $\mu$ incognito), anche $\bar{X}$ (contenuto medio delle lattine) si distribuisce normalmente.

Pertanto, nota la media campionaria, è possibile calcolare l'intervallo che contiene $\mu$ con un'affidabilità del 95%. Standardizzando si ha:

$$P\left(x_1 < \bar{X} < x_2\right) = 0.95$$




$$P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < +z_{\alpha/2}\right) = 0.95$$

dove $+z_{\alpha/2}$ è quel valore di Z (v.a. normale standard) che lascia alla destra un'area pari a $0{,}025$.

# Generico intervallo di confidenza per la media μ

$$P\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < +z_{\alpha/2}\right) = 0.95$$

Sulle tavole della Z i valori corrispondenti a $z_{0.025}$ equivalgono a +/- 1.96 e pertanto si ha:

$$P\left(-1.96 < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < +1.96\right) = 0.95$$

$$P(-1.96 \cdot \sigma/\sqrt{n} < \overline{X} - \mu < +1.96 \cdot \sigma/\sqrt{n}) = 0.95$$

$$P(-1.96 \cdot \sigma/\sqrt{n} - \overline{X} < -\mu < +1.96 \cdot \sigma/\sqrt{n} - \overline{X}) = 0.95$$

$$P(\overline{X} - 1.96 \cdot \sigma/\sqrt{n} < \mu < \overline{X} + 1.96 \cdot \sigma/\sqrt{n}) = 0.95$$

## Intervallo di confidenza per la media μ al 95%

Gli estremi dell'intervallo di confidenza per la media μ di una popolazione al livello di confidenza 0.95 (95%) sono, nel nostro esempio

$$\overline{X} - 1.96\frac{0.71}{\sqrt{36}}, \overline{X} + 1.96\frac{0.71}{\sqrt{36}} \longrightarrow \left(\overline{X} - 0.23, \overline{X} + 0.23\right)$$

σ=0,71 e n=36

Se si considerano tutti i possibili campioni di ampiezza *36*, il 95% degli intervalli di confidenza ottenuti contiene la media della popolazione.

Il 5% degli intervalli di confidenza non contiene la media della popolazione.

vedi file excel

# Intervallo di confidenza per la media (σ noto)

In generale, possiamo definire l'intervallo di confidenza per la media μ di una popolazione al generico **livello di confidenza** (1-α):

$$\Pr(\bar{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n} < \mu < \bar{X} + z_{\alpha/2} \cdot \sigma/\sqrt{n}) = (1-\alpha)$$

Dove $z_{\alpha/2}$, detto **valore critico**, è quel valore che lascia a destra un'area pari a α/2.

Se si considerano tutti i possibili campioni di ampiezza n, e per ciascuno si calcolano la media campionaria e l'intervallo centrato su questa, una percentuale pari a (1-α) degli intervalli così ottenuti contiene la media della popolazione e solo una percentuale α di essi non la contiene.

# Esempio

**Un'azienda produce laminati che dovrebbero avere una lunghezza media pari a 50 cm e uno s.q.m pari a 0.10 cm. Volendo controllare se il processo produttivo è sotto controllo, si estrae un campione casuale di n=100 laminati con media campionaria pari a 50.5. Calcolare un intervallo di confidenza al 95% per la media della popolazione.**

$$\bar{X} - 1.96 \cdot \sigma / \sqrt{n} < \mu < \bar{X} + 1.96 \cdot \sigma / \sqrt{n}$$

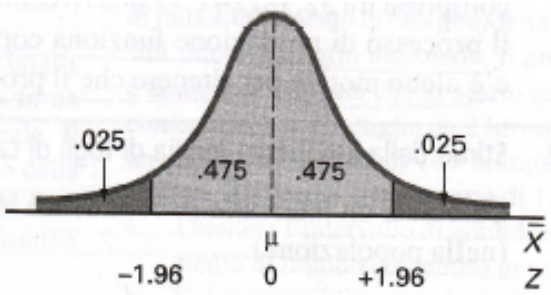
$$50.5 - 1.96 \cdot 0.10 / \sqrt{100} < \mu < 50.5 + 1.96 \cdot 0.10 / \sqrt{100}$$

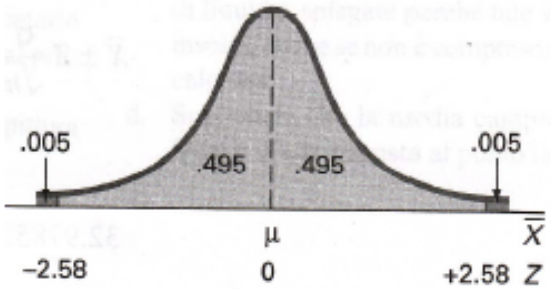
$$50.4804 < \mu < 50.5196$$

Con un livello di confidenza del 95 %, stimiamo che la media della popolazione sia compresa nell'intervallo (50.4804; 50.5196). Visto che 50 non cade all'interno dell'intervallo, il processo non è sotto controllo.

# Da notare che…

- Maggiore è il livello di confidenza (1-$\alpha$), maggiore è l'ampiezza dell'intervallo e minore è la precisione della stima.




- Maggiore è la numerosità del campione *n*, minore è, al contrario, l'intervallo.

# 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

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

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

# 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; $\sigma^2$=26,6; 1-$\alpha$=0,95

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?

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

# La popolazione ha una forma normale, ma lo sqm non è noto

In genere anche lo s.q.m della popolazione, così come la media, non è noto. In questo caso, se X ha una distribuzione normale, è possibile utilizzare la seguente variabile casuale

$$t_{n-1} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Il simbolo $t_{n-1}$ sta ad indicare la variabile aleatoria continua *t* di *Student* con *n-1* gradi di libertà (dove *n è* numerosità del campione*)*. ***S*** rappresenta lo s.q.m. del campione.

# La variabile aleatoria *t* di Student

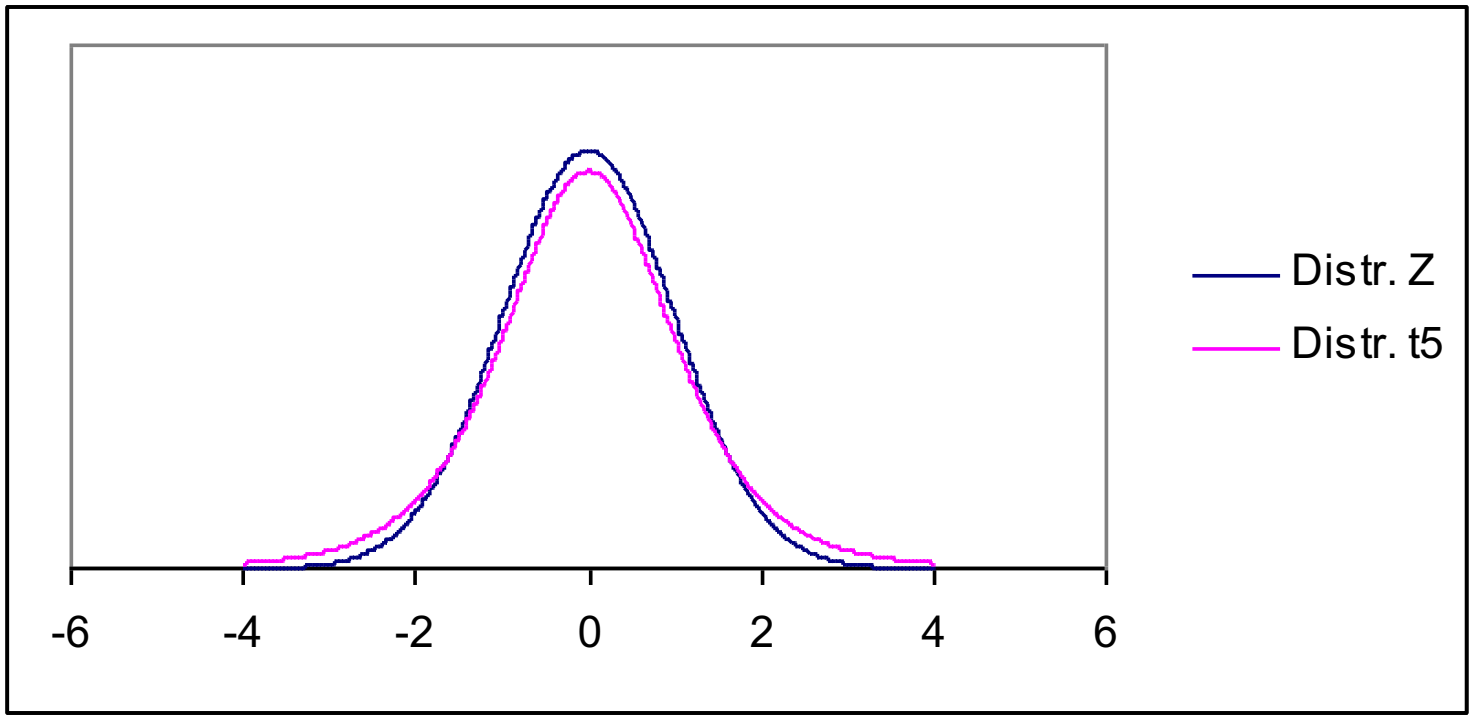
La variabile aleatoria $t_v$ di Student è una v.a. continua simile alla *Z* (v.a. normale standard).

Ha una forma campanulare con media (valore atteso) pari a 0, ovvero $E(t_v)=0$, ed è simmetrica.

La sua forma cambia al cambiare del parametro *v*.

Con $v \rightarrow \infty$, la v.a. $t_v$ tende alla *Z*.

# La $t_5$ di Student e la normale standard



# Generico intervallo di confidenza per la media $\mu$ ($\sigma$ non noto)

L'intervallo di confidenza per la media $\mu$ di una popolazione al generico livello di confidenza (1-$\alpha$), quando la varianza $\sigma^2$ non è nota, è

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

dove $t_{n-1,\alpha/2}$ è quel valore della $t_{n-1}$ che lascia alla destra un'area pari a $\alpha/2$.

La tavola t di Student riporta nella colonna madre i gradi di libertà, sulla riga madre l'area nella coda destra ed all'interno i corrispondenti valori critici $\boldsymbol{t_{n-1,\alpha/2}}$.

# Esempio

**Un'azienda produce laminati che dovrebbero avere una lunghezza media pari a 50 cm. Volendo verificare se il processo produttivo è sotto controllo, si estrae un campione casuale di n=100 laminati con media campionaria pari a 50.5 e s.q.m. pari a 0.20. Calcolare un intervallo di confidenza al 95% per la media della popolazione.**

| | Area nella coda di destra | | | | | |
|---|---|---|---|---|---|---|
| Gradi di libertà | 0.25 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 |
| 1 | 1.0000 | 3.0777 | 6.3138 | 12.7062 | 31.8207 | 63.6574 |
| 2 | 0.8165 | 1.8856 | 2.9200 | 4.3027 | 6.9646 | 9.9248 |
| 3 | 0.7649 | 1.6377 | 2.3534 | 3.1824 | 4.5407 | 5.8409 |
| 4 | 0.7407 | 1.5332 | 2.1318 | 2.7764 | 3.7469 | 4.6041 |
| 5 | 0.7267 | 1.4759 | 2.0150 | 2.5706 | 3.3649 | 4.0322 |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| 96 | 0.6771 | 1.2904 | 1.6609 | 1.9850 | 2.3658 | 2.6280 |
| 97 | 0.6770 | 1.2903 | 1.6607 | 1.9847 | 2.3654 | 2.6275 |
| 98 | 0.6770 | 1.2902 | 1.6606 | 1.9845 | 2.3650 | 2.6269 |
| 99 | 0.6770 | 1.2902 | 1.6604 | 1.9842 | 2.3646 | 2.6264 |
| 100 | 0.6770 | 1.2901 | 1.6602 | 1.9840 | 2.3642 | 2.6259 |

$$\Pr(\bar{X} - 1.9842 \cdot S/\sqrt{100} < \mu < \bar{X} + 1.9842 \cdot S/\sqrt{100}) = 0.95$$

# Esempio

**Un'azienda produce laminati che dovrebbero avere una lunghezza media pari a 50 cm. Volendo verificare se il processo produttivo è sotto controllo, si estrae un campione casuale di n=100 laminati con media campionaria pari a 50.5 e s.q.m. pari a 0.20. Calcolare un intervallo di confidenza al 95% per la media della popolazione.**

$$50.5 - 1.9842 \cdot 0.20/\sqrt{100} < \mu < 50.5 + 1.9842 \cdot 0.20/\sqrt{100}$$

$$50.46032 < \mu < 50.53968$$

Dal momento che 50 non cade all'interno dell'intervallo, possiamo concludere che il processo non è sotto controllo.

Nota che l'intervallo è maggiore rispetto al caso in cui σ è noto, quindi la precisione della stima della media della popolazione è inferiore.

# Gradi di libertà

All'aumentare dei gradi di libertà, la distribuzione t si avvicina progressivamente alla distribuzione normale, poiché all'aumentare dell'ampiezza del campione S diventa uno stimatore sempre più affidabile di σ.

**Se l'ampiezza campionaria è uguale o maggiore di 120, la distribuzioni t e la distribuzione Z coincidono.**

## Stima della proporzione *p* di una popolazione

Un campione può essere utilizzato anche per stimare la proporzione (quota) di popolazione che possiede una certa caratteristica.

Si indica con

$p$ = proporzione della popolazione (indicata anche con $\pi$)

$p_s$ = proporzione del campione

$n$ = numerosità del campione

La stima puntuale di $p$ ($\pi$) è data da $p_s$.

Per la stima intervallare si deve naturalmente costruire un intervallo di confidenza.

# Stima della proporzione *p* di una popolazione

La proporzione del campione è uno stimatore non distorto del parametro della popolazione p ($\pi$)

Si indica con

$$p_s = X/n$$

X= numero successi nel campione

*n* = numerosità del campione

Considerata la presenza della caratteristica pari ad 1 e l'assenza pari a 0, la proporzione campionaria può essere rappresentata da una **distribuzione binomiale.**

# Conseguenze del TLC

La v.c. X ~ Binomiale(n;π) risulta dalla somma di n v.c. iid di Bernoulli
In virtù del TLC, per n grande la sua distribuzione può essere approssimata da una N(nπ; nπ(1-π))

Di conseguenza

$$Z = \frac{X - n\pi}{\sqrt{n\pi \cdot (1-\pi)}} \sim N(0,1)$$

# Ancora sul teorema del limite centrale

X ~ Binomiale conta il numero di successi in n prove

Spesso si è interessati alla proporzione di successi, definita da

$$P = \frac{X}{n}$$

Al crescere di n (se *np* e *n(1-p)* sono maggiori o uguali a 5), P tende a distribuirsi come una Normale con media π e varianza π(1-π)/n

Quindi

$$\frac{P-\pi}{\sqrt{\frac{\pi\cdot(1-\pi)}{n}}} \sim \mathrm{N}(0,1)$$

$$P \sim \mathrm{N}\left(\pi, \frac{\pi\cdot(1-\pi)}{n}\right)$$

# Distribuzione della proporzione campionaria

Considerando, allo stesso modo, $\pi$=p e P=$p_s$, allora la distribuzione binomiale può essere approssimata nel modo seguente

$$p_s \sim N\left(p, \frac{p(1-p)}{n}\right)$$

È possibile standardizzare, ottenendo:

$$Z = \frac{p_s - p}{\sqrt{\frac{p(1-p)}{n}}}$$

# Intervallo di confidenza per la proporzione *p* di una popolazione al 95%

$$P\left(p_s - 1.96\sqrt{\frac{p(1-p)}{n}} < p < p_s + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

Intervallo di confidenza = stima intervallare della proporzione *p*.

In pratica, gli estremi sono

$$\left(p_s - 1.96\sqrt{\frac{p_s(1-p_s)}{n}}, p_s + 1.96\sqrt{\frac{p_s(1-p_s)}{n}}\right)$$

# Generico intervallo di confidenza per la proporzione di una popolazione

Gli estremi dell'intervallo di confidenza per la proporzione *p* di una popolazione al livello generico di confidenza (1-$\alpha$) sono

$$\left( p_s - z_{\alpha/2}\sqrt{\frac{p_s(1-p_s)}{n}}, p_s + z_{\alpha/2}\sqrt{\frac{p_s(1-p_s)}{n}} \right)$$

# Esempio

In un'indagine campionaria viene offerto un prodotto analcolico (A) e si chiede di riconoscerne la marca scegliendo tra due alternative (A e B). Si selezionano 200 consumatori e tra questi 100 indovinano la marca (A).

Costruire l'intervallo di confidenza al 95% per la proporzione dei consumatori che individuano correttamente la marca.

# Esempio

n=200 $p_s$=0.5 $\alpha$=0.05 $z_{\alpha/2}$ =1.96 ⟶ P(?<p<?)=0.95

$$P\left(p_s - 1.96\sqrt{\frac{p_s(1-p_s)}{n}} < p < p_s + 1.96\sqrt{\frac{p_s(1-p_s)}{n}}\right) = 0.95$$

$$P\left(0.5 - 1.96\sqrt{\frac{0.5(1-0.5)}{200}} < p < 0.5 + 1.96\sqrt{\frac{0.5(1-0.5)}{200}}\right) = 0.95$$

$$\boxed{P(0.43 < p < 0.57) = 0.95}$$

Selezionando tutti i possibili campioni di 200 consumatori, nel 95% dei casi la proporzione dei soggetti che indovina la marca è compresa tra il 43% e il 57%.