

# Statistica e genealogia: un amore possibile?

## Un'applicazione del teorema limite centrale per la ricerca genealogica

Vincenzo Alfano<sup>1</sup>

IAGI

5 dicembre 2021

---

<sup>1</sup>Ph.D., MA, MSc., Socio IAGI

## Introduzione

Motivazioni

La scienza delle decisioni

## La proposta

Il Teorema Limite centrale

La variabile casuale normale

## Il Tool

## Conclusione

Bibliografia

# Introduzione

## Motivazioni

Quanti in sala hanno desiderato nella propria vita di avere più tempo per la ricerca genealogica?

# Introduzione

## Motivazioni

Quanti in sala hanno desiderato nella propria vita di avere più tempo per la ricerca genealogica?

C'è una maniera (più o meno) semplice di farlo:

# Introduzione

## Motivazioni

Quanti in sala hanno desiderato nella propria vita di avere più tempo per la ricerca genealogica?

C'è una maniera (più o meno) semplice di farlo:  
ottimizzare le ricerche nel tempo dedicato alla genealogia.

# Introduzione

## Statistica

Cos'è la statistica?

# Introduzione

## Statistica

Cos'è la statistica?

Una definizione, è *la scienza delle decisioni in condizioni d'incertezza*.

# Introduzione

## Il problema

Scienze Umane  $\neq$  Scienze dure

# Introduzione

## Il problema

Scienze Umane  $\neq$  Scienze dure  
Non è sempre stato così!

# Introduzione

## Il problema

Scienze Umane  $\neq$  Scienze dure

Non è sempre stato così!

Leonardo, Michelangelo, Donatello, ma anche Pitagora, Galileo Galilei...

# Introduzione

Quante decisioni!

Quante decisioni in condizioni d'incertezza prendiamo in genealogia?

# Introduzione

## Quante decisioni!

Quante decisioni in condizioni d'incertezza prendiamo in genealogia?

Tantissime! Si pensi alla ricerca dell'atto di morte di un avo.

# Introduzione

## Quante decisioni!

Quante decisioni in condizioni d'incertezza prendiamo in genealogia?

Tantissime! Si pensi alla ricerca dell'atto di morte di un avo. Si potrebbe dire che (anche) la genealogia è *la scienza delle decisioni in condizioni d'incertezza*.

# La pratica

## Ricerca di un atto di morte

Come cercate il documento attestante la morte di un avo?

# La pratica

## Ricerca di un atto di morte

Come cercate il documento attestante la morte di un avo?  
Sostanzialmente due approcci:

# La pratica

## Ricerca di un atto di morte

Come cercate il documento attestante la morte di un avo?  
Sostanzialmente due approcci:  
Forza bruta, o...

# La pratica

## Ricerca di un atto di morte

Come cercate il documento attestante la morte di un avo?

Sostanzialmente due approcci:

Forza bruta, o...

Stima dell'anno più probabile?

## La proposta

Quel che vi propongo, è una maniera di stimare meglio ciò che già stimate!

# Il Teorema Limite Centrale

Uno dei pilastri su cui si costruiscono tanti risultati statistici è il TLC.

# Il Teorema Limite Centrale

Uno dei pilastri su cui si costruiscono tanti risultati statistici è il TLC.

Questo afferma che la somma di un gran numero di variabili casuali normali indipendenti tra loro genera a sua volta una variabile distribuita come una casuale normale.

# Il Teorema Limite Centrale

Uno dei pilastri su cui si costruiscono tanti risultati statistici è il TLC.

Questo afferma che la somma di un gran numero di variabili casuali normali indipendenti tra loro genera a sua volta una variabile distribuita come una casuale normale.

In altri termini, le variabili che sono frutto di (tante) altre variabili casuali normali, tendono a distribuirsi come variabili casuali normali.

# Il Teorema Limite Centrale

Uno dei pilastri su cui si costruiscono tanti risultati statistici è il TLC.

Questo afferma che la somma di un gran numero di variabili casuali normali indipendenti tra loro genera a sua volta una variabile distribuita come una casuale normale.

In altri termini, le variabili che sono frutto di (tante) altre variabili casuali normali, tendono a distribuirsi come variabili casuali normali. E' un risultato molto importante,

# Il Teorema Limite Centrale

Uno dei pilastri su cui si costruiscono tanti risultati statistici è il TLC.

Questo afferma che la somma di un gran numero di variabili casuali normali indipendenti tra loro genera a sua volta una variabile distribuita come una casuale normale.

In altri termini, le variabili che sono frutto di (tante) altre variabili casuali normali, tendono a distribuirsi come variabili casuali normali. E' un risultato molto importante, perché pur non sapendo nulla sulla variabile che ci interessa, arriviamo a conoscerne la distribuzione sapendo solo il processo che la genera!

# La variabile casuale normale

Ma cos'è una variabile casuale normale?

## La variabile casuale normale

Ma cos'è una variabile casuale normale?

Una variabile casuale normale, o *gaussiana*, ha una distribuzione di frequenza dalla caratteristica forma a campana,

# La variabile casuale normale

Ma cos'è una variabile casuale normale?

Una variabile casuale normale, o *gaussiana*, ha una distribuzione di frequenza dalla caratteristica forma a campana, con la maggioranza delle osservazioni attorno al centro, che rappresenta anche la media, e man mano che ci si allontana da questa media sempre meno osservazioni.

# La normale

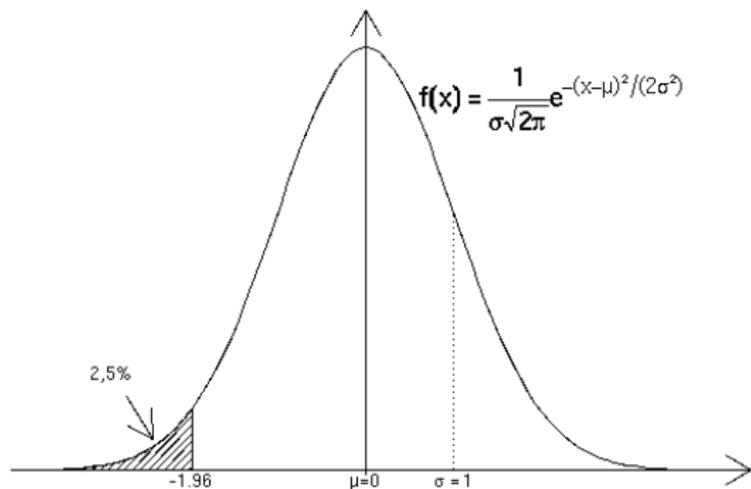


Figura: Tipica distribuzione della Casuale Normale (o Gaussiana).

## La normale

Un esempio di variabile casuale normale è l'altezza.

## La normale

Un esempio di variabile casuale normale è l'altezza.

La maggioranza delle persone al mondo, avrà un'altezza pari alla media mondiale.

## La normale

Un esempio di variabile casuale normale è l'altezza.

La maggioranza delle persone al mondo, avrà un'altezza pari alla media mondiale.

Man mano che ci si allontana dalla media, tendono a diminuire le osservazioni, cioè le persone di quell'altezza.

## La normale

Un esempio di variabile casuale normale è l'altezza.

La maggioranza delle persone al mondo, avrà un'altezza pari alla media mondiale.

Man mano che ci si allontana dalla media, tendono a diminuire le osservazioni, cioè le persone di quell'altezza.

Se dunque avremo poche persone in meno leggermente sotto e sopra la media,

## La normale

Un esempio di variabile casuale normale è l'altezza.

La maggioranza delle persone al mondo, avrà un'altezza pari alla media mondiale.

Man mano che ci si allontana dalla media, tendono a diminuire le osservazioni, cioè le persone di quell'altezza.

Se dunque avremo poche persone in meno leggermente sotto e sopra la media, allontanandoci da questo valore medio il numero di persone in cui si riscontra quello specifico valore diventa sempre più piccolo.

## La normale

Un esempio di variabile casuale normale è l'altezza.

La maggioranza delle persone al mondo, avrà un'altezza pari alla media mondiale.

Man mano che ci si allontana dalla media, tendono a diminuire le osservazioni, cioè le persone di quell'altezza.

Se dunque avremo poche persone in meno leggermente sotto e sopra la media, allontanandoci da questo valore medio il numero di persone in cui si riscontra quello specifico valore diventa sempre più piccolo.

In altri termini, avremo pochissimi nani e giganti, e tantissime persone più o meno vicine alla media.

## Una normale genealogica

Sapete cos'altro è una VCN?

## Una normale genealogica

Sapete cos'altro è una VCN?

La durata della vita, e cioè distanza tra l'atto di nascita e quello di morte, dice il teorema limite centrale!

## Una normale genealogica

Sapete cos'altro è una VCN?

La durata della vita, e cioè distanza tra l'atto di nascita e quello di morte, dice il teorema limite centrale!

La durata della vita è dovuta:

## Una normale genealogica

Sapete cos'altro è una VCN?

La durata della vita, e cioè distanza tra l'atto di nascita e quello di morte, dice il teorema limite centrale!

La durata della vita è dovuta:

- ▶ alla lunghezza della vita dei nostri genitori;

## Una normale genealogica

Sapete cos'altro è una VCN?

La durata della vita, e cioè distanza tra l'atto di nascita e quello di morte, dice il teorema limite centrale!

La durata della vita è dovuta:

- ▶ alla lunghezza della vita dei nostri genitori;
- ▶ alla lunghezza della vita dei nostri nonni, ecc..;

## Una normale genealogica

Sapete cos'altro è una VCN?

La durata della vita, e cioè distanza tra l'atto di nascita e quello di morte, dice il teorema limite centrale!

La durata della vita è dovuta:

- ▶ alla lunghezza della vita dei nostri genitori;
- ▶ alla lunghezza della vita dei nostri nonni, ecc..;
- ▶ all'alimentazione seguita;

## Una normale genealogica

Sapete cos'altro è una VCN?

La durata della vita, e cioè distanza tra l'atto di nascita e quello di morte, dice il teorema limite centrale!

La durata della vita è dovuta:

- ▶ alla lunghezza della vita dei nostri genitori;
- ▶ alla lunghezza della vita dei nostri nonni, ecc..;
- ▶ all'alimentazione seguita;
- ▶ alle ore di sport praticato;

## Una normale genealogica

Sapete cos'altro è una VCN?

La durata della vita, e cioè distanza tra l'atto di nascita e quello di morte, dice il teorema limite centrale!

La durata della vita è dovuta:

- ▶ alla lunghezza della vita dei nostri genitori;
- ▶ alla lunghezza della vita dei nostri nonni, ecc..;
- ▶ all'alimentazione seguita;
- ▶ alle ore di sport praticato;
- ▶ ai luoghi che frequentiamo;

## Una normale genealogica

Sapete cos'altro è una VCN?

La durata della vita, e cioè distanza tra l'atto di nascita e quello di morte, dice il teorema limite centrale!

La durata della vita è dovuta:

- ▶ alla lunghezza della vita dei nostri genitori;
- ▶ alla lunghezza della vita dei nostri nonni, ecc..;
- ▶ all'alimentazione seguita;
- ▶ alle ore di sport praticato;
- ▶ ai luoghi che frequentiamo;
- ▶ alla probabilità di incorrere in un incidente (ognuno con la sua specifica probabilità);

## Una normale genealogica

Sapete cos'altro è una VCN?

La durata della vita, e cioè distanza tra l'atto di nascita e quello di morte, dice il teorema limite centrale!

La durata della vita è dovuta:

- ▶ alla lunghezza della vita dei nostri genitori;
- ▶ alla lunghezza della vita dei nostri nonni, ecc..;
- ▶ all'alimentazione seguita;
- ▶ alle ore di sport praticato;
- ▶ ai luoghi che frequentiamo;
- ▶ alla probabilità di incorrere in un incidente (ognuno con la sua specifica probabilità);
- ▶ e così via...

## Una normale genealogica

La maggioranza della popolazione vivrà esattamente l'aspettativa di vita media, e man mano diminuendo fino a pochissimi nati morti e ultracentenari.

## Una normale genealogica

La maggioranza della popolazione vivrà esattamente l'aspettativa di vita media, e man mano diminuendo fino a pochissimi nati morti e ultracentenari.

Conoscendo l'*AVM*, sapremo che la probabilità massima di trovare l'atto di morte è tanti anni dopo l'anno di nascita quanto è il valore in anni dell'*AVM*.

## Una normale genealogica

La maggioranza della popolazione vivrà esattamente l'aspettativa di vita media, e man mano diminuendo fino a pochissimi nati morti e ultracentenari.

Conoscendo l'*AVM*, sapremo che la probabilità massima di trovare l'atto di morte è tanti anni dopo l'anno di nascita quanto è il valore in anni dell'*AVM*.

Difatti in corrispondenza di questo anno la variabile casuale normale anni di vita vissuta raggiunge la frequenza massima (il picco della campana).

## Una normale genealogica

La maggioranza della popolazione vivrà esattamente l'aspettativa di vita media, e man mano diminuendo fino a pochissimi nati morti e ultracentenari.

Conoscendo l'*AVM*, sapremo che la probabilità massima di trovare l'atto di morte è tanti anni dopo l'anno di nascita quanto è il valore in anni dell'*AVM*.

Difatti in corrispondenza di questo anno la variabile casuale normale anni di vita vissuta raggiunge la frequenza massima (il picco della campana).

In termini formali:

$$P_{max} = a^*$$

$$a^* = n + AVM$$

## Una normale genealogica

La maggioranza della popolazione vivrà esattamente l'aspettativa di vita media, e man mano diminuendo fino a pochissimi nati morti e ultracentenari.

Conoscendo l'*AVM*, sapremo che la probabilità massima di trovare l'atto di morte è tanti anni dopo l'anno di nascita quanto è il valore in anni dell'*AVM*.

Difatti in corrispondenza di questo anno la variabile casuale normale anni di vita vissuta raggiunge la frequenza massima (il picco della campana).

In termini formali:

$$P_{max} = a^*$$
$$a^* = n + AVM$$

dove  $n$  è l'anno di nascita del soggetto ed  $AVM$  è l'aspettativa di vita media.

## Una normale genealogica

Altro risultato che ci suggerisce il TLC è che la probabilità di trovare l'atto degrada allontanandosi dalla media nei due versi.

## Una normale genealogica

Altro risultato che ci suggerisce il TLC è che la probabilità di trovare l'atto degrada allontanandosi dalla media nei due versi. La probabilità più alta, una volta esperita la possibilità che l'atto sia in corrispondenza di  $a^*$ , sarà quella di trovare l'atto in  $a^* - 1$  ed in  $a^* + 1$  (equiprobabili).

## Una normale genealogica

Altro risultato che ci suggerisce il TLC è che la probabilità di trovare l'atto degrada allontanandosi dalla media nei due versi. La probabilità più alta, una volta esperita la possibilità che l'atto sia in corrispondenza di  $a^*$ , sarà quella di trovare l'atto in  $a^* - 1$  ed in  $a^* + 1$  (equiprobabili).

Una volta esperita anche questa possibilità, la probabilità maggiore sarà in corrispondenza di  $a^* - 2$  ed in  $a^* + 2$ , e così via...

## Una normale genealogica

Altro risultato che ci suggerisce il TLC è che la probabilità di trovare l'atto degrada allontanandosi dalla media nei due versi. La probabilità più alta, una volta esperita la possibilità che l'atto sia in corrispondenza di  $a^*$ , sarà quella di trovare l'atto in  $a^* - 1$  ed in  $a^* + 1$  (equiprobabili).

Una volta esperita anche questa possibilità, la probabilità maggiore sarà in corrispondenza di  $a^* - 2$  ed in  $a^* + 2$ , e così via...

Cosa significa questo? Che calcolato  $a^*$ , se inizio a cercare in quell'anno l'atto di morte, e proseguo cercando nei registri corrispondenti ad un anno prima ed un anno dopo  $a^*$ , due anni prima e due dopo, sino a rinvenirlo, sto massimizzando le probabilità di trovare l'atto di mio interesse, e dunque conseguentemente sto minimizzando il tempo necessario a trovarlo.

## Aspettativa di vita media

*AVM* è tanto migliore quante più informazioni si prendano in considerazione per il soggetto.

## Aspettativa di vita media

*AVM* è tanto migliore quante più informazioni si prendano in considerazione per il soggetto.

Ad esempio le donne vivono più degli uomini, o in Campania si vive in media meno che in Lombardia. *AVM*, essendo una stima, sarà tanto migliore quanto più calata su di un archetipo simile al soggetto è stata fatta la stima.

## Aspettativa di vita media

*AVM* è tanto migliore quante più informazioni si prendano in considerazione per il soggetto.

Ad esempio le donne vivono più degli uomini, o in Campania si vive in media meno che in Lombardia. *AVM*, essendo una stima, sarà tanto migliore quanto più calata su di un archetipo simile al soggetto è stata fatta la stima.

Ad esempio l'aspettativa di vita media degli italiani del 1980 è una variabile peggiore per predire l'anno di morte di un ragazzo del 1983 a Napoli che l'aspettativa di vita media degli uomini nati in Campania nel 1983.

## Aspettativa di vita media

Non resta dunque che conoscere l'*AVM* migliore per il soggetto che ci interessa.

## Aspettativa di vita media

Non resta dunque che conoscere l'*AVM* migliore per il soggetto che ci interessa.

Questo è stato fatto per voi dal sottoscritto. Recuperando dati da:

## Aspettativa di vita media

Non resta dunque che conoscere l'*AVM* migliore per il soggetto che ci interessa.

Questo è stato fatto per voi dal sottoscritto. Recuperando dati da:

- ▶ ISTAT (2019), Sistemi di indicatori territoriali;

## Aspettativa di vita media

Non resta dunque che conoscere l'*AVM* migliore per il soggetto che ci interessa.

Questo è stato fatto per voi dal sottoscritto. Recuperando dati da:

- ▶ ISTAT (2019), Sistemi di indicatori territoriali;
- ▶ Emanuele Felice (2007), *I divari regionali in Italia sulla base degli indicatori sociali (1871-2001)*, Rivista di Politica Economica;

## Aspettativa di vita media

Non resta dunque che conoscere l'*AVM* migliore per il soggetto che ci interessa.

Questo è stato fatto per voi dal sottoscritto. Recuperando dati da:

- ▶ ISTAT (2019), Sistemi di indicatori territoriali;
- ▶ Emanuele Felice (2007), *I divari regionali in Italia sulla base degli indicatori sociali (1871-2001)*, Rivista di Politica Economica;
- ▶ Galloway, Patrick (1994), *A Reconstruction of the Population of North Italy from 1650 to 1881*, European Journal of Population.

## Aspettativa di vita media

Non resta dunque che conoscere l'*AVM* migliore per il soggetto che ci interessa.

Questo è stato fatto per voi dal sottoscritto. Recuperando dati da:

- ▶ ISTAT (2019), Sistemi di indicatori territoriali;
- ▶ Emanuele Felice (2007), *I divari regionali in Italia sulla base degli indicatori sociali (1871-2001)*, Rivista di Politica Economica;
- ▶ Galloway, Patrick (1994), *A Reconstruction of the Population of North Italy from 1650 to 1881*, European Journal of Population.

Abbiamo le stime dell'*AVM* per gli italiani dal 1650 ad oggi!

# I dati

Anni	Fonte	Note
1650-1870	Galloway, 1994	Inverse projection
1871	Felice, 2007	Log ratio
1872-1890	Elaborazione da Felice, 2007	Media armonica
1891	Felice, 2007	Log ratio
1892-1910	Elaborazione da Felice, 2007	Media armonica
1911	Felice, 2007	Log ratio
1912-1937	Elaborazione da Felice, 2007	Media armonica
1938	Felice, 2007	Log ratio
1939-1950	Elaborazione da Felice, 2007	Media armonica
1951	Felice, 2007	Log ratio
1952-1970	Elaborazione da Felice, 2007	Media armonica
1971	Felice, 2007	Log ratio
1974-2017	ISTAT, 2019	Distingue per genere

# Il tool

Come funziona?

## Il tool

Come funziona?

Lanciamo l'applicazione (in formato Excel).

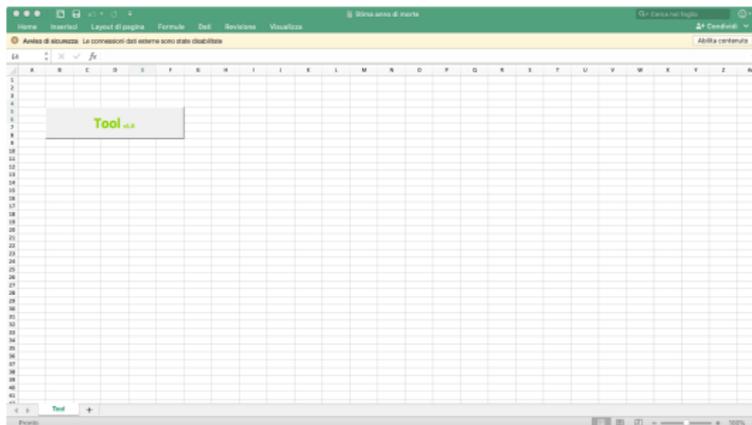


Figura: La schermata di lancio.

## Il tool

Cliccando su Tool, si aprirà una schermata con tre diversi menù a tendina:

## Il tool

Cliccando su Tool, si aprirà una schermata con tre diversi menù a tendina:

- ▶ Anno di nascita

## Il tool

Cliccando su Tool, si aprirà una schermata con tre diversi menù a tendina:

- ▶ Anno di nascita
- ▶ Territorio

## Il tool

Cliccando su Tool, si aprirà una schermata con tre diversi menù a tendina:

- ▶ Anno di nascita
- ▶ Territorio
- ▶ Sesso

# Il tool

--- Stima statistica dell'anno di morte più probabile ---

Anno di nascita  
*(dal 1871 al 2017)*

Territorio

Sesso

Stima

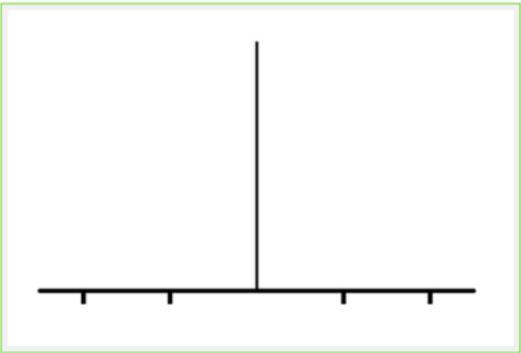


Figura: La schermata di scelta.

## Il tool

Inseriti questi dati, il programma calcola l'anno di morte più probabile, mostrando anche la distribuzione normale di probabilità in cui è possibile trovare l'anno di morte.

# Il tool

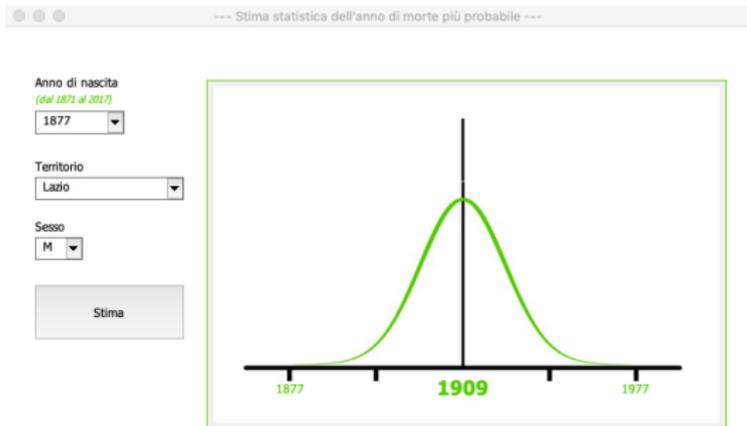


Figura: La schermata con i risultati.

*Et voilà!* Il tempo dedicato alla genealogia è ottimizzato!

## Bibliografia minima consigliata

Barbiera, Irene & Dalla Zuanna, Gianpiero (2007), *Le dinamiche di popolazione dell'Italia medievale*, Working paper series, Dipartimento di Scienze Statistiche, Università di Padova.

Billingsley, Patrick (1995), *Probability and Measure*. John Wiley & Sons.

Caratti di Valfrei, Lorenzo (1993), *Metodologia della ricerca genealogica*, Patròn editore.

Durrett, Richard (2004), *Probability: theory and examples*. Cambridge University Press.

Felice Emanuele (2007), *I divari regionali in Italia sulla base degli indicatori sociali (1871-2001)*, *Rivista di Politica Economica*.

Galloway, Patrick (1994), *A Reconstruction of the Population of North Italy from 1650 to 1881*, *European Journal of Population*.

ISTAT (2019), *Sistemi di indicatori territoriali*, [www.istat.it](http://www.istat.it), url consultato il 29/4/2019.

Piccolo, Domenico (2004), *Statistica per le decisioni*, Il Mulino.

Grazie per l'attenzione!

Grazie per l'attenzione!

Vincenzo Alfano - vinc.alfano@gmail.com - 334.3094311