# Creazione di dati verosimili su foglio di calcolo

Questo progetto consiste nel creare dati casuali, ma generati ad hoc in un foglio di calcolo e farvi delle manipolazioni atte alla consolidazione dei concetti appresi nel corso.

## Generare il dataset

La prima scheda del foglio di calcolo è la scheda dei parametri (Parameters) e conterrà 3 righe e 2 colonne (una per il titolo e una per il valore):

- `Probability`: valore della probabilità di tua scelta
- `Mean`: valor medio di tua scelta
- `StdDev`: standard deviation di tua scelta

La seconda scheda del dataset (`Data`) contiene una sola colonna, che rappresenta le età di una popolazione di Luggnagg.
I dati sono generati casualmente seguendo una distribuzione normale:

- `Age (250 individuals)`: colonna contenente le età dei 250 individui presi a campione

## Manipolazione del dataset

Per entrare dentro alla semantica dei dati che hai generato, dovrai modificare il foglio di calcolo. In particolare, il foglio avrà la seguente struttura:

### Primo tab o scheda

- Nome: "Parameters"
- Stile:
    - La colonna di intestazione ha carattere Comics Sans MS, dimensione 12pt e colore blu, riquadrata su 4 lati con bordo doppio
    - Ogni cella non del titolo è riquadrata su 4 lati in nero con il bordo sottile
    - Non sono presenti celle vuote (vengono colorate di bianco senza bordi righe e colonne a contorno dei veri e propri dati)
- Contenuto:
    - I parametri scelti per generare la distribuzione delle età della popolazione di Luggnagg

### Secondo tab o scheda

- Nome: "Data"
- Stile:
    - La prima riga di intestazione ha carattere Comics Sans MS, dimensione 12pt e colore blu, riquadrata su 4 lati con bordo doppio
    - Ogni cella non del titolo è riquadrata su 4 lati in nero con il bordo sottile

    - Non sono presenti celle vuote (vengono colorate di bianco senza bordi righe e colonne a contorno dei veri e propri dati)
    - La riga di intestazione contiene i seguenti titoli:
        - `Data`
        - `Groups`
- Contenuto:
  Nella prima colonna avrai i dati generati, mentre nella seconda interi appartenenti all'intervallo [1, 4] (estremi inclusi) generati casualmente, che ti serviranno per selezionare un sotto-campione

## Terzo tab o scheda

- Nome: "Sample"
- Stile:
    - La prima riga di intestazione ha carattere Comics Sans MS, dimensione 12pt e colore blu, riquadrata su 4 lati con bordo doppio
    - Ogni cella non del titolo è riquadrata su 4 lati in nero con il bordo sottile
    - Non sono presenti celle vuote (vengono colorate di bianco senza bordi righe e colonne a contorno dei veri e propri dati)
    - La riga di intestazione contiene i seguenti titoli:
        - `Data`
        - `Groups`
        - `Sample data`
- Contenuto:
    - Nella colonna A sono riportati i dati calcolati (hint: incolla speciale. Se salti questo passaggio i dati cambiano sempre)
    - Nella colonna B sono riportati i gruppi generati casualmente
    - Nella colonna C riporterai soltanto i valori dell'età appartenenti ad uno dei 4 gruppi (a tua scelta fra 1, 2, 3 o 4). Hint: utilizza il condizionale SE e poi...

## Quarto tab o scheda

- Nome: "Statistical insight"
- Stile:
    - La colonna di intestazione ha carattere Comics Sans MS, dimensione 12pt e colore blu, riquadrata su 4 lati con bordo doppio
    - Ogni cella non del titolo è riquadrata su 4 lati in nero con il bordo sottile
    - Non sono presenti celle vuote (vengono colorate di bianco senza bordi righe e colonne a contorno dei veri e propri dati)
    - La colonna di intestazione contiene i seguenti titoli:
        - `STDDEV`
        - `EXPECTED VALUE`
        - `COUNT`
        - `CONFIDENCE RATE`

        - `Estimation of p parameter`
        - `Confidence interval`
- Contenuto - 6 righe e 2 colonne (una per il titolo e una per il valore):
    - `STDDEV`: dove calcolerai la deviazione standard del campione
    - `EXPECTED VALUE`: dove è calcolato il valore atteso del campione (hint: media)
    - `COUNT`: che conta i numeri presenti nel campione
    - `CONFIDENCE RATE`: la confidenza con cui cerchi un intervallo di confidenza (valore tra 0 e 1)
    - `Estimation of p parameter`: valore calcolato per la confidenza
    - `Confidence interval`: estremi sinistro e destro dell'intervallo di confidenza per i valori calcolati
    - **Una casella di testo che spiega i valori ottenuti e ne giustifica il valore**

## Quinto tab o scheda

- Nome: "(Un)correlated variables"
- Stile:
    - La prima riga di intestazione ha carattere Comics Sans MS, dimensione 12pt e colore blu, riquadrata su 4 lati con bordo doppio
    - Ogni cella non del titolo è riquadrata su 4 lati in nero con il bordo sottile
    - Non sono presenti celle vuote (vengono colorate di bianco senza bordi righe e colonne a contorno dei veri e propri dati)
    - La riga di intestazione contiene i seguenti titoli:
        - `Sample data`
        - `Number of cats`
        - `Age of partner`
- Contenuto:
    - `Sample data`: copia delle età, riportate per praticità
    - `Number of cats`: numeri da 1 a 7 generati casualmente (hint: dopo la generazione copiare i valori)
    - Nella cella G2 si riporta a mo' di titolo la seguente `Correlation age and cats`
    - Nella cella H2 si calcola tale correlazione
    - Nella cella I2 si motiva il valore ottenuto
    - Nella cella G3 si riporta a mo' di titolo `Desired correlation(r)`
    - Nella cella H3 si scrive un valore inferiore e prossimo a 1, che rappresenta la correlazione che si ritiene possa esserci fra l'età degli individui campionati e l'età del partner
    - `Age of partner`: valori interi calcolati come $r\,(correlation)\ *\ age\ +\ SQRT(1-r^2)\ *\ num\_cats$
    - Nella cella G4 si riporta a mo' di titolo `Actual correlation(r')`
    - Nella cella H4 si calcola tale correlazione
    - Nella cella I4 si motiva il valore ottenuto

## Sesto tab o scheda

- Nome: "Linear regression"
- Stile:
    - La prima riga di intestazione ha carattere Comics Sans MS, dimensione 12pt e colore blu, riquadrata su 4 lati con bordo doppio
    - Ogni cella non del titolo è riquadrata su 4 lati in nero con il bordo sottile
    - Non sono presenti celle vuote (vengono colorate di bianco senza bordi righe e colonne a contorno dei veri e propri dati)
    - La riga di intestazione contiene i seguenti titoli:
        - `Y (age)`
        - `X (rank)`
- Contenuto:
    - `Y (age):` copia delle età del campione e dei partner concatenate (approssimativamente tra i 140 e i 150 valori) **ordinate** in senso crescente
    - `X (rank):` un numero intero progressivo (da 1 in su), che indica l'ordine con cui si sono presentate le persone a farsi censire
    - Al centro del foglio uno *scatterplot* delle due grandezze in esame
    - Nella cella E10 si scriva la regressione lineare per il 160esimo partecipante
    - Sotto lo scatterplot, si motivino i risultati ottenuti