UNIVERSITÀ
DEGLI STUDI DI TRIESTE

dipartimento
di ingegneria
e architettura

Corso di Laurea in Ingegneria Clinica e Biomedica
Informatica Medica I

IL LINGUAGGIO PYTHON PER L'ANALISI DEI DATI: ANALISI DEL CASO DI STUDIO

Prof. Sara Renata Francesca Marceglia

Machine learning

IRIS SETOSA

Deep learning vs machine learning

Machine Learning

Feature extraction

Classification

IRIS SETOSA

Output

Deep Learning

Feature extraction + Classification

IRIS SETOSA

Output

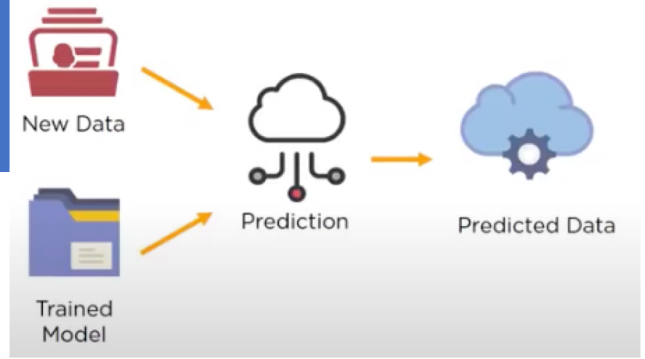
MACHINE LEARNING: AZIONI

FASE 1 - TRAINING

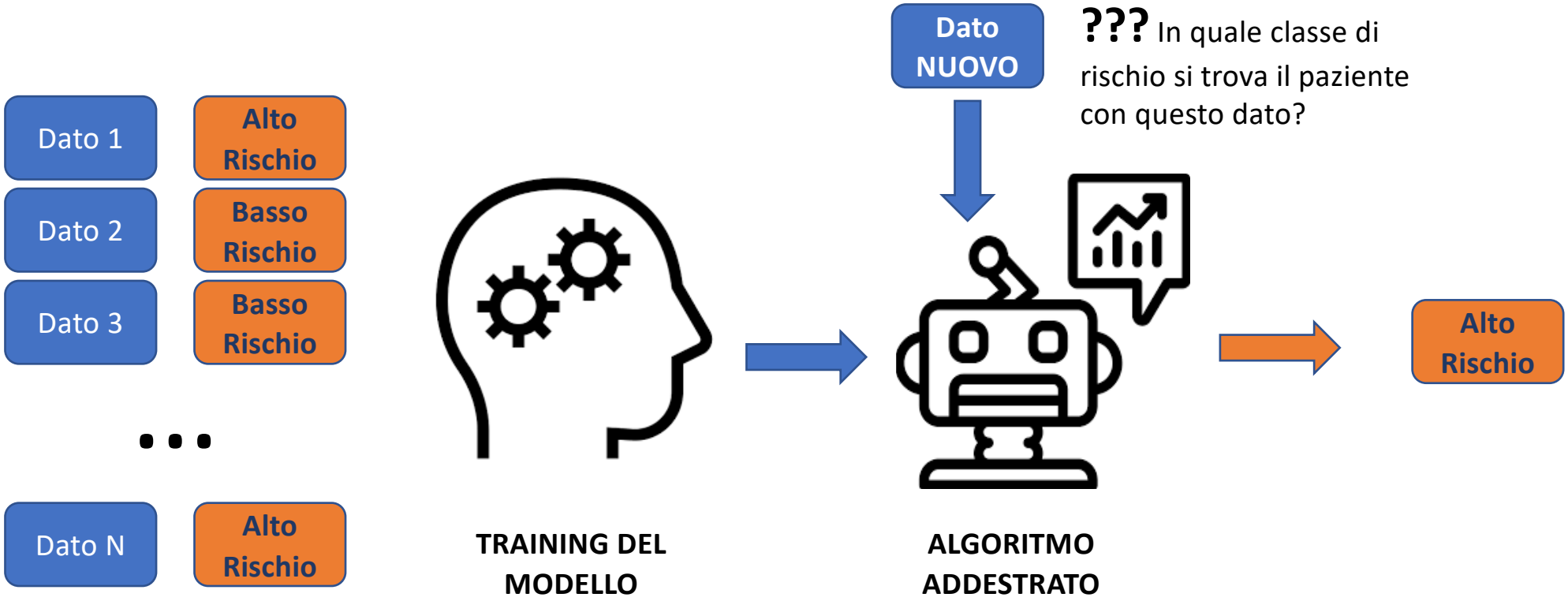
- Data collection
- Data Preprocessing
- Learning
- Testing

FASE 2 – PREDICTION

- Nuovi dati + modello
- Predizione dati

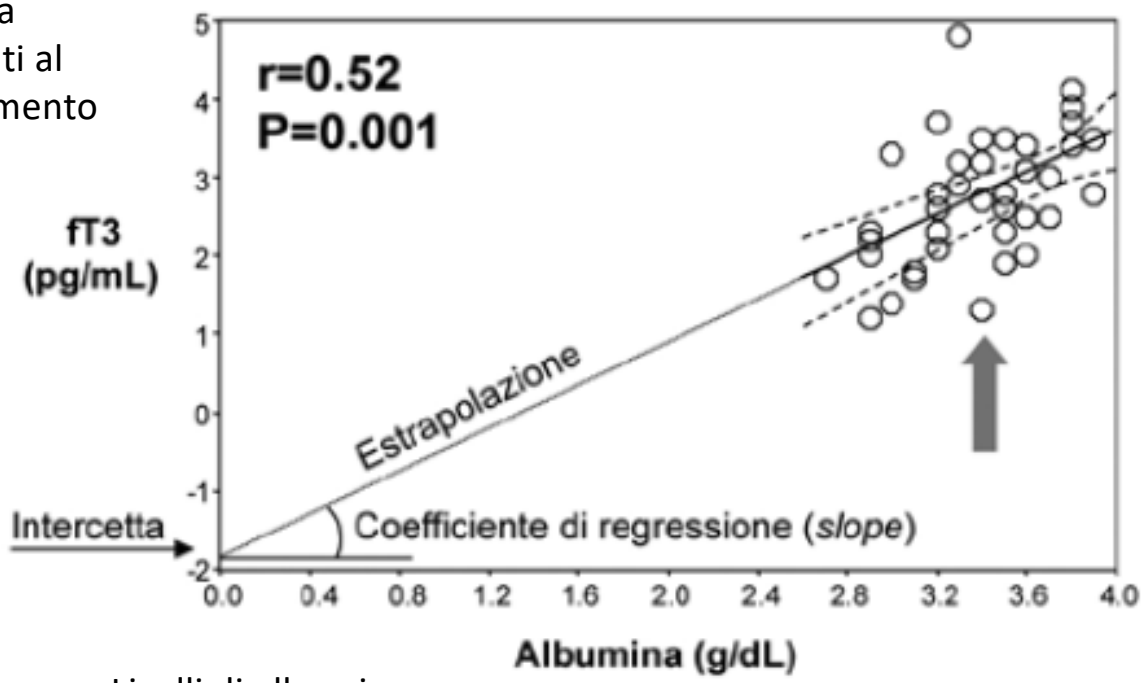


Classificazione: predizione dell'appartenenza ad una classe



Regressione: predizione di un valore

Livelli di triodotironina libera collegati al malfunzionamento tiroideo



Livelli di albumina libera collegati alla malnutrizione

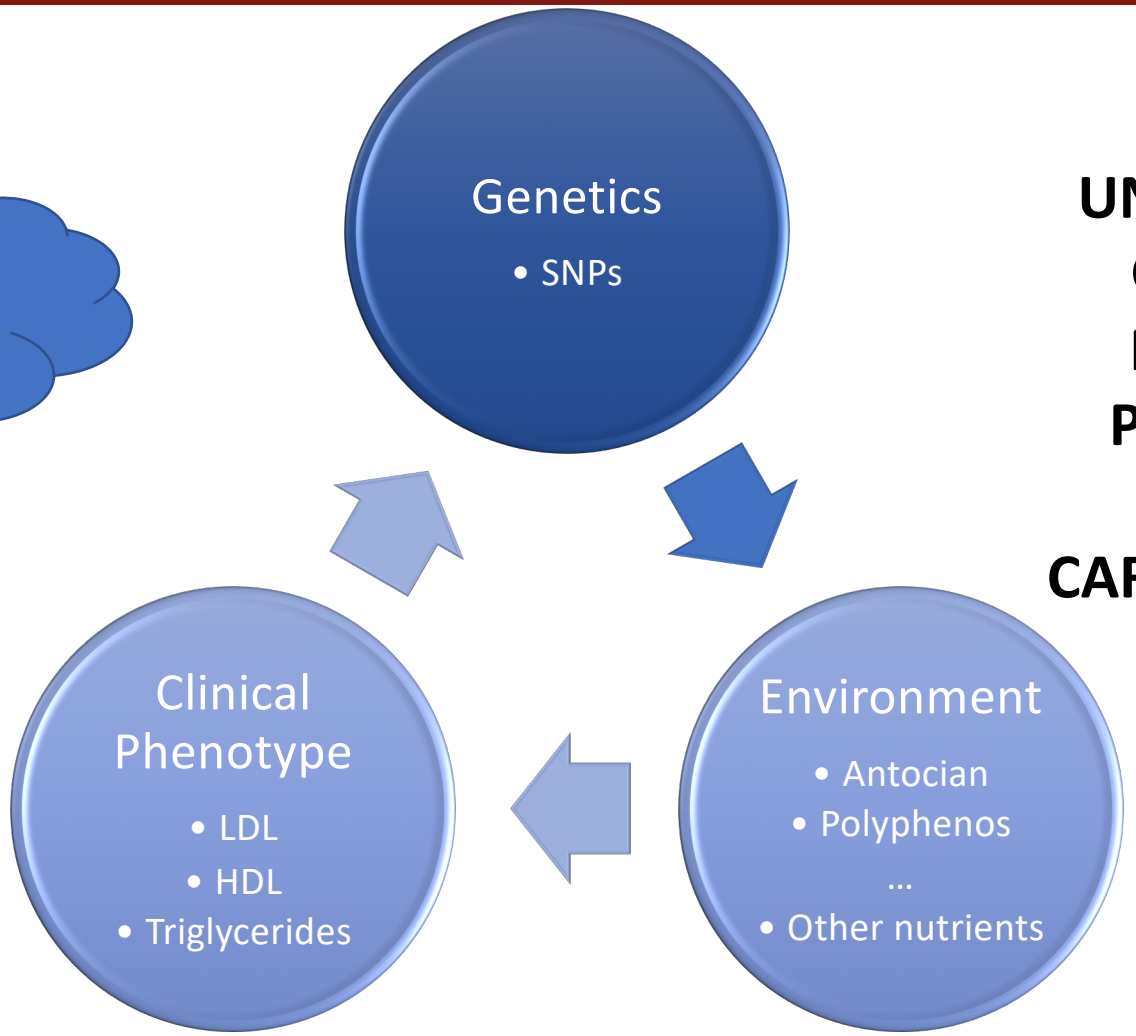
Il calcolo della regressione permette di predire il valore di ft3 a partire dai livelli di albumina libera:

Modello: retta di regressione – y = intercept + slope*x

Dati di training: ◯

Enia et al, Nephrol Dial Transplant 2007

CASO DI STUDIO: Gene x Environment analysis



CONFOUNDING FACTORS

Fattori che interagiscono/correlano sia con l'output dell'analisi (fenotipo) sia con I fattori di cui vogliamo testare il coinvolgimento:

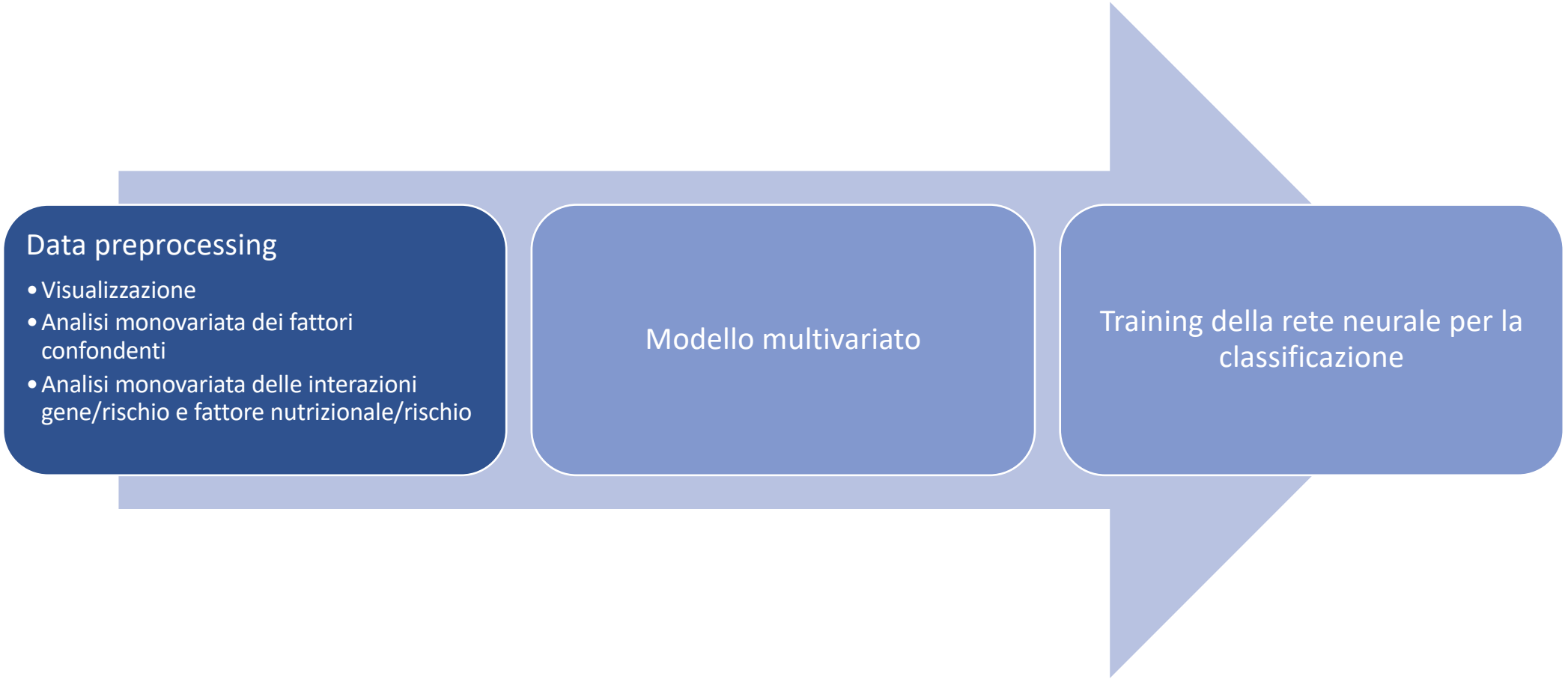
- Age
- Gender
- Smoke
- Physical activity

ESISTE UN'ASSOCIAZIONE GENE-FATTORI NUTRIZIONALI PREDITTIVA DEL RISCHIO CARDIOVASCOLARE?

CASO DI STUDIO

- La risposta alla domanda di ricercar generale dovrebbe comprendere tutti gli SNP, tutti i fattori nutrizionali e tutti i fattori di rischio cardiovascolare
- Per il nostro caso di studio consideriamo solo alcuni elementi:
  - Fattore nutrizionale: antociani
  - SNP = RS5888
  - Fattori di rischio: LDL, HDL, trigliceridi
- Obiettivo: **classificare il livello di rischio cardiovascolare (in base ai fattori di rischio noti) di un paziente data la presenza di una certa variante dello SNP e di un certo livello di assunzione del fattore nutrizionale**

WORKFLOW


Data preprocessing
• Visualizzazione
• Analisi monovariata dei fattori confondenti
• Analisi monovariata delle interazioni gene/rischio e fattore nutrizionale/rischio
Modello multivariato
Training della rete neurale per la classificazione

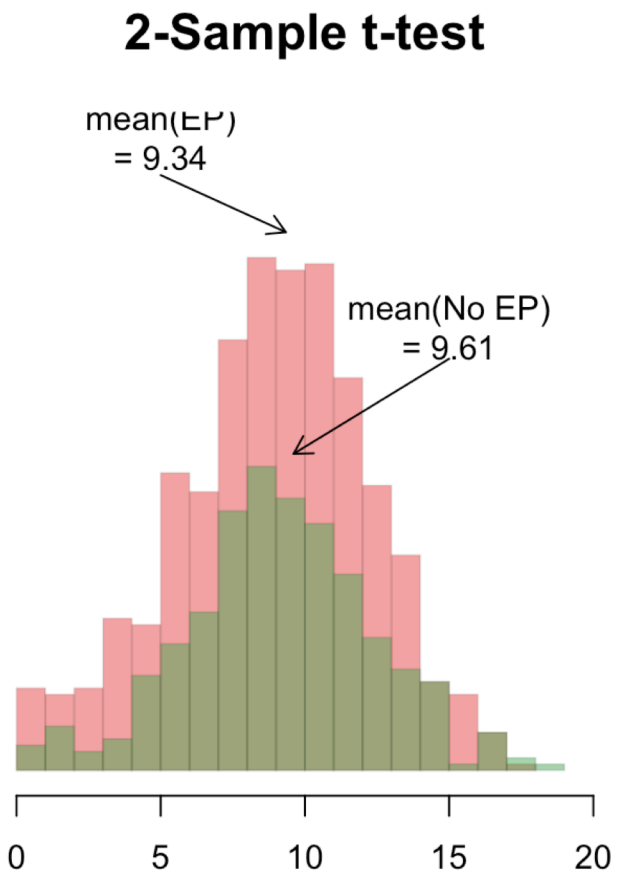
DATA PREPROCESSING

- Serve per farsi un'idea dei dati che abbiamo a disposizione
- Visualizzazione: scatterplot e grafici a barre sono utili per vedere il comportamento dei dati
- Analisi dei fattori confondenti:
  - Analisi monovariata (su singolo fattore) del fattore nutrizionale rispetto a ciascun fattore confondente
  - Analisi monovariata (su singolo fattore) della variante genetica rispetto a ciascun fattore confondente
  - Consideriamo come fattori confondenti da includere nel modello multivariate I fattori che vengono significativi
- Analisi preliminare delle interazioni:
  - Analisi monovariata (su singolo fattore) delle interazioni
    - Gene – fattori di rischio
    - Fattore ambientale – fattori di rischio

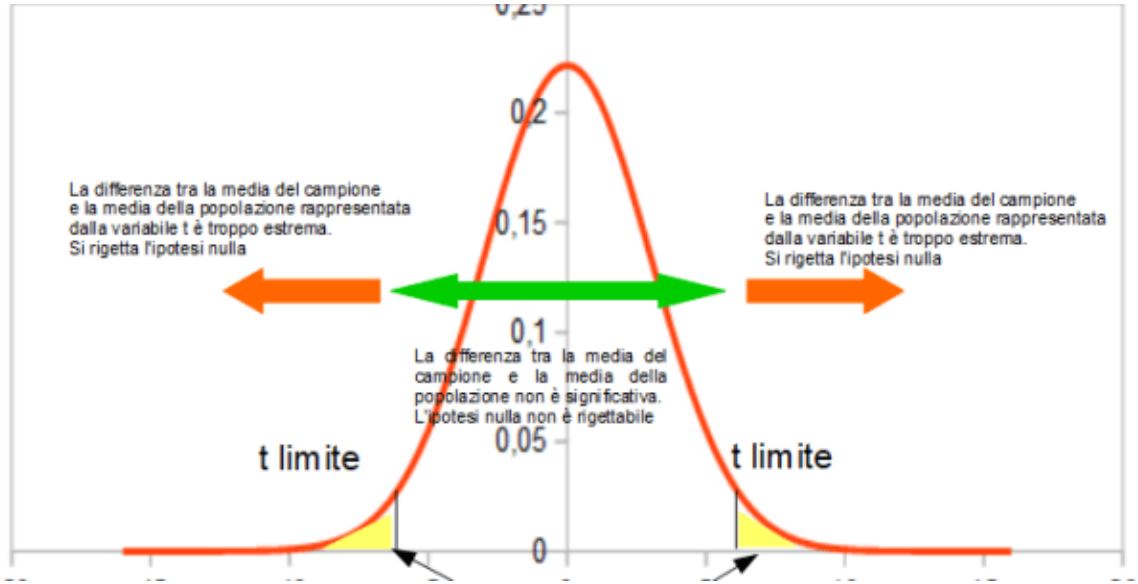
ANALISI MONOVARIATA

- Se il fattore da controllare è **categorico:**
  - **Stratificazione** = divisione della popolazione in "strati" ossia categorie in base ad una certa variabile (ad es: stratificazione per genere = dividere i dati tra maschi e femmine)
  - La visualizzazione di un box-plot (medie o mediane degli strati) è utile per avere un'idea qualitativa dell'andamento
  - Si valuta la significatività dell'influenza del fattore sulla variabile di interesse tramite test statistico:
    - T-test/ANOVA (o analogo non parametrico): se la variabile è continua
    - Chi-squared test: se la variabile è discreta
- Se il fattore da controllare è continuo:
  - Correlazione = valutazione della quantità di varianza di una certa variabile in funzione di un'altra
  - La visualizzazione di uno scatter-plot è utile per una valutazione qualitativa della correlazione
  - La significatività si valuta tramite il calcolo dei coefficienti di correlazione di Spearman (variabili non gaussiane) o Pearson (variabili gaussiane)

Test T



Verifica se due popolazioni appartengono alla stessa distribuzione usando la statistica T



Chi-squared test

OBSERVED (O)

	F1	F2	
G1	a	b	G1tot
G2	c	d	G2tot
	F1tot	F2tot	N

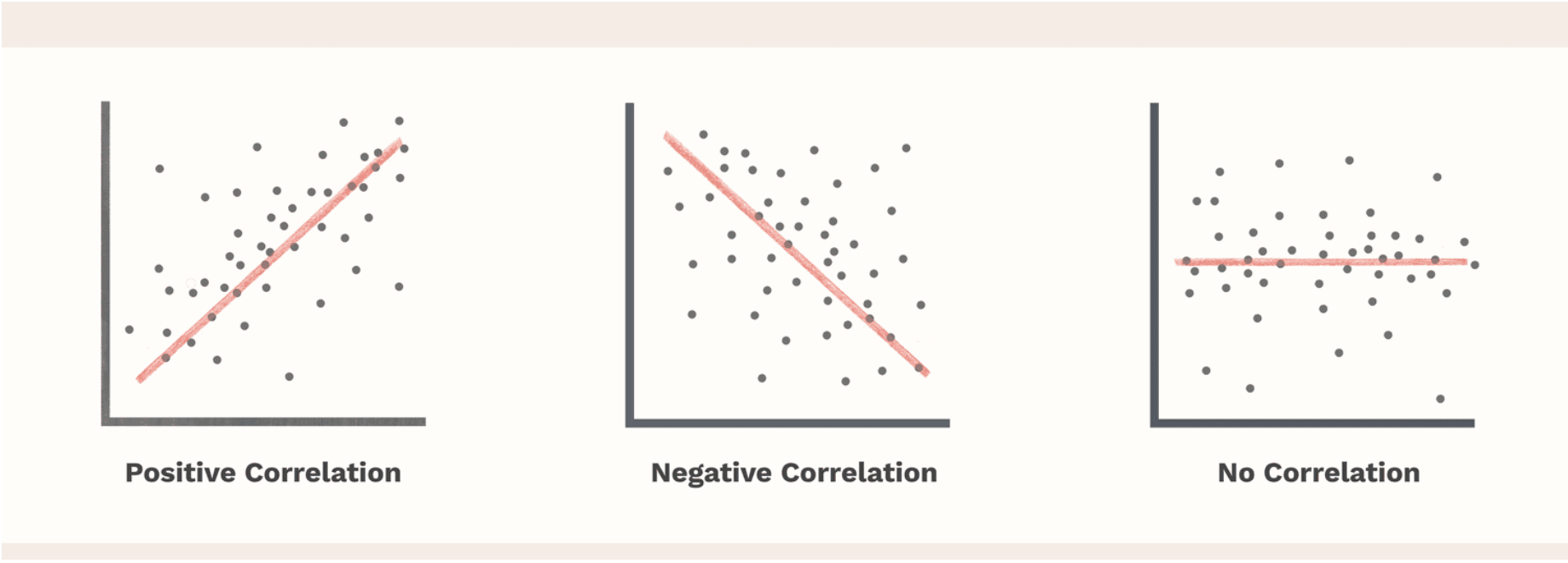
EXPECTED (E)

	F1	F2	
G1	Ea = (F1tot/N) *G1tot	Eb = (F2tot/N) *G1tot	G1tot
G2	Ec = (F1tot/N) *G2tot	Ed = (F2tot/N) *G2tot	G2tot
	F1tot	F2tot	N

Chi-squared → effettua una statistica sullo scostamento quadratico di ciascun valore osservato rispetto al valore atteso

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(a-Ea)^2}{Ea} + \frac{(b-Eb)^2}{Eb} + \frac{(c-Ec)^2}{Ec} + \frac{(d-Ed)^2}{Ed}$$

Cofficiente di correlazione



ANALISI PRELIMINARE ANTOCIANI

- Antociani:
  - Variabile continua
  - Si può discretizzare in alta- e bassa- assunzione: si stabiliscono dei valori limite a partire dalla distribuzione della popolazione (ad es: alta assunzione = ultimo terzile, bassa assunzione = primo terzile)
- Analisi monovariata fattori confondenti:
  - relazione tra antociani (variabile continua) e fattori confondenti: sesso (t-test), età(correlazione), BMI(correlazione), smoke (t-test), attività fisica (t-test)
  - Relazione tra antociani (discretizzati) e fattori confondenti: sesso (chi-squared), età (t-test), BMI (t-test), smoke (chi-squared), attività fisica (chi-squared)
- Analisi monovariata fattori di rischio:
  - Relazione tra antociani (variabile continua) e fattori di rischio (LDL, HDL, colesterolo, trigliceridi): correlazione
  - Relazione tra antociani (variabile discreta) e fattori di rischio (LDL, HDL, colesterolo, trigliceridi): t-test

ANALISI PRELIMINARE SNPs

- Analisi delle frequenze: Scegliamo 1 SNP e calcoliamo la frequenza delle varianti
- Analisi dei fattori confondenti: interazione tra i fattori confondenti e la presenza di una certa variante (ad es. Verifichiamo che i maschi non abbiano maggiore probabilità di avere una certa variante)
- Analisi monovariata fattori di rischio:
  - LDL, HDL, trigliceridi e colesterolo sono variabili continue mentre le varianti sono categoriche.
  - Se le varianti sono >2, si applica l'analisi della varianza (ANOVA) a 1 via invece che il t-test
- Interazione con antociani:
  - Verifichiamo che l'assunzione di antociani non sia in qualche modo correlata alla variante
  - Si effettua usando: t-test antociani continui vs variante (se varianti = 2), altrimenti ANOVA 1 via

ANALISI MULTIVARIATA

- Analisi multivariata: regressione multipla

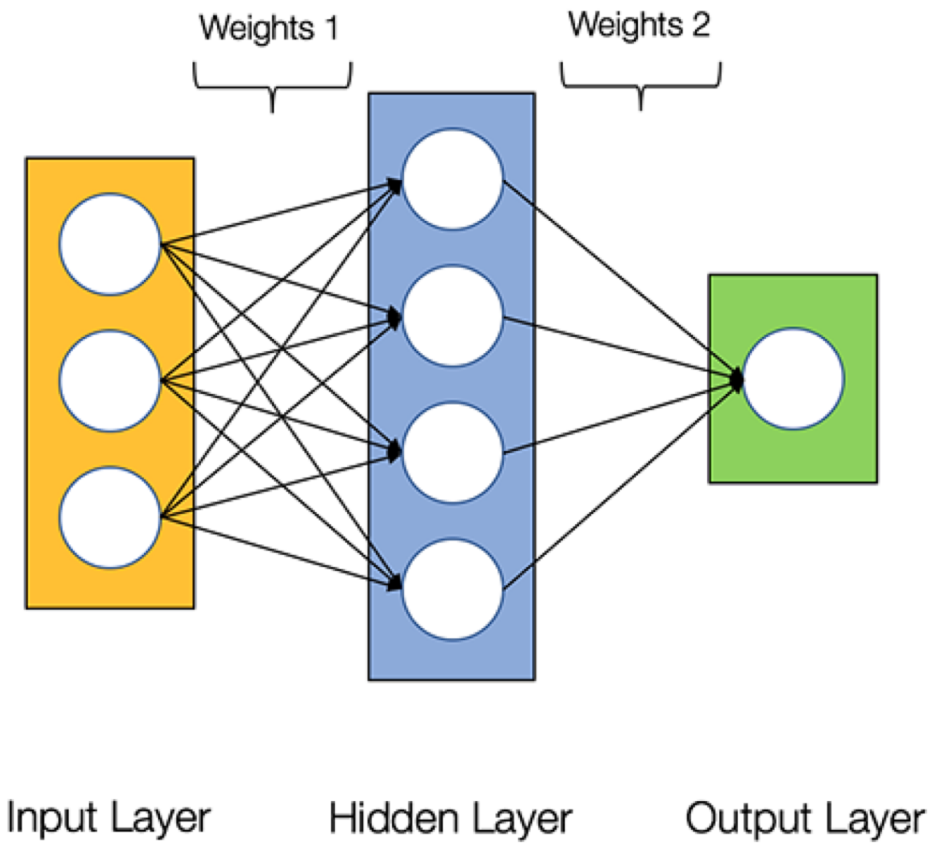
$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

  - x1, x2, .., xn = potenziali fattori interagenti che spiegano il valore della variabile y
  - Bb1, b2, …, bn = peso dei fattori
  - b0 = costante
  - Nel calcolo della regressione multipla viene assegnato un valore di significatività al peso di ciascun fattore. Vanno esclusi i fattori non significativi
- In questo caso va utilizzato un «modello lineare generalizzato» perché c'è copresenza di fattori continui e fattori categorici
- Input della regressione multipla: tutti i fattori confondenti significativi alla monovariata, il livello di antociani, le varianti dello SNP
- output: il livello del fattore di rischio (LDL, HDL, …)

RETE NEURALE

- INPUT:
  - Fattori confondenti significativi all'analisi multivariata
  - Antociani (discretizzati o continui)
  - Variante dello SNP
- OUTPUT:
  - Livello di rischio: appartenenza alla classe alta/media/bassa del livello di colesterolo LDL

RETE NEURALE



Ogni neurone nello strato nascosto è caratterizzato da una funzione di attivazione (solitamente sigmoide) → definisce quando il neurone “spara”