UNIVERSITÀ DEGLI STUDI DI TERAMO

CL in BIOTECNOLOGIE

Anno Accademico 2022/2023

CHIMICA ANALITICA

Elaborazione dei dati

Di che cosa si occupa la Statistica?

- Fisica: fenomeni naturali
- Sociologia: fenomeni sociali
- Geologia: fenomeni che riguardano la crosta terrestre
- Astronomia: fenomeni celesti
- Biologia: fenomeni della vita (biologici)
- Medicina: fenomeni che riguardano lo stato di salute
- Economia: fenomeni di gestione delle risorse
- Chimica: fenomeni sulla composizione e trasformazioni della materia
- . . . . . .
- La Statistica si occupa di fenomeni reali!
  Si presta dunque a tutte le altre discipline.
  La Statistica studia i dati.

- Il punto di partenza di una indagine statistica è costituito da un insieme (che chiamiamo popolazione di riferimento), disomogeneo all'interno (ovvero non tutti gli elementi sono uguali tra di loro) e che costituisce la parte del mondo che ci interessa.
- Gli elementi di questo insieme (che di volta in volta nei problemi concreti saranno persone, animali, batteri, immagini raccolte da un satellite,...) vengono convenzionalmente indicati come unità statistiche.
- In genere, i ricercatori studiano un sottoinsieme della popolazione relativamente piccolo (campione) e desiderano trarre conclusioni circa l'intera popolazione.
- Inferenza: come utilizzare le informazioni nel campione per trarre conclusioni sulla distribuzione delle variabili di interesse nella popolazione.
- È importante anche poter associare alle analisi condotte su un campione una valutazione dell'affidabilità dei risultati.

- La prima fase di ogni analisi statistica è rappresentata dall'organizzazione e dalla sintesi dei DATI, le informazioni raccolte sulle UNITÀ STATISTICHE che compongono il CAMPIONE.
- Concetti e strumenti fondamentali dell'analisi esplorativa sono:
  - Variabili e tipi di variabili (qualitative sconnesse o ordinali, quantitative discrete o continue).
  - Frequenze (assolute, relative, percentuali, cumulate) e tabelle.
  - Grafici (a torta, a barre, istogramma).
  - Misure di posizione (media, mediana, moda, quantili).
  - Misure di variabilità (varianza, scarto interqurtile, campio di variazione).

- I DATI sono una raccolta di informazioni (espresse in forma numerica).
- Le entità (individui, ore del giorno, ...) che vengono osservate nello studio sono dette UNITÀ STATISTICHE (casi).
- L'insieme di tutte le unità statistiche di interesse per lo studio è detto POPOLAZIONE di riferimento.
- Invece, un sottoinsieme di unità statistiche selezionate (spesso casualmente) da una popolazione è detto CAMPIONE. La dimensione del campione può variare da poche unità a molte migliaia di osservazioni.
- Una quantità di interesse nella popolazione è detta parametro, mentre la quantità calcolata sul campione è detta statistica.

ESEMPIO: La popolazione oggetto di studio è l'insieme di tutti i pazienti affetti da patologia simile, anche in futuro (si tratta di una popolazione virtuale).
Il campione è costituito dai $n = 47$ pazienti che sono entrati nell'esperimento.

DEF: Una VARIABILE (o CARATTERE) è una caratteristica di interesse rilevata sulle unità statistiche (ad esempio, età, peso, trattamento, ...).

Il termine 'variabile' evidenzia che la caratteristica di interesse può assumere una pluralità di valori. L'insieme dei valori possibili si può pensare noto, ma prima di fare l'osservazione su una unità statistica, non sappiamo quale valore si osserverà.

DEF: I valori distinti assunti da una variabile sono detti MODALITÀ della variabile. Le modalità si presumono note preliminarmente.

Esempio: nello studio sul trattamento con la realtà virtuale, la variabile *FIM* può assumere valori nell'intervallo (0, 130). Le modalità sono dunque tutti i numeri reali appartenenti a questo intervallo.

Esempio: in uno studio sulla biodiversità, si può osservare la variabile *numero di esemplari di lupo* avvistati in una settimana da un certo punto di osservazione. Le modalità sono i valori $0, 1, 2, 3, \ldots$ (i numeri naturali), anche se difficilmente si osserveranno valori grandi.

Una variabile può essere:

• QUALITATIVA o CATEGORIALE quando le sue modalità sono espresse in forma verbale (*sesso*, *livello di istruzione*, *trattamento*, ...).

A sua volta una variabile qualitativa può essere:
– SCONNESSA o NOMINALE se non esiste nessun ordinamento tra le modalità.
Esempi:
la variabile *sesso* con modalità M e F;

la variabile *modo di somministrazione* con modalità ORALE, ENDOVENA, ...

– ORDINALE se è possibile individuare un ordinamento naturale delle modalità.
Esempi:
la variabile *livello di istruzione* con modalità ELEMENTARE, MEDIA INFERIORE, MEDIA SUPERIORE, ...;

la variabile *giudizio* con modalità INSUFFICIENTE, SUFFICIENTE, DISCRETO, OTTIMO.

• Se le modalità sono solo due si parla di variabili DICOTOMICHE o BINARIE (*sesso*, *presenza*, . . .). A volte le due modalità sono espresse con valori numerici (0,1, oppure 1,2,. . .), ma il valore del numero non vuol dire assolutamente nulla!!

Oppure, una variabile può essere:

• QUANTITATIVA (o NUMERICA) quando le modalità sono espresse da numeri (età, peso,. . .). A sua volta una variabile quantitativa può essere:

– DISCRETA quando l'insieme delle modalità è finito o numerabile (stessa cardinalità dell'insieme dei naturali). Esempi:
- la variabile *numero di 'teste' in 10 lanci di una moneta*, con modalità 0,1, . . . , 10;
- le variabili *numero di sedute*, *numero di figli*, . . . con modalità 0, 1, 2, . . . ;

– CONTINUA quando l'insieme delle modalità è un intervallo, ossia un sottoinsieme, eventualmente illimitato, dei numeri reali. Esempi:
- la variabile *peso* (in kg) che ha come modalità possibili tutti i valori positivi,
- la variabile *dose* di un dato farmaco (in mg) con modalità da zero a 1000mg.
- eventuale suddivisione in classi.

- VARIABILI QUALITATIVE vs QUANTITATIVE
  - A seconda del tipo di variabili osservate, sono possibili diverse analisi statistiche.
  - Ci sono degli strumenti statistici appositi per studiare tipi diversi di variabili.
  - Tra le varie tipologie di dati è implicita una gerarchia (le variabili quantitative possono essere discretizzate, le variabili quantitative discrete possono essere tradotte in variabili qualitative ordinali, quelle ordinali possono essere considerate nominali). Le analisi statistiche sono più ricche, per così dire, ascendendo la gerarchia.
- DATI UNIVARIATI vs MULTIVARIATI
  - Le analisi univariate considerano una sola variabile rilevata sulle unità.
  - Nello studio congiunto di due variabili si parla di analisi bivariata.
  - Lo studio congiunto di due o più variabili è detto analisi multivariata (ovviamente il multivariato include il bivariato).

Per la rappresentazione grafica dei dati statistici, sono possibili vari tipi di grafici. Di seguito potete trovarne vari tipi.

Diagramma a barre

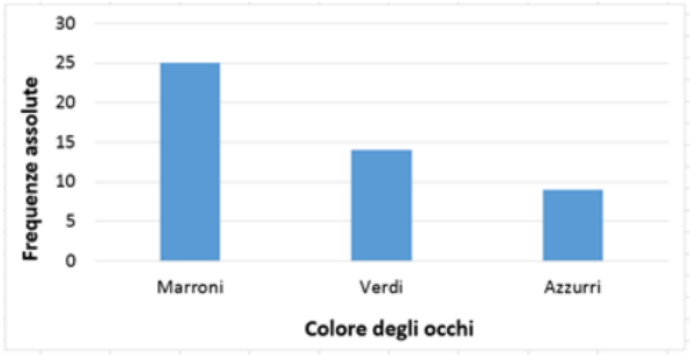
Nel diagramma a barre vengono indicate sull'asse orizzontale tutte le diverse modalità di un carattere e sull'asse verticale le corrispondenti frequenze. Per ogni modalità compare una barra rettangolare o semplicemente una linea verticale di altezza pari alla frequenza.

Esempio

Riprendiamo la tabella delle frequenze associate al colore degli occhi:

Colore degli occhi	Frequenze assolute
Marroni	25
Verdi	14
Azzurri	9

Il corrispondente diagramma a barre è il seguente:



Il diagramma a barre appena presentato riguarda le frequenze assolute, ma si possono anche considerare, allo stesso modo, le frequenze relative o percentuali.

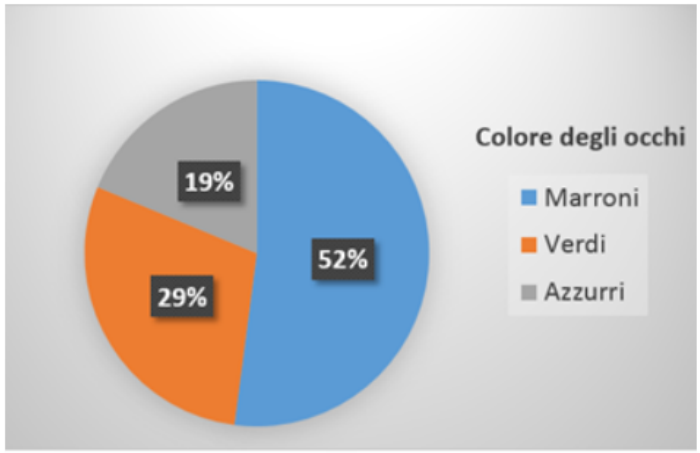
In generale il diagramma a barre si utilizza per caratteri qualitativi.

Diagramma a torta

Il diagramma a torta è un'altra forma di rappresentazione grafica che si basa essenzialmente sulla stessa costruzione del diagramma a barre. In questo caso però tutte le diverse modalità del carattere sono riportate all'interno di un cerchio (la "torta") e le frequenze corrispondenti sono rappresentate sottoforma di "fette" di diversa ampiezza.

Esempio

Se riprendiamo sempre lo stesso esempio relativo al colore degli occhi avremo il seguente diagramma a torta, riferito alle frequenze percentuali:

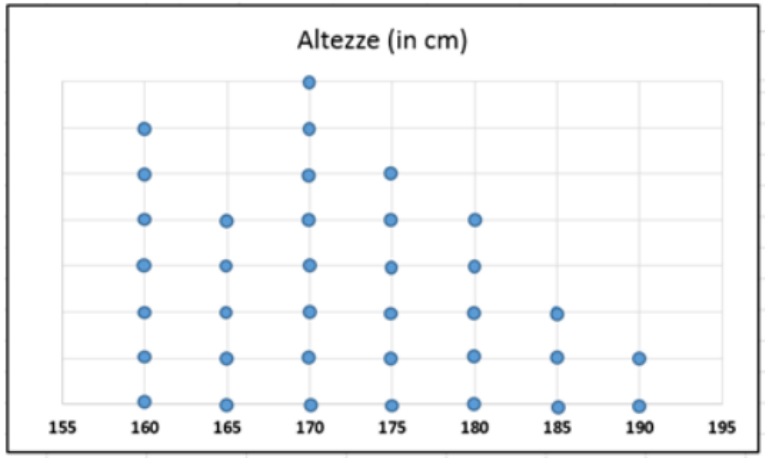


Il diagramma a dispersione

Il diagramma a dispersione si costruisce per caratteri di tipo quantitativo: sull'asse orizzontale vengono riportati i diversi valori (rispettando le proporzioni) e si disegna un punto in corrispondenza di ogni dato raccolto, effettuando eventualmente piccole approssimazioni sui valori.

Esempio

Ecco un grafico a dispersione costruito per rappresentare diversi valori di altezze registrati su un campione di 36 persone:

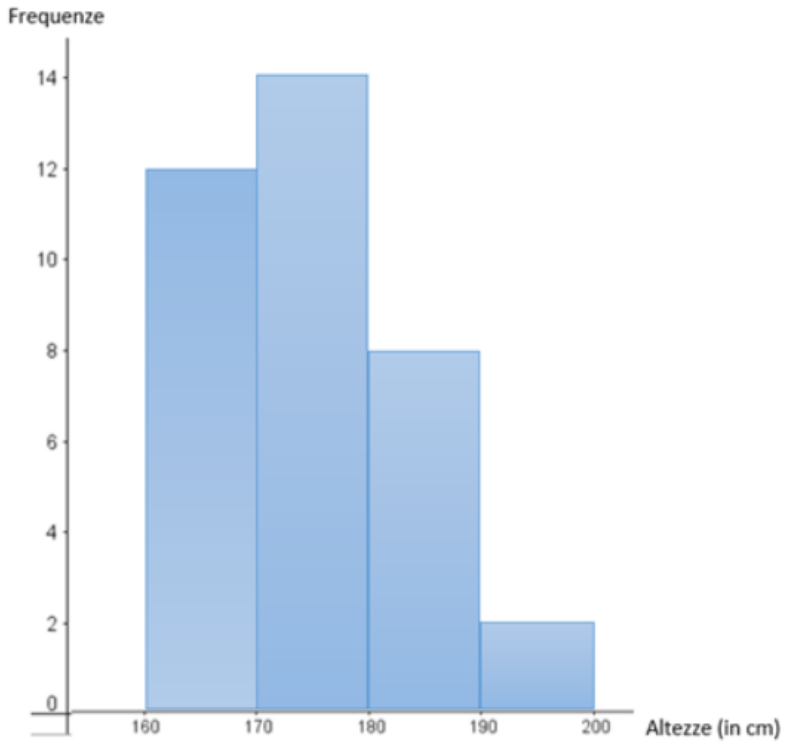


Istogramma

Una diversa forma di rappresentazione per caratteri di tipo quantitativo è l'istogramma.

Esempio

Riferendoci ai dati sulle altezze riportati nel precedente diagramma a dispersione, si può costruire il seguente istogramma:



L'istogramma è stato costruito suddividendo l'intervallo in cui variano le altezze in 4 classi di uguale ampiezza e associando ad ogni classe il numero di dati registrati in essa. Attenzione: nel caso in cui un valore coincida con un separatore di due classi, tale valore viene contato nella classe superiore.

Classi di altezze (in cm)	Frequenze
160-170	12
170-180	14
180-190	8
190-200	2

Si osserva che l'istogramma è simile, anche se solo in apparenza, al diagramma a barre. Ci sono i seguenti aspetti da tener conto:

- i rettangoli dell'istogramma devono essere adiacenti e avere come vertici i punti che separano le classi
- nell'istogramma ogni rettangolo deve avere area proporzionale alla frequenza della classe corrispondente (nel nostro caso, avendo classi di uguale ampiezza, si ha l'altezza di ogni rettangolo proporzionale alla frequenza della classe corrispondente).

T-test

Il test t (o t test) è un test parametrico di significatività statistica che utilizza la distribuzione t di Student per valutare la stima di una variabile o di un valore.
La distribuzione dei valori t ha come suoi parametri la media e l'errore standard della media, e, per campioni grandi, non è sensibile agli scostamenti dalla normalità della forma della distribuzione.
Inoltre, sempre per campioni grandi (maggiori di circa 30 casi), la distribuzione t tende a coincidere con la distribuzione normale standard (dei valori Z): come abbiamo visto e più volte ripetuto, infatti, all'aumentare delle dimensioni del campione, l'errore standard del campione tende a coincidere con la deviazione standard della popolazione.
Come si vede nella figura sopra, infatti, la curva con 3 gradi di libertà (df, degrees of freedom) è più larga e bassa rispetto alla curva con 30 df — che coincide quasi con la distribuzione normale.

I valori t di una variabile si calcolano come segue:

$$t = \frac{x_i - \mu}{ES}$$

ANOVA: l'analisi della varianza spiegata semplice

L'analisi della varianza (ANOVA, dall'inglese Analysis of Variance) comprende una serie di test statistici che rientrano nell'ambito della statistica inferenziale. Scopri in questo articolo quando si può usare, quale test scegliere e come si interpretano i risultati.

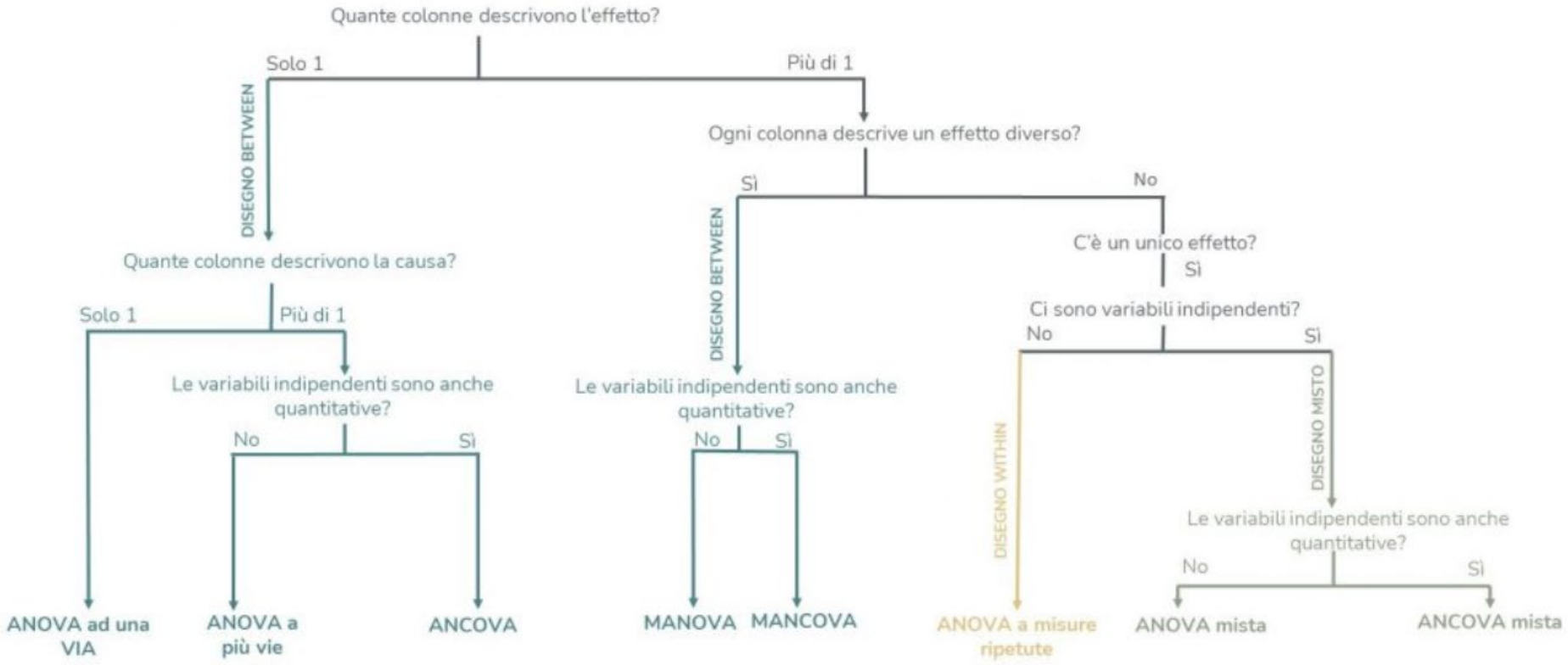
L'ANOVA è una generalizzazione del test t. Entrambe le tecniche si utilizzano infatti per il confronto di valori medi. La differenza è che:

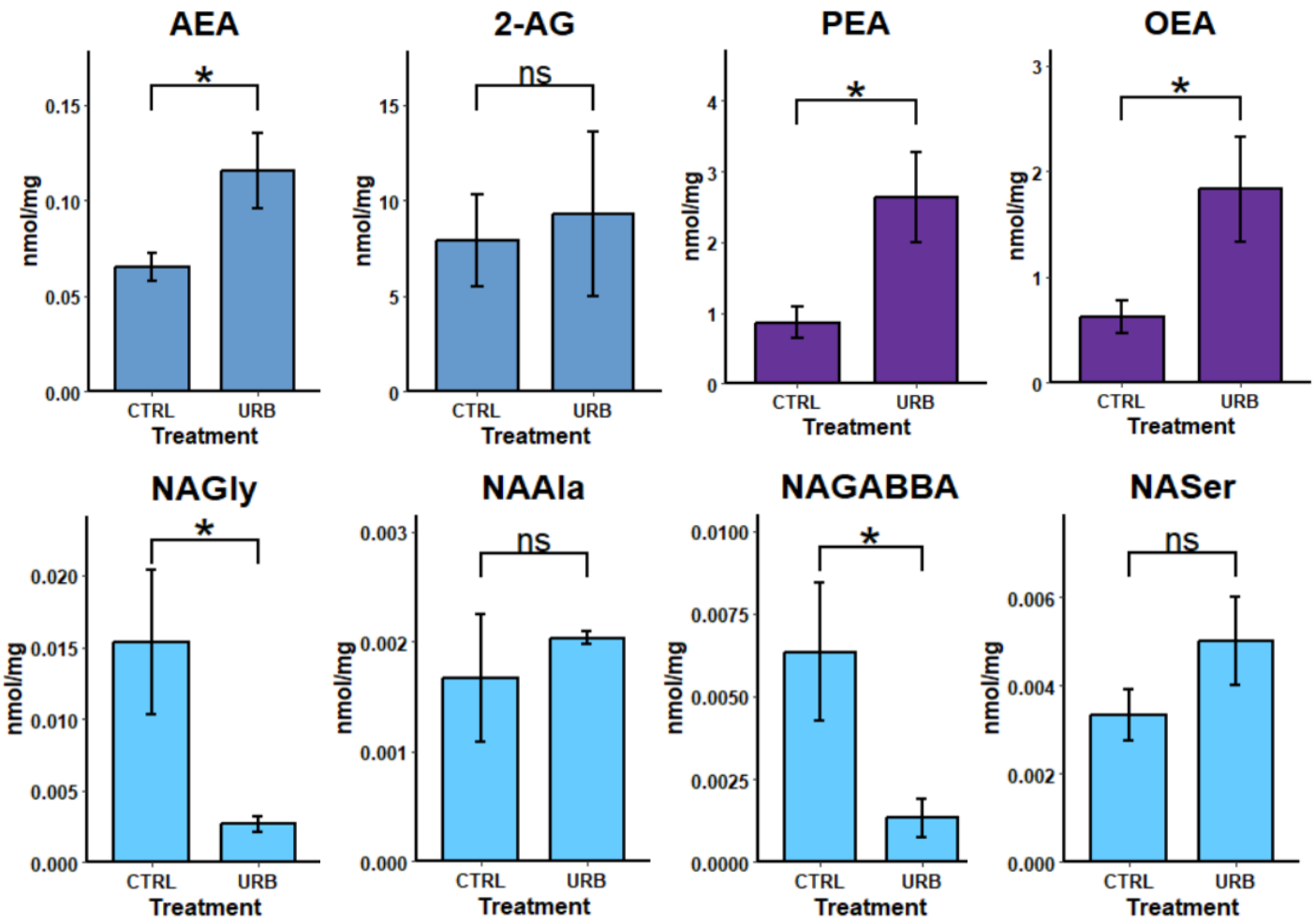
il test t permette di confrontare solo due gruppi
l'ANOVA permette di confrontare un numero qualsiasi di gruppi

Ad esempio, immagina che il tuo obiettivo sia confrontare il punteggio medio conseguito all'esame di statistica tra chi ha frequentato il corso in presenza e chi online. In questo caso, i gruppi sono solo due (frequentanti in presenza e frequentanti online). Pertanto, per il confronto delle medie puoi usare indifferentemente il test t o l'ANOVA.

Immagina ora di voler approfondire l'analisi, suddividendo lo stesso campione di studenti in tre gruppi: chi ha frequentato solo in presenza, chi ha frequentato solo online, e chi ha frequentato un po' online ed un po' in presenza. In questo caso, per il confronto delle medie non puoi più usare il test t ma devi necessariamente ricorrere all'ANOVA.

L’obiettivo dell’ANOVA è valutare gli effetti su una variabile di interesse (variabile dipendente-risposta di tipo continuo) di uno o più fattori di controllo (variabili indipendenti categoriali con due o più modalità).




AEA
*
0.15
0.10
0.05
0.00
nmol/mg
CTRL
URB
Treatment
2-AG
ns
15
10
5
0
nmol/mg
CTRL
URB
Treatment
PEA
*
4
3
2
1
0
nmol/mg
CTRL
URB
Treatment
OEA
*
3
2
1
0
nmol/mg
CTRL
URB
Treatment
NAGly
*
0.020
0.015
0.010
0.005
0.000
nmol/mg
CTRL
URB
Treatment
NAAla
ns
0.003
0.002
0.001
0.000
nmol/mg
CTRL
URB
Treatment
NAGABBA
*
0.0100
0.0075
0.0050
0.0025
0.0000
nmol/mg
CTRL
URB
Treatment
NASer
ns
0.006
0.004
0.002
0.000
nmol/mg
CTRL
URB
Treatment

L'Analisi fattoriale e la PCA

L'analisi fattoriale consiste in un insieme di tecniche statistiche che permettono di ottenere una **riduzione della complessità del numero di fattori che spiegano un fenomeno.**

Si propone quindi di **determinare un certo numero di variabili "latenti"** (fattori non direttamente misurabili nella realtà) più ristretto e riassuntivo rispetto al numero di variabili di partenza.

ES:

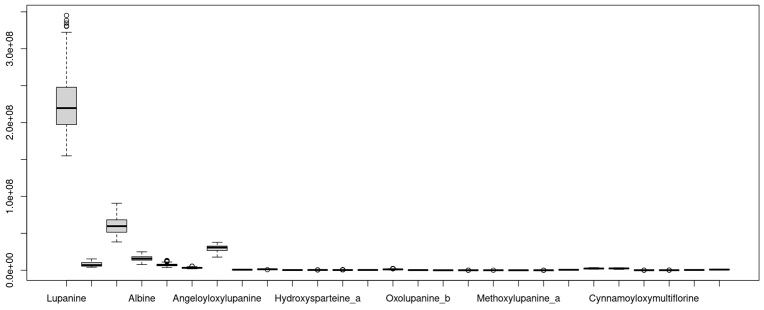
Si pensi, ad esempio, all'insieme dei voti di una popolazione di studenti di una certa scuola.

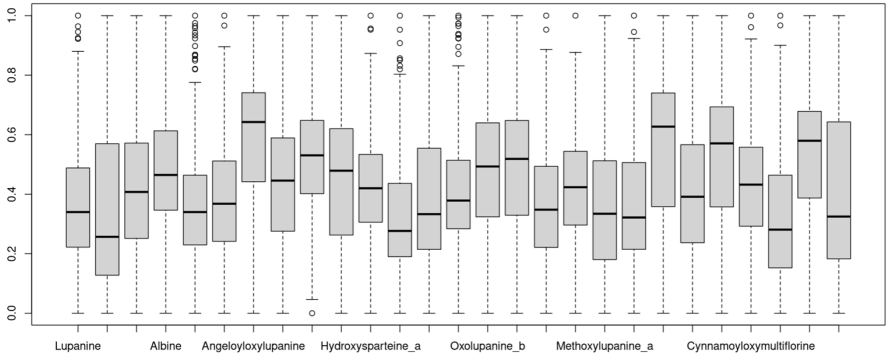
I voti riguardano il rendimenti degli stessi nelle diverse materie (italiano, matematica, scienze, geografia, storia, ecc.). È lecito supporre che le abilità di apprendimento possano distinguersi in **due fattori**: ***abilità nelle materie scientifiche*** **e** ***abilità nelle materie umanistiche*****.**

Con l'analisi fattoriale è possibile misurare queste due abilità attraverso la costruzione di due variabili latenti di sintesi (combinazione lineare) delle variabili originarie (i voti nelle diverse materie) ognuna pesata sulla base dell'importanza "*u*" (del contributo) nel discriminare gli individui sulla base delle loro abilità scientifiche e umanistiche.

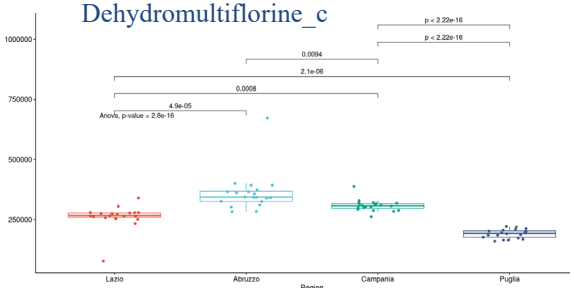
MS semi-Target: 7 target and 27 semi-Target alkaloids to classify 400 samples.
Origin Classification (Campania, Lazio, Abruzzo, Puglia)

- Lupanine
- Sparteine
- Multiflorine
- Albine
- Angustifoline
- Hydroxylupanine
- Angeloyloxylupanine
- Dehydromultiflorine_a
- Dehydromultiflorine_b
- Dehydromultiflorine_c
- Hydroxysparteine_a
- Hydroxysparteine_b
- Hydroxysparteine_c
- Oxolupanine_a
- Oxolupanine_b
- Oxolupanine_c
- Methoxymultiflorine_a
- Methoxymultiflorine_b
- Methoxylupanine_a
- Methoxylupanine_b
- Propyonyloxylupanine
- Angeloyloxymultiflorine
- Methylbutyryloxolupanine
- Cynnamoyloxymultiflorine
- Ammodendrine
- Benzoyloxylupanine_a
- Benzoyloxylupanine_b



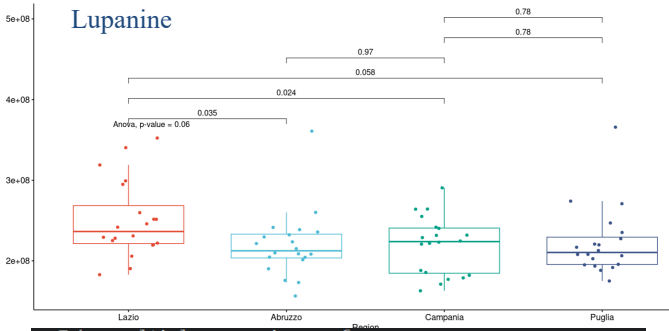


Dehydromultiflorine_c

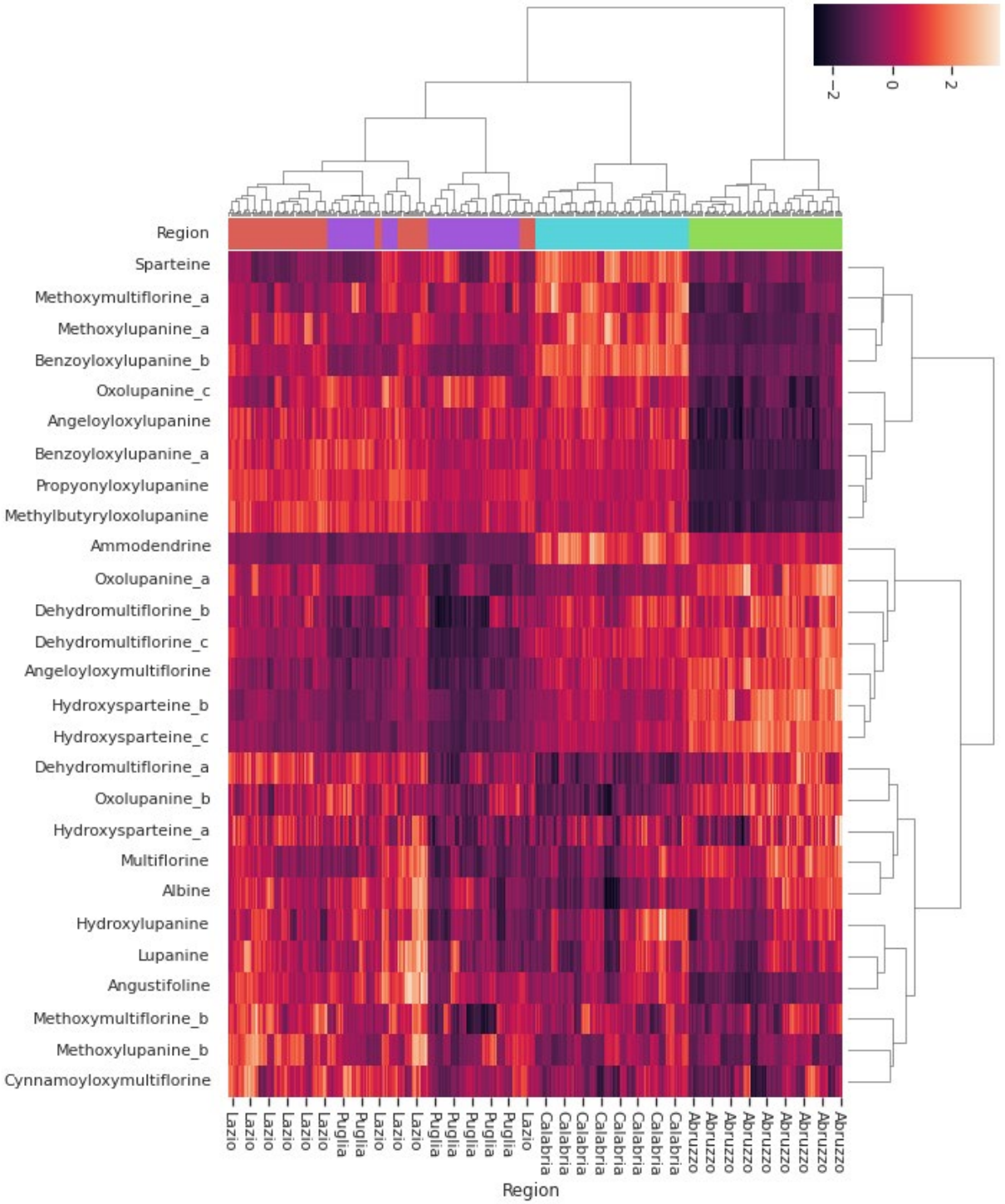


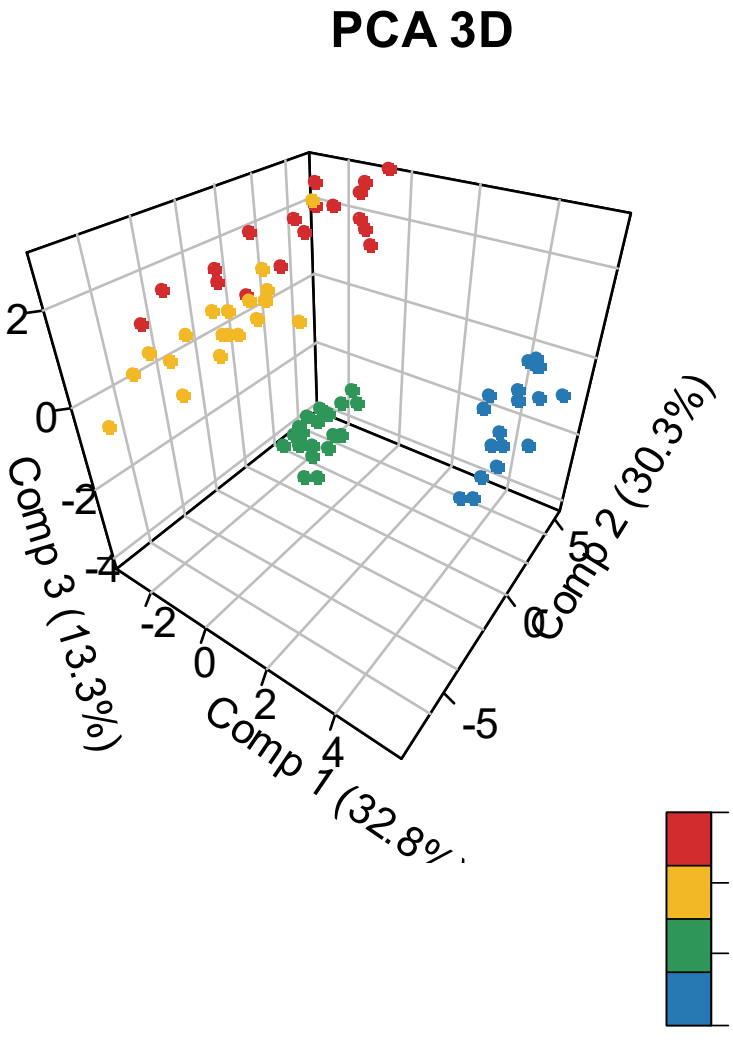

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = V10 ~ Region, data = XY)

$Region
                     diff        lwr        upr     p adj
Campania-Abruzzo   -53500  -94590.69  -12409.312 0.0054660
Lazio-Abruzzo      -99070 -140160.69  -57979.312 0.0000001
Puglia-Abruzzo    -170550 -211640.69 -129459.312 0.0000000
Lazio-Campania     -45570  -86660.69   -4479.312 0.0237823
Puglia-Campania   -117050 -158140.69  -75959.312 0.0000000
Puglia-Lazio       -71480 -112570.69  -30389.312 0.0001078
```

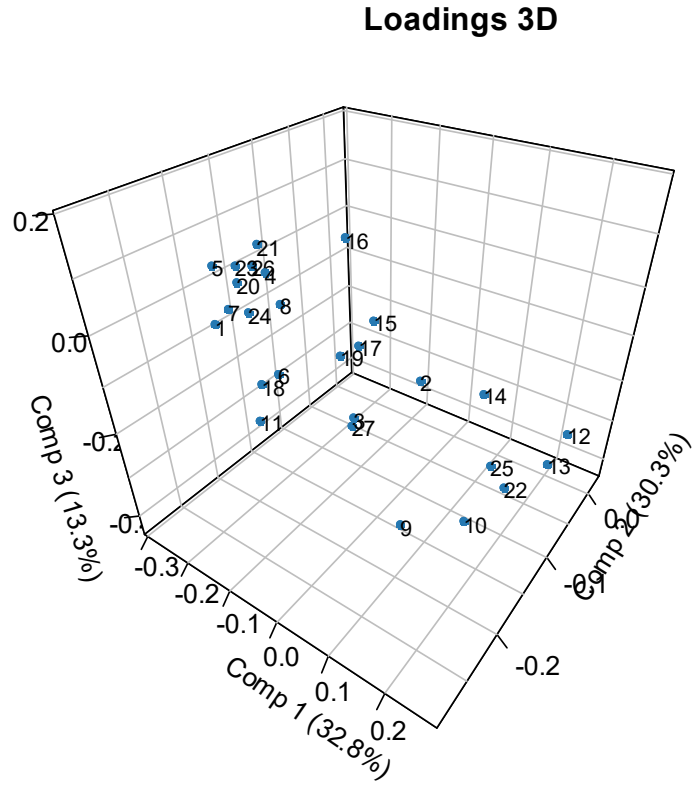
Lupanine

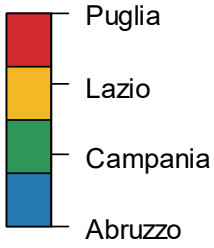



```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = V1 ~ Region, data = XY)

$Region
                     diff       lwr      upr     p adj
Campania-Abruzzo   -390000 -35265396 34485396 0.9999908
Lazio-Abruzzo    30905000  -3970396 65780396 0.1007436
Puglia-Abruzzo    3160000 -31715396 38035396 0.9952162
Lazio-Campania   31295000  -3580396 66170396 0.0943841
Puglia-Campania   3550000 -31325396 38425396 0.9932563
Puglia-Lazio    -27745000 -62620396  7130396 0.1657430
```

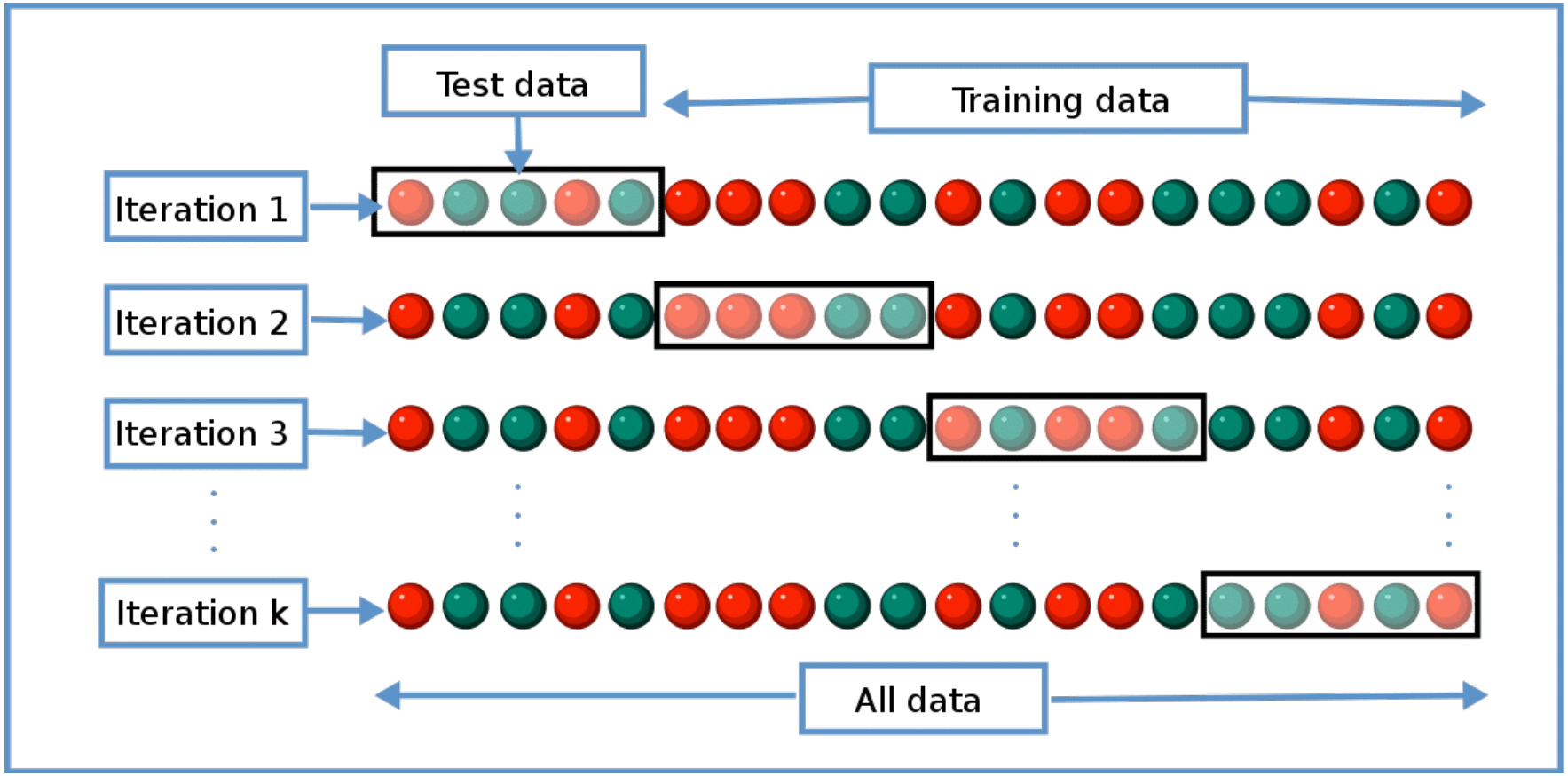

−2
0
2
Region
Sparteine
Methoxymultiflorine_a
Methoxylupanine_a
Benzoyloxylupanine_b
Oxolupanine_c
Angeloyloxylupanine
Benzoyloxylupanine_a
Propyonyloxylupanine
Methylbutyryloxolupanine
Ammodendrine
Oxolupanine_a
Dehydromultiflorine_b
Dehydromultiflorine_c
Angeloyloxymultiflorine
Hydroxysparteine_b
Hydroxysparteine_c
Dehydromultiflorine_a
Oxolupanine_b
Hydroxysparteine_a
Multiflorine
Albine
Hydroxylupanine
Lupanine
Angustifoline
Methoxymultiflorine_b
Methoxylupanine_b
Cynnamoyloxymultiflorine
Lazio
Puglia
Calabria
Abruzzo
Region

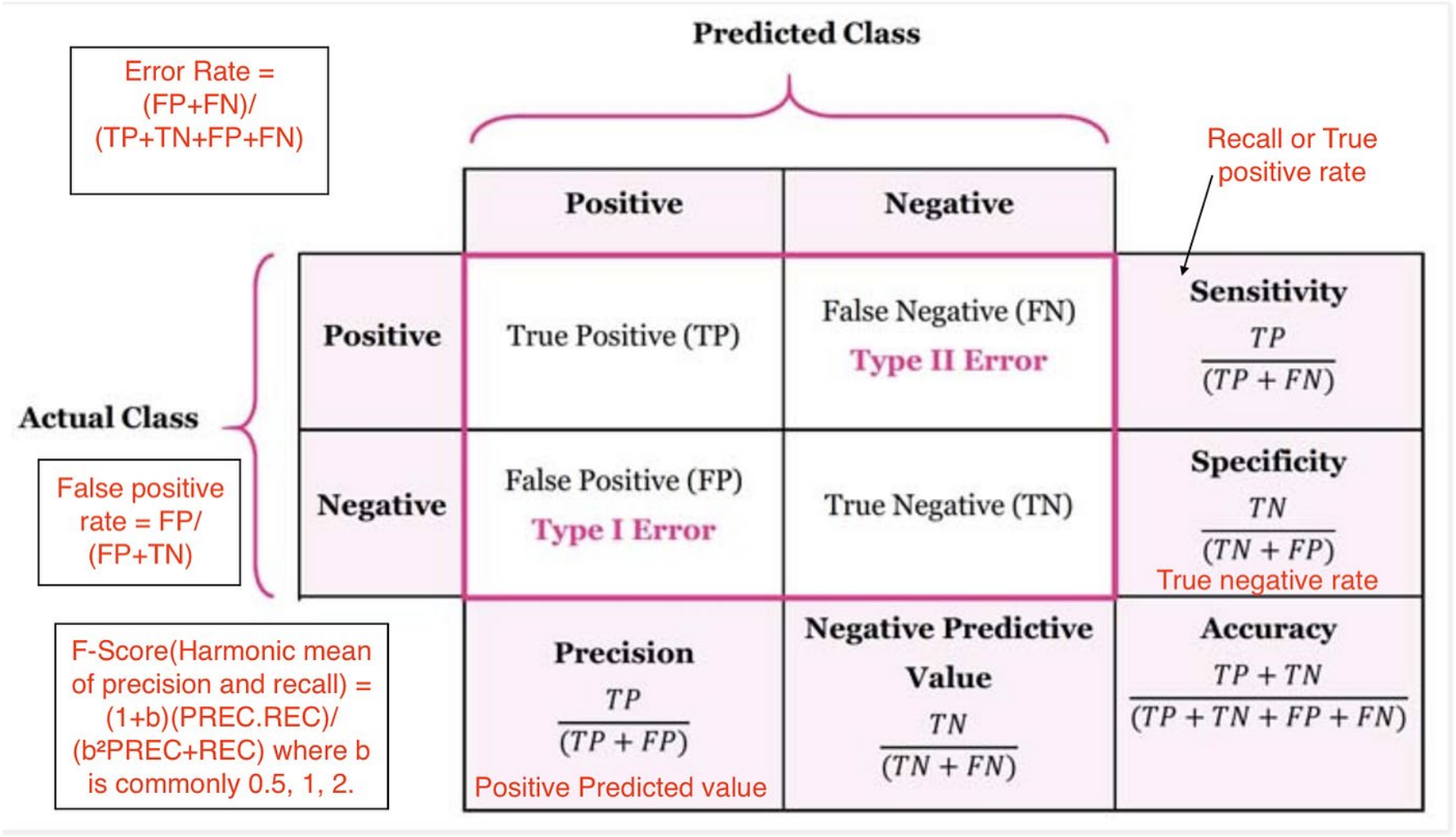

PCA 3D
Comp 3 (13.3%)
Comp 1 (32.8%)
Comp 2 (30.3%)
2
0
-2
-4
-2
0
2
4
-5
0
5


Loadings 3D
Comp 3 (13.3%)
Comp 1 (32.8%)
Comp 2 (30.3%)
0.2
0.0
-0.3
-0.2
-0.1
0.0
0.1
0.2
-0.2
0.0
0.2
1
2
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
24
25
27


Puglia
Lazio
Campania
Abruzzo

PLS-DA è una tecnica di riduzione di dimensionalità, una variante della regressione dei minimi quadrati parziali (PLS-R), che viene utilizzata quando la variabile di risposta è categorica. Si tratta di un compromesso tra l'analisi discriminante usuale e un'analisi discriminante sui componenti principali delle variabili predittive. In particolare, PLS-DA, anziché trovare iperpiani di massima varianza tra la risposta e le variabili indipendenti, trova un modello di regressione lineare proiettando le variabili previste e le variabili osservate in uno spazio nuovo. PLS-DA può fornire una buona comprensione delle cause della discriminazione attraverso pesi e carichi, conferendogli un ruolo unico nell'analisi esplorativa dei dati, ad esempio nella metabolomica tramite la visualizzazione di variabili significative come metaboliti o picchi spettroscopici.


Test data
Training data
Iteration 1
Iteration 2
Iteration 3
Iteration k
All data


Error Rate = (FP+FN)/ (TP+TN+FP+FN)
Predicted Class
Positive
Negative
Recall or True positive rate
Actual Class
Positive
True Positive (TP)
False Negative (FN)
Type II Error
Sensitivity
$\frac{TP}{(TP+FN)}$
False positive rate = FP/ (FP+TN)
Negative
False Positive (FP)
Type I Error
True Negative (TN)
Specificity
$\frac{TN}{(TN+FP)}$
True negative rate
F-Score(Harmonic mean of precision and recall) = (1+b)(PREC.REC)/ (b²PREC+REC) where b is commonly 0.5, 1, 2.
Precision
$\frac{TP}{(TP+FP)}$
Positive Predicted value
Negative Predictive Value
$\frac{TN}{(TN+FN)}$
Accuracy
$\frac{TP+TN}{(TP+TN+FP+FN)}$

PLS- Discriminant Analysis Cross-Validation (10-Fold Summary)

Semi-Target 27 alkaloids

	Abruzzo	Campania	Lazio	Puglia
Sensitivity	1.00	0.96	0.91	0.97
Specificity	1.00	1.00	0.99	0.96
Pos Pred Value	1.00	1.00	0.96	0.90
Neg Pred Value	1.00	0.99	0.97	0.99
Balanced Accuracy	1.00	0.98	0.95	0.97