**ANALISI MULTIVARIATA DEI DATI, Experimental design (Design of Experiments, DoE)**

*Come progettare esperimenti per ottenerne la massima*

quantità di informazioni della migliore qualità

L'*experimental design* o Design of Experiments (DoE) è una metodologia statistica utile per pianificare serie di esperimenti in modo da ottenere la massima quantità di informazioni della migliore qualità con il minimo numero di esperimenti

**Massima quantità di informazioni della migliore qualità?**

$$y = b_0 + b_1X_1 + b_2X_2 + \cdots + bpX_p$$

**Possibile definizione di efficienza di una metodologia di ricerca?**

$$E = (p+1) / N \leq 1$$

**E' dimostrabile che studiare gli effetti di un dato numero di fattori facendo variare *un fattore alla volta e mantenendo tutti gli altri costanti ogni volta è una* metodologia NON efficiente.**

**Si può ottenere una stima migliore degli effetti dei fattori usando piani sperimentali in cui tutti i fattori sono fatti variare simultaneamente.**

**La metodologia DoE consiste nella costruzione di piani sperimentali in cui i fattori *che si ritiene influenzino la risposta sono fatti variare tutti insieme in modo tale da* ricavare la massima quantità di informazioni della migliore qualità eseguendo il numero minimo di esperimenti indispensabile allo scopo.**

**Tali piani sperimentali sono costruiti per avere la massima efficienza.**

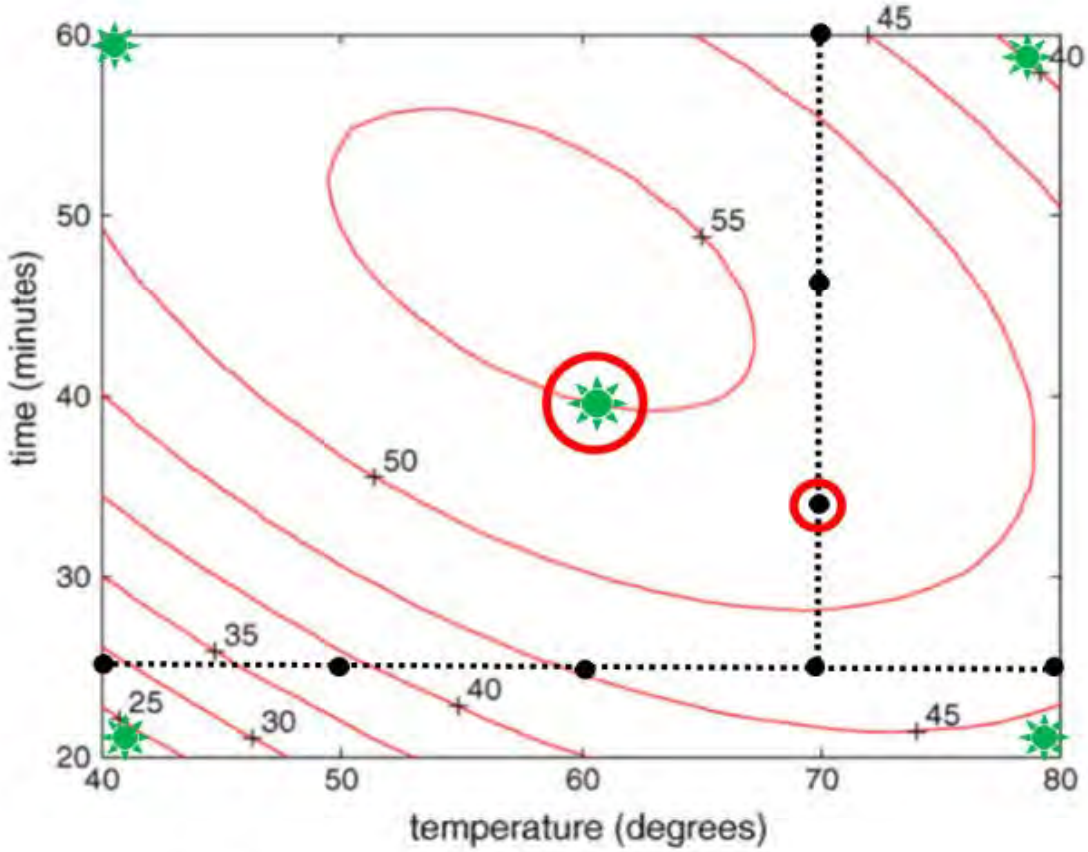



**Fig. 1.** Isoresponse plot (yield of the chemical reaction).

| N | T (°C) | t(min) | Resa(%) |
|---|---|---|---|
| 1 | 40 | 25 | 28 |
| 2 | 50 | 25 | 38 |
| 3 | 60 | 25 | 45 |
| 4 | 70 | 25 | 47 |
| 5 | 80 | 25 | 47 |
| 6 | 70 | 35 | 53 |
| 7 | 70 | 45 | 53 |
| 8 | 70 | 60 | 47 |

| N | T (°C) | t(min) | Resa(%) |
|---|---|---|---|
| 1 | 40 | 20 | 23 |
| 2 | 40 | 60 | 53 |
| 3 | 80 | 20 | 20 |
| 4 | 80 | 60 | 38 |
| 5 | 60 | 40 | 55 |

R.Leardi, Experimental design in chemistry. A tutorial. Analytical Chimica Acta (2009) 652:161-172

# L'analisi multivariata dei dati

(Prof. Riccardo Leardi, Università di Genova)

- Un esempio paradossale:

- L'altezza di 10 uomini:
- Come possiamo trattare questi dati?

| Individuo | Altezza (cm) |
|---|---|
| 1 | 175 |
| 2 | 198 |
| 3 | 168 |
| 4 | 182 |
| 5 | 178 |
| 6 | 185 |
| 7 | 177 |
| 8 | 171 |
| 9 | 188 |
| 10 | 176 |

- Interpretazione classica del dato:

| | | |
|---|---|---|
| **Media** | **179.8** | |
| **dev std** | 8.8 | |
| **min** | 168 | -1.3 σ |
| **max** | 198 | 2.1 σ |

# L'analisi multivariata dei dati

(Prof. Riccardo Leardi, Università di Genova)

- Il peso dei 10 volontari precedenti:

| Individuo | Peso (Kg) |
|---|---|
| 1 | 73 |
| 2 | 110 |
| 3 | 65 |
| 4 | 95 |
| 5 | 81 |
| 6 | 99 |
| 7 | 80 |
| 8 | 105 |
| 9 | 83 |
| 10 | 74 |

- Applichiamo la stessa analisi dati:

| | | |
|---|---|---|
| **Media** | **86.5** | |
| **dev std** | 14.9 | |
| **min** | 65 | -1.4 σ |
| **max** | 110 | 1.6 σ |

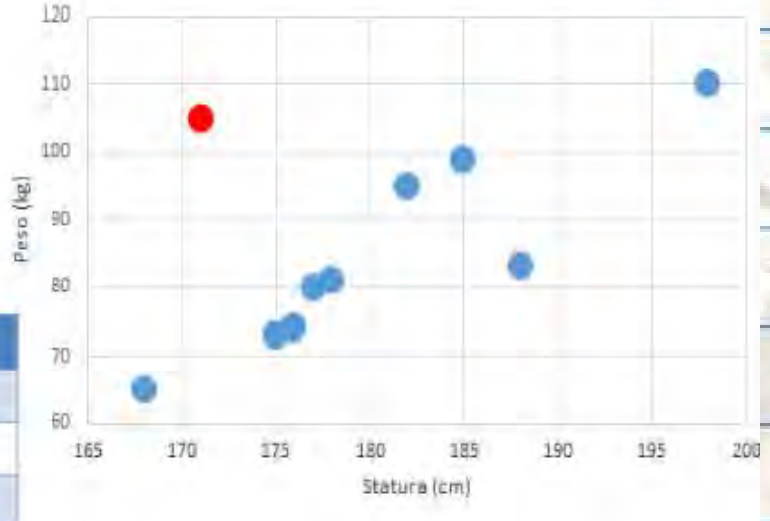
- Cosa possiamo concludere?
- Considerando un dato alla volta sembra tutto nella norma...

# L'analisi multivariata dei dati

(Prof. Riccardo Leardi, Università di Genova)

- Se però guardiamo i dati considerando la loro relazione ci accorgiamo immediatamente di un individuo con caratteristiche "fuori standard"

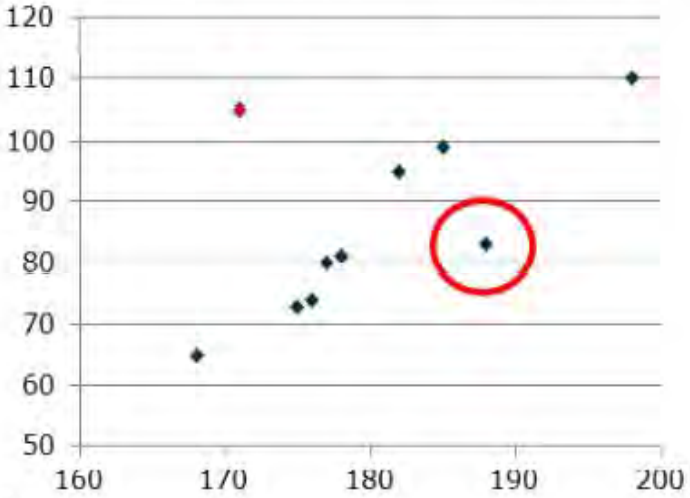
| Individuo | Altezza (cm) | Peso (kg) |
|---|---|---|
| 1 | 175 | 73 |
| 2 | 198 | 110 |
| 3 | 168 | 65 |
| 4 | 182 | 95 |
| 5 | 178 | 81 |
| 6 | 185 | 99 |
| 7 | 177 | 80 |
| 8 | 171 | 105 |
| 9 | 188 | 83 |
| 10 | 176 | 74 |



# L'analisi multivariata dei dati

(Prof. Riccardo Leardi, Università di Genova)

- Quanto tempo avete impiegato ad individuare l'individuo #8 come fuori standard?
- Se gli individui da analizzare fossero stati 100 o 1000 pensate che ce l'avreste fatta con una analisi di una variabile alla volta?
- Tutti gli altri individui sono nella norma?



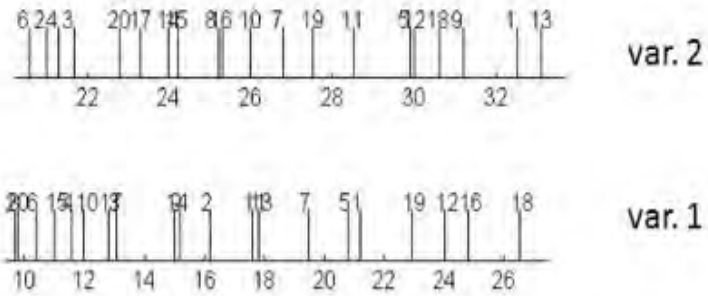

05/04/2017

- **Non considerare una variabile alla volta!**
- **Mettere i dati in grafici!**

# L'analisi multivariata dei dati

(Prof. Riccardo Leardi, Università di Genova)

- Se complichiamo (leggermente) il problema che conclusioni possiamo trarre dall'osservare una variabile alla volta?

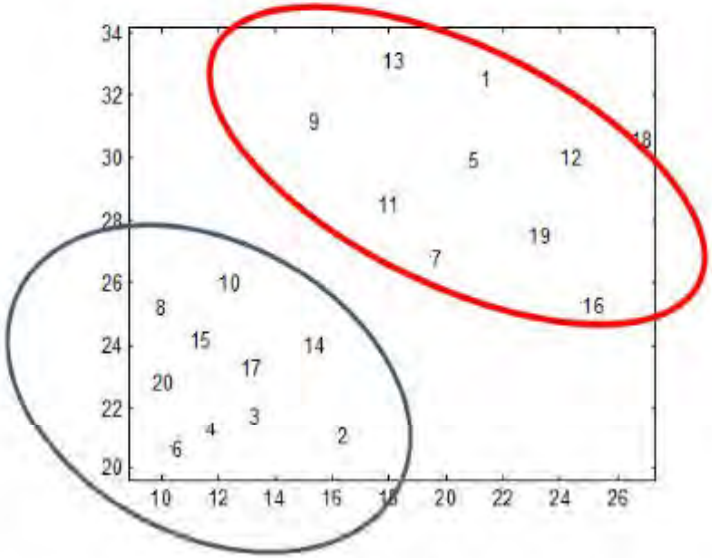



| campione | var. 1 | var. 2 |
|---|---|---|
| 1 | 21.2 | 32.5 |
| 2 | 16.2 | 21.0 |
| 3 | 13.1 | 21.7 |
| 4 | 11.6 | 21.3 |
| 5 | 20.8 | 29.9 |
| 6 | 10.4 | 20.6 |
| 7 | 19.5 | 26.8 |
| 8 | 9.8 | 25.2 |
| 9 | 15.2 | 31.2 |
| 10 | 12.0 | 26.0 |
| 11 | 17.6 | 28.5 |
| 12 | 24.0 | 30.0 |
| 13 | 17.8 | 33.1 |
| 14 | 15.0 | 24.0 |
| 15 | 11.0 | 24.2 |
| 16 | 24.8 | 25.3 |
| 17 | 12.8 | 23.3 |
| 18 | 26.5 | 30.6 |
| 19 | 22.9 | 27.5 |
| 20 | 9.7 | 22.8 |

# L'analisi multivariata dei dati

(Prof. Riccardo Leardi, Università di Genova)

- Guardiamo invece questo grafico:




- Osserviamo dei gruppi di campioni, chiamati clusters che non si riescono ad identificare analizzando in maniera semplice i dati

**I risultati degli esperimenti sono misurazioni (numeri!)**

**I risultati devono essere UTILI = dare le informazioni che rispondano al quesito posto all'inizio dello studio**

**Le informazioni devono essere UTILIZZABILI, «di buona qualità», ossia tali che da esse si possano prendere decisioni con il minimo rischio di commettere errori**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 113 | 15/06/2016 | EV61 | DGML | 2727 | 1 | acqua deio | SI | y= 0,0159x+0,1457 | 0.748 | 37.88050314 | | |
| 114 | 15/06/2016 | EV61 | DGML | 2727 | 1 | acqua deio | SI | y= 0,0159x+0,1457 | 0.796 | 40.89937107 | | |

# I DATI NON SONO INFORMAZIONE

# SENZA INTERPRETAZIONE SONO RUMORE

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 125 | 15/06/2016 | EV66 | DGML | 0 | 0 | PBS w/o Ca e Mg | SI | y= 0,0159x+0,1457 | 0.485 | 21.33962264 | | |
| 126 | 15/06/2016 | EV66 | DGML | 0 | 0 | PBS w/o Ca e Mg | SI | y= 0,0159x+0,1457 | 0.571 | 26.74842767 | | |
| 127 | 15/06/2016 | EV66 | DGML | 0 | 0 | PBS w/o Ca e Mg | SI | y= 0,0159x+0,1457 | 0.503 | 22.47169811 | 23.52 | |

**Nessun esperimento isolato ha un contenuto di informazioni utile.**

**L'informazione portata da un esperimento dipende dalla sua posizione nello spazio dei fattori rispetto alla posizione degli altri esperimenti.**

**La qualità dell'informazione dipende dalla distribuzione degli esperimenti nel dominio sperimentale.**

**La distribuzione degli esperimenti dipende dal modello scelto**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 141 | 13/07/2016 | EV16 | DGL | 0 | 0 | TRIS | SI | y=0,01235x+0,2245 | 0.225 | 0.04048583 | 4.048582996 | |

**Assegnato il modello, note le limitazioni sperimentali e il budget disponibile (= il numero massimo di esperimenti che si è disposti ad eseguire), il DoE consente di stabilire quale è il gruppo di esperimenti che darà la massima informazione possibile della migliore qualità.**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 150 | 13/07/2016 | EV20 | DGL | 664 | 1 | TRIS | SI | y=0,01235x+0,2245 | 0.224 | -0.04048583 | -4.048582996 | |
| 151 | 13/07/2016 | EV20 | DGL | 664 | 1 | TRIS | SI | y=0,01235x+0,2245 | 0.212 | -1.012145749 | 0 | |

curva BSA | campioni | Dati convalida

**Le misurazioni possono essere**

**singole, singolo oggetto: in tale caso per migliorare la qualità delle informazioni ottenute, è sufficiente aumentare il numero di prove (es. pesata di una unica massa su bilancia digitale).**

**Multiple, su più oggetti diversi tra loro: per migliorare la qualità delle informazioni ottenute dalle misure, oltre alla qualità fornita da ogni misurazione, sarà anche importante studiare le condizioni sperimentali (come è condotta la misura, es. pesata di un oggetto a temperature differenti, resa di reazione come funzione delle quantità dei reagenti e delle condizioni di reazione; forma di un picco cromatografico come risultato della scelta delle condizioni di eluizione).**

**Da questi dati è possibile ricavare un modello**

# Un modello? Perché?

Perché attraverso un modello è possibile

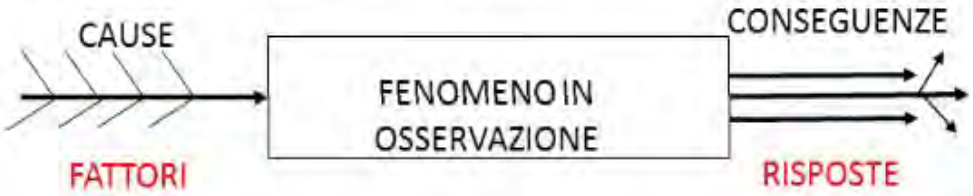
rappresentare i risultati degli esperimenti in un modo più completo e fedele alla realtà osservata = ottenere una comprensione approfondita del fenomeno in esame

fare previsioni di buona qualità (precise e accurate) sul valore della risposta in tutto il dominio sperimentale = possibilità di prevedere il valore della risposta anche in condizioni sperimentali che non sono state indagate senza necessità di condurre l'esperimento

## CHE TIPI DI MODELLI?

$$y = f(x_1, x_2, \ldots, x_p)$$

**MODELLI EMPIRICI – «soft» - il sistema è una «black box»**
**BASATI SUI RISULTATI DEGLI ESPERIMENTI**
**SEMPLICI**
**LOCALI**
**FALSI (non descrivono la realtà, mettono in relazione y con le $X_i$)**
**UTILI**
**LINEARI**




**MODELLI TEORICI, MECCANICISTICI – «hard»- il sistema è descritto completamente**
**GLI ESPERIMENTI CONFERMANO PREVISIONI DEI MODELLI (fatte a priori)**
**NON LINEARI (es. equazioni differenziali, forme esponenziali)**
**GENERALI (o validi nelle condizioni di contorno)**
**ESATTI**
**«VERI» (non confutati dall'esperienza..., sono validi fino a quando non si esegue un esperimento che confuta il modello)**

# Quali modelli semplici?

Lineari: $y = b_0 + b_1X_1 + b_2X_2 + \ldots + e$

Lineari con interazioni: $y = b_0 + b_1X_1 + b_2X_2 + b_{12}X_1X_2 + \ldots + e$

Quadratici: $y = b_0 + b_1X_1 + b_2X_2 + b_{12}X_1X_2 + b_{11}X^2{}_1 + b_{22}X^2{}_2 + \ldots + e$

# COME SI INIZIA?

**Esperienza preliminare**

**Definizione del problema: ad es. separazione di un composto A in un prodotto B**

**Informazioni di base: esperienza dell'operatore, ricerca di letteratura...**

**Prove iniziali: si ha qualche informazione sul comportamento (ad es. cromatografico) di A**

**Si scelgono**

1. **le condizioni sperimentali «migliori» sulla base dell'esperienza preliminare acquisita;**
2. **l'intervallo di variazione dei fattori intorno alle condizioni «migliori»;**
3. **il piano sperimentale più semplice che indaghi il dominio sperimentale di interesse.**