

## STATISTICAL APPROACH

Generare un modello probabilistico delle distribuzio  
ne dei dati e poi usarlo per identificare gli  
oggetti nelle regioni a bassa probabilità come  
outlier.

## PARAMETRICO VS NON-PARAMETRICO

Come capire se i dati seguono la distribuzione normale?  
La risposta è usare un test come il Q-Q.

## QUANTILE-QUANTILE PLOT

Il grafico Q-Q è un grafico che confronta due quantile, dove per quantile si intende la percentuale di dati che stanno sotto una certa percentuale (25-percentile ovvero, dividendo la distribuzione dei dati in 4 parti uguali il 25% dei dati è sotto quel punto percentile)

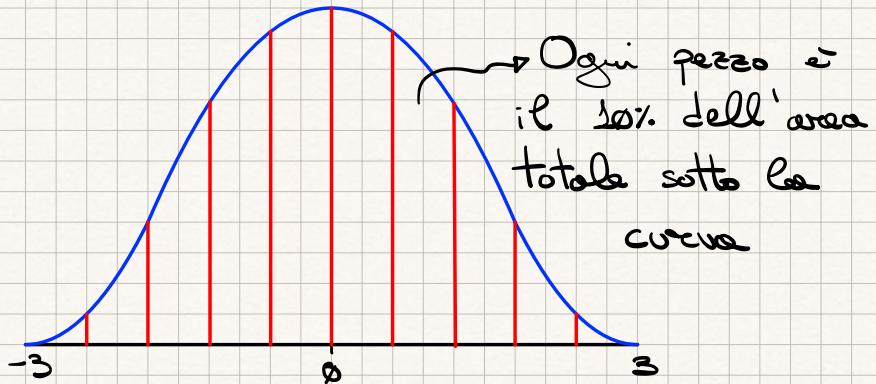
L'idea è quella di mettere nel grafico Q-Q la distri  
buzione dei dati reali contro quella dei dati generati  
dalla distribuzione normale, e se esce fuori un grafico  
più o meno lineare, allora possiamo concludere che la  
distribuzione dei dati è normale.

## Possi

1) Ordinare gli esempi presi dalla distribuzione normale (7.19, 6.31, 5.89, 4.50, 3.77, 4.25, 5.19, 5.79, 6.79):

(3.77, 4.25, 4.50, 5.19, 5.79, 5.89, 6.31, 6.79, 7.19)

2) Disegnare la distribuzione normale e dividerla in  $n+1$  parti, dove  $n$  è il numero dei nostri dati:



3) Trovare le z-value che rappresenta i quantili (cut-point) della distribuzione normale (le linee rosse nel grafico).

$$10\% = -1.28$$

$$20\% = -0.84$$

$$30\% = -0.52$$

$$40\% = -0.25$$

$$50\% = 0$$

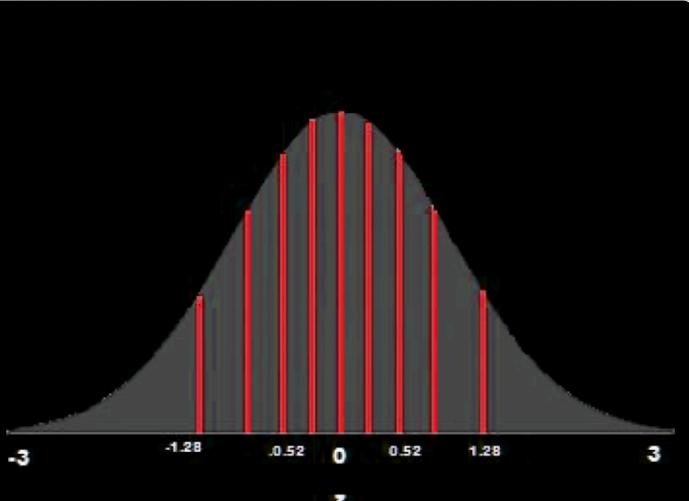
$$60\% = 0.25$$

$$70\% = 0.52$$

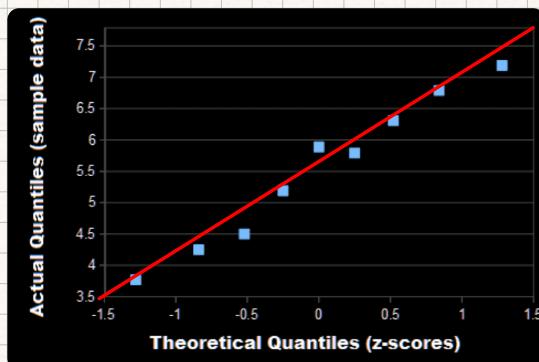
$$80\% = 0.84$$

$$90\% = 1.28$$

$$100\% = 3.0$$



4) Disegnare il Q-Q



### METODO ↴

Modellare i dati secondo la distribuzione normale

Usare la **Maximum Likelihood** per stimare  $\mu$  e  $\sigma^2$

La probabilità che un punto è generato dal modello gaussiano è dato:

$$P(x_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{\sigma^2}}$$

Quindi la probabilità totale, che considera tutti i punti  $x_i \in X$ :

$$\mathcal{L}(N(\mu, \sigma^2) : X) = P(X | \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{\sigma^2}}$$

Ora il prodotto delle singole probabilità.

Bisogna trovare i parametri  $\mu$  e  $\sigma^2$  tale per cui la probabilità sia massimizzata.

$$N(\mu, \sigma^2) = \operatorname{argmax} \{ \mathcal{L}(N(\mu, \sigma^2) : X) \}$$

Ora facendo il logaritmo naturale di  $\mathcal{L}(\cdot)$  possiamo dire che

$$\begin{aligned}\ln \mathcal{L}(\mu, \sigma^2) &= \sum \ln(f(x_i | \mu, \sigma^2)) = \\ &= -\frac{m}{2} \ln(2\pi) - \frac{m}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2\end{aligned}$$

Ma se deriviamo questa ultima formula prima per  $\mu$  poi per  $\sigma^2$  ricaviamo proprio le formule di media e deviazione standard

$$\frac{\partial \ln(\mathcal{L})}{\partial \mu} = \mu = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\frac{\partial \ln(\mathcal{L})}{\partial \sigma^2} = \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

Quindi possiamo approssimare la distribuzione dei dati alla gaussiana con  $\mu$  e  $\sigma^2$  calcolati dal dataset.

Dove sono gli outlier?

Avendo capito di poter approssimare il dataset alla distribuzione normale, possiamo dire che il 99.7% dei dati stanno nella regione  $\mu \pm 3\sigma$ , gli altri sono outlier.

## METODO 2

- 1) Ordinare i punti in ordine crescente:  $x_1 \dots x_m$
- 2) Trovare la media  $\mu$  e la deviazione standard  $\sigma^2$
- 3) Calcolare il GRUBB test  $G$ :

$$G = \frac{\max_{i=1 \dots m} |x_i - \mu|}{\sigma}$$

- 4) Trovare il valore critico di  $G$  in una delle tabelle dei valori critici divisi per numero di punti e livelli  $\alpha$  (è tipo il p-value) accettati
- 5) per  $G_{TEST} < G_{CRITICO}$  Non è un outlier.  
per  $G_{TEST} \geq G_{CRITICO}$  è un outlier.

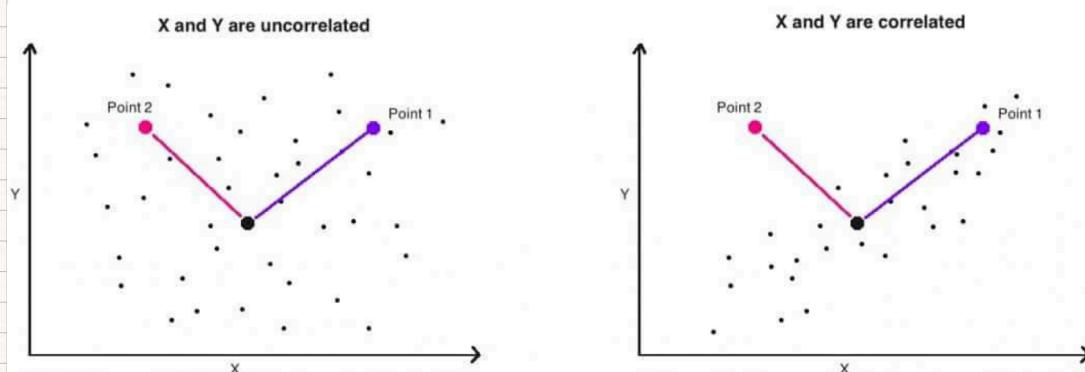
## MULTIVARIATE OUTLIER

Sono presenti in quei dataset che hanno più di un attributo. L'idea è trasformare il problema da multivariate a UNIVariante.

### METODO 1: MAHALANOBIS DISTANCE

Con mahalanobis distance si intende la distanza che separa un punto da una distribuzione probabilistica. È l'equivalente della distanza euclidea per spazi multivariabili, ovvero con più variabili aleatorie.

La semplice distanza euclidea fra un punto e il centro di una distribuzione di punti è spesso forzante, soprattutto se prendiamo in considerazione variabili dipendenti fra loro:



Differenze fra mahalanobis e euclideo:

- 1) Trasforma le colonne in variabili indipendenti.
- 2) Scala le colonne così la loro varianza è 1.
- 3) Solo dopo 1 e 2 calcola la distanza euclidea

## Definizione

$$MDist^2(o, \bar{o}) = (o - \bar{o})^T S^{-1} (o - \bar{o})$$

Dove  $\bar{o}$  è il vettore medio del dataset multivario.

Dove  $S$  è la matrice delle covarianze, questo perché dividere per la matrice  $S$  è la versione multivariante della normalizzazione usando lo z-score. Se il punto è correlato con la distribuzione allora  $MDist^2$  sarà basso perché alta sarà la covarianza.

## Passi

- 1) Calcolare il vettore medio  $\bar{o}$ .
- 2) Ho calcolare  $MDist(o, \bar{o})$
- 3) Cercare gli outlier nel nuovo dataset trasformato in uno univariante:

$$\{MDist(o, \bar{o}) \mid o \in D\}$$

- 4) Se la  $MDist(o, \bar{o})$  è troppo grande come outlier anche l'oggetto  $o$  lo è.

## Drawback

È computazionalmente pesante calcolare  $S$  e  $S^{-1}$ .  
 $S$  è grande per da essere salvata.

## METODO 2: $\chi^2$ - STATISTIC

ASSUNZIONI: La popolazione del dataset  $O$  è multivariata e ha il vettore medio  $\bar{o}$  e la matrice delle covarianze.

Usiamo una distanza basata sul  $\chi^2$ , usando la media dei valori al posto degli expected value  $\bar{\sigma}$ :

$$\chi^2 = \sum_{i=1}^n \frac{(x_i - \bar{x}_i)^2}{\bar{\sigma}_i}$$

Per il teorema del limite centrale per  $n$  molto grande (per esempio 30) allora il  $\chi^2$  segue la distribuzione normale.

### Rilevazione Outlier

Per rilevare gli outlier imponiamo un limite superiore, ma dato che  $\chi^2$  segue la normale in determinate condizioni, il limite superiore è  $\chi^2 + 3 \cdot S_{\chi^2}$ . (che è praticamente  $\chi^2 \pm 3 \cdot \sigma$ ).