Parte I

Statistica descrittiva

1. Popolazioni, individui e caratteri. Indicatori sintetici di campioni monovariati

La statistica descrittiva si occupa dell'analisi di dati raccolti da una popolazione, ovvero da un insieme di individui. In sintesi, dato un insieme molto grande di dati, così grande che non è *utile* guardarlo dato per dati, si cerca di estrarne delle informazioni sintetiche e tuttavia significative.

Gli oggetti con cui abbiamo a che fare sono dunque

- gli **individui** oggetto dell'indagine: ciascun individuo è un oggetto singolo dell'indagine.
- la **popolazione**, ovvero l'insieme degli individui oggetto dell'indagine.
- il **carattere** osservato o variabile, che è la quantità misurata o la qualità rilevata su ciascun individuo della popolazione.

Esempio 1.0.1. Rilevo l'altezza di ciascun abitante del Comune di Firenze. Ogni residente del Comune di Firenze è un individuo; la popolazione è l'insieme di tutti i residenti nel Comune di Firenze; il carattere in esame è l'altezza misurata, per esempio, in centimetri.

Esempio 1.0.2. Rilevo il reddito annuo di ciascun nucleo familiare del Comune di Firenze. Ogni nucleo familiare è un individuo; la popolazione è l'insieme dei nuclei familiari registrati all'Anagrafe del Comune di Firenze; il carattere osservato è il reddito annuo familiare misurato in Euro.

Esempio 1.0.3. Rilevo il numero dei componenti di ciascun nucleo familiare del Comunque di Firenze. Come nell'esempio precedente ogni nucleo familiare è un individuo; la popolazione è l'insieme dei nuclei familiari registrati all'Anagrafe del Comune di Firenze. Il carattere osservato è il numero dei componenti di ciascun nucleo familiare, cioè un numero intero maggiore-uguale di 1.

Esempio 1.0.4. Per ogni studente presente in aula rilevo il colore degli occhi. Ogni studente presente in aula è un individuo. La popolazione è l'insieme degli studenti presenti ed il carattere osservato è il colore degli occhi.

In questi esempi abbiamo incontrato i due tipi fondamentali di carattere:

- **caratteri numerici o quantitativi** come l'altezza, il reddito familiare, il numero dei componenti del nucleo familiare;

- **caratteri qualitativi** come il colore degli occhi.

I caratteri numerici a loro volta si possono suddividere in

- **caratteri numerici discreti** che possono assumere solo un insieme discreto di valori, come il numero dei componenti dei nuclei familiari;
- **caratteri numerici continui** che variano con continuità ovvero con una estrema accuratezza, eccessiva rispetto ai fini dell'indagine, come l'altezza delle persone o il reddito annuo familiare.

1.1. Campione statistico, modalità e classi modali

Supponiamo di aver osservato un certo carattere su una popolazione di n individui. Abbiamo un *vettore delle osservazioni*

$$x = (x_1, x_2, \dots, x_n)$$

che chiamiamo **campione statistico** di cardinalità n.

Se il campione è relativo ad un carattere qualitativo o numerico discreto, chiamo **modalità** i valori che esso assume su un campione.

Se il campione è relativo ad un carattere numerico continuo si procede nel seguente modo: la popolazione in esame è comunque un insieme finito, quindi il carattere, per quanto continuo, nel campione assume solo un numero finito di valori. Sia $[a, b)$ un intervallo che contiene tutti i valori x_i, $i = 1, \dots, n$ assunti dal carattere sugli individui della popolazione. Suddividiamo l'intervallo $[a, b)$ in N parti uguali (N sarà suggerito dall'esperienza). Otteniamo N intervalli

$$I_j := \left[a + (j-1)\frac{b-a}{N}, a + j\frac{b-a}{N} \right], \qquad j = 1, \dots, N.$$

Chiamo ciascuno di questi intervalli **classe di modalità**.

1.2. Frequenza assoluta e frequenza relativa

Consideriamo un campione $x = (x_1, x_2, \dots, x_n)$ relativo ad un carattere qualitativo o numerico discreto. Nel campione, cioè nella popolazione in esame, il carattere osservato assume un certo numero di valori distinti

$$z_1, z_2, \dots, z_k, \qquad 1 \le k \le n.$$

Per ogni $j = 1, \dots, k$ chiamo **effettivo** o **frequenza assoluta** della modalità z_j il numero

$$N_j := \#\{i \in \{1, \dots, n\} \colon x_i = z_j\}$$

mentre chiamo **frequenza relativa** della modalità z_j il numero

$$p_j := \frac{N_j}{n}.$$

Se il carattere osservato è numerico continuo, si considera ciascuna classe di modalità individuata

$$I_j := \left[a + (j-1)\frac{b-a}{N}, a + j\frac{b-a}{N}\right], \qquad j = 1, \dots, N$$

e si chiama **frequenza assoluta o effettivo** della classe di modalità I_j il numero

$$N_j := \#\{i \in \{1, \dots, n\} \colon x_i \in I_j\}.$$

Come prima definiamo **frequenza relativa** della classe I_j il numero $p_j := \dfrac{N_j}{n}$.

1.3. Moda e valori modali

Sia $x = (x_1, x_2, \dots, x_n)$ un campione statistico e siano $z_1, z_2, \dots, z_k$ le modalità assunte (o $I_1, I_2, \dots, I_k$ le classi di modalità assunte) e siano $p_1, p_2, \dots, p_k$ le relative frequenze relative.

Se esiste uno ed un solo indice $\overline{j} \in \{1, 2, \dots, k\}$ tale che la modalità $z_{\overline{j}}$ (o la classe $I_{\overline{j}}$) ha frequenza massima, ovvero se esiste un unico $\overline{j} \in \{1, 2, \dots, k\}$ tale che $p_{\overline{j}} \geq p_j$ $\forall j = 1, \dots, k$, allora la modalità $z_{\overline{j}}$ (o la classe $I_{\overline{j}}$) si dice **moda** del campione x.

Se esistono due o pi indici $\overline{j_1}, \overline{j_2}, \dots, \overline{j}_s$ tali che le modalità $z_{\overline{j_1}}, z_{\overline{j_2}}, \dots, z_{\overline{j}_s}$ (o le classi $I_{\overline{j_1}}, I_{\overline{j_2}}, \dots, I_{\overline{j}_s}$) hanno frequenza massima, allora queste modalità (o classi) si dicono **valori (o classi) modali**.

1.4. Mediana

D'ora innanzi consideriamo solo caratteri numerici.
Sia dunque $x = (x_1, x_2, \dots, x_n)$ un campione relativo ad un carattere numerico. Ordiniamo i dati del campione in ordine crescente:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}$$

e distinguiamo due casi:

- n dispari: $n = 2m + 1$

  $$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(m)} \leq x_{(m+1)} \leq x_{(m+2)} \leq \ldots \leq x_{(2m)} \leq x_{(2m+1)}$$

  Il dato $x_{(m+1)}$ è maggioreuguale di m dati e minoreuguale di altrettanti dati. Diciamo che il dato $x_{(m+1)}$ è la **mediana** del campione.

- n pari: $n = 2m$

  $$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(m-1)} \leq x_{(m)} \leq x_{(m+1)} \leq x_{(m+2)} \leq \ldots \leq x_{(2m)}$$

  Il dato $x_{(m)}$ è maggioreuguale di $m - 1$ dati e minoreuguale di m dati. Il dato $x_{(m+1)}$ è maggioreuguale di m dati e minoreuguale di $m - 1$ dati.

  Chiamiamo **mediana** del campione il numero $\dfrac{x_{(m)} + x_{(m+1)}}{2}$.

1.5. Media e varianza campionaria. Scarto quadratico medio (o deviazione standard)

Consideriamo un campione relativo ad un carattere numerico

$$x = (x_1, x_2, \dots, x_n)\,.$$

Chiamo **media aritmetica** o, pi semplicemente, **media** il numero

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Supponiamo che nel campione siano presnti k modalità $z_1, z_2, \dots, z_k$ con rispettive frequenze assolute $N_1, N_2, \dots N_k$ e frequenze relative $p_1, p_2, \dots p_k$. Allora

$$\begin{aligned}\overline{x} &= \frac{1}{n}\left(x_1 + x_2 + \dots x_n\right) = \frac{1}{n}\left(N_1 z_1 + N_2 z_2 + \dots N_k z_k\right) = \\ &= p_1 z_1 + p_2 z_2 + \dots p_k z_k = \sum_{j=1}^{k} p_j z_j.\end{aligned}$$

Chiamo **varianza campionaria** di x il numero nonnegativo

$$\sigma_x^2 = \mathrm{Var}\,[x] := \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2.$$

Osserviamo che la media è un valore centrale attorno al quale si dispongono i dati $x_1,\ x_2, \dots, x_n$ mentre la varianza è un *indice di dispersione*: la varianza è nulla se e solo se tutti i dati del campioni sono uguali (e dunque coincidono con la media). Una varianza bassa indica che comunque i dati sono *vicini* al valore medio $\overline{x}$ mentre una varianza alta indica una maggiore dispersione dei dati.

La radice quadrata della varianza campionaria

$$\sigma_x = \mathrm{Std}\,[x] := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

si chiama **scarto quadratico medio** o **deviazione standard** del campione x.

Anche per la varianza campionaria possiamo scrivere una formula che coinvolga solo le modalità e le rispettive frequenze.

$$\begin{aligned}\sigma_x^2 &= \frac{1}{n-1}\left((x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \dots (x_n - \overline{x})^2\right) = \\ &= \frac{1}{n-1}\left(N_1(z_1 - \overline{x})^2 + N_2(z_2 - \overline{x})^2 + \dots N_k(z_k - \overline{x})^2\right) = \\ &= \frac{n}{n-1}\left(p_1(z_1 - \overline{x})^2 + p_2(z_2 - \overline{x})^2 + \dots p_k(z_k - \overline{x})^2\right) = \\ &= \frac{n}{n-1}\sum_{j=1}^{k} p_j (z_j - \overline{x})^2.\end{aligned}$$

Esempio 1.5.1. Nella tabella che segue, tratta da [3], riportiamo alcuni dati relativi a campioni di laterizio e che useremo per fare alcuni esmpi relativi alle nozioni introdotte mediante il software R `http://cran.r-project.org/`. Per una introduzione si rimanda ai manuali [4] e [2].

Sample Code	Porosità totale (%)	Raggio medio del poro (μm)	Volume dei pori su dimensione dei pori 0.3–0.8 μm	Densità (g/cm^3)	Resistenza alla trazione (MPa)	CO_2/SBW	Temperatura di cottura (DTA)
AS1	41.460	0.528	80.0	1.550	0.403	0.38	740
AS2	47.210	0.467	81.2	1.650	0.645	0.70	740
AS3	43.670	0.697	78.5	1.710	0.527	0.46	740
AS4	52.390	0.422	77.3	1.520	0.143	0.48	740
AS5	44.700	0.411	87.4	1.500	0.593	0.29	740
AS6	51.330	0.422	88.6	1.480	0.463	0.33	740
AS7	31.460	0.718	80.6	1.900	0.955	0.23	740
AS8	40.900	0.458	80.4	1.680	0.195	0.41	740
AS9	45.540	0.492	80.8	1.620	1.328	0.50	750
AS10	45.620	0.734	86.2	1.620	1.405	0.34	750
AS11	44.140	0.730	85.7	1.590	0.256	0.42	750
AS12	40.710	0.543	87.8	1.750	0.309	0.20	750
AS13	35.700	0.686	84.3	1.520	0.472	0.05	740
C1	40.290	0.306	43.5	1.760	0.520	0.43	740
C2	36.570	0.625	42.3	1.750	0.738	0.36	740
C3	42.130	0.249	63.2	1.630	0.410	0.25	740
C4	37.830	0.731	47.9	2.020	0.601	0.28	740
C5	42.180	0.407	59.4	1.580	0.376	0.34	740
C6	41.600	0.446	42.8	1.850	0.473	0.26	740
C7	32.660	0.664	64.3	1.850	0.695	0.25	740
C8	36.070	0.673	58.2	1.780	0.624	0.29	740
C9	36.040	1.397	55.6	1.730	0.582	0.38	740
C10	36.640	0.861	45.2	1.750	0.650	0.47	740
R1	42.890	0.785	10.2	1.540	0.453	1.04	850
R2	26.850	0.315	14.7	2.010	1.124	1.86	960
R3	28.550	0.158	18.6	1.920	0.937	1.96	850
R4	29.860	0.158	15.3	1.890	1.020	1.48	850
R5	45.700	0.984	12.8	1.500	0.328	–	800
R6	54.640	1.525	12.5	1.340	0.267	0.67	750
R7	27.550	2.657	14.6	1.920	0.892	0.40	730
R8	40.820	0.622	15.3	1.570	0.502	1.94	860

Inseriamo la tabella in R

```
> table2 <-  read.table("table2.csv", header = TRUE)
> table2
   Code Totpor   PRA   PV Densi TenStr CO2SBW FirTemp
1   AS1  41.46 0.528 80.0  1.55  0.403   0.38     740
2   AS2  47.21 0.467 81.2  1.65  0.645   0.70     740
```

```
3    AS3  43.67 0.697 78.5  1.71  0.527   0.46     740
4    AS4  52.39 0.422 77.3  1.52  0.143   0.48     740
5    AS5  44.70 0.411 87.4  1.50  0.593   0.29     740
6    AS6  51.33 0.422 88.6  1.48  0.463   0.33     740
7    AS7  31.46 0.718 80.6  1.90  0.955   0.23     740
8    AS8  40.90 0.458 80.4  1.68  0.195   0.41     740
9    AS9  45.54 0.492 80.8  1.62  1.328   0.50     750
10 AS10  45.62 0.734 86.2  1.62  1.405   0.34     750
11 AS11  44.14 0.730 85.7  1.59  0.256   0.42     750
12 AS12  40.71 0.543 87.8  1.75  0.309   0.20     750
13 AS13  35.70 0.686 84.3  1.52  0.472   0.05     740
14   C1  40.29 0.306 43.5  1.76  0.520   0.43     740
15   C2  36.57 0.625 42.3  1.75  0.738   0.36     740
16   C3  42.13 0.249 63.2  1.63  0.410   0.25     740
17   C4  37.83 0.731 47.9  2.02  0.601   0.28     740
18   C5  42.18 0.407 59.4  1.58  0.376   0.34     740
19   C6  41.60 0.446 42.8  1.85  0.473   0.26     740
20   C7  32.66 0.664 64.3  1.85  0.695   0.25     740
21   C8  36.07 0.673 58.2  1.78  0.624   0.29     740
22   C9  36.04 1.397 55.6  1.73  0.582   0.38     740
23  C10  36.64 0.861 45.2  1.75  0.650   0.47     740
24   R1  42.89 0.785 10.2  1.54  0.453   1.04     850
25   R2  26.85 0.315 14.7  2.01  1.124   1.86     960
26   R3  28.55 0.158 18.6  1.92  0.937   1.96     850
27   R4  29.86 0.158 15.3  1.89  1.020   1.48     850
28   R5  45.70 0.984 12.8  1.50  0.328     --     800
29   R6  54.64 1.525 12.5  1.34  0.267   0.67     750
30   R7  27.55 2.657 14.6  1.92  0.892   0.40     730
31   R8  40.82 0.622 15.3  1.57  0.502   1.94     860
```

Per ciascun carattere definiamo una variabile che contenga la mediana, una per la media, una per la Varianza e una per la deviazione standard e poi stampiamo i valori (tratteremo il carattere di nome CO2SBW a parte perché su un individuo non è stato rilevato)

```
> medianaTotPor <- median(table2$Totpor);
> meanTotPor <- mean(table2$Totpor);
> VarTotPor <- var(table2$Totpor);
> StdTotPor <- sd(table2$Totpor)
> medianaTotPor; meanTotPor; VarTotPor; StdTotPor
[1] 40.9
[1] 40.11935
[1] 49.52185
[1] 7.037176
> medianaPRA <- median(table2$PRA);
> meanPRA <- mean(table2$PRA);
```

```
VarPRA <- var(table2$PRA);
> StdPRA <- sd(table2$PRA)
> medianaPRA; meanPRA; VarPRA; StdPRA
[1] 0.622
[1] 0.6732581
[1] 0.226613
[1] 0.4760389
> medianaPV <- median(table2$PV);
> meanPV <- mean(table2$PV);
> VarPV <- var(table2$PV);
> StdPV <- sd(table2$PV)
> medianaPV; meanPV; VarPV; StdPV
[1] 59.4
[1] 55.32903
[1] 815.0935
[1] 28.54984
> medianaDensi <- median(table2$Densi);
> meanDensi <- mean(table2$Densi);
> VarDensi <- var(table2$Densi);
> StdDensi <- sd(table2$Densi)
> medianaDensi; meanDensi; VarDensi; StdDensi
[1] 1.68
[1] 1.692903
[1] 0.02894129
[1] 0.1701214
> medianaTenStr <- median(table2$TenStr);
> meanTenStr <- mean(table2$TenStr);
> VarTenStr <- var(table2$TenStr);
> StdTenStr <- sd(table2$TenStr)
> medianaTenStr; meanTenStr; VarTenStr; StdTenStr
[1] 0.527
[1] 0.6092258
[1] 0.09882738
[1] 0.3143682
> medianaFirTemp <- median(table2$FirTemp);
> meanFirTemp <- mean(table2$FirTemp);
> VarFirTemp <- var(table2$FirTemp);
> StdFirTemp <- sd(table2$FirTemp)
> medianaFirTemp; meanFirTemp; VarFirTemp; StdFirTemp
[1] 740
[1] 764.8387
[1] 2805.806
[1] 52.96986
```

Introduciamo la tabella togliendo i dati relativi al campione R5 e calcoliamo la media del carattere

```
> table2noR5 <-  read.table("table2_noR5.csv", header = TRUE)
> medianaCO2SBW <- median(table2noR5$CO2SBW);
> meanCO2SBW <- mean(table2noR5$CO2SBW);
> VarCO2SBW <- var(table2noR5$CO2SBW);
> StdCO2SBW <- sd(table2noR5$CO2SBW)
> medianaCO2SBW; meanCO2SBW; VarCO2SBW;  StdCO2SBW
[1] 0.39
[1] 0.5816667
[1] 0.2765868
[1] 0.5259152
```

2. Campioni bivariati: covarianza, coefficiente di correlazione e retta di regressione

2.1. Covarianza e coefficiente di correlazione

Supponiamo di avere un **campione bivariato** cioè di rilevare due caratteri sugli individui di una medesima popolazione.

Abbiamo dunque due vettori di dati

$$x = (x_1, x_2, \dots x_n)\,, \qquad y = (y_1, y_2, \dots y_n)\,.$$

x_i e y_i sono le rilevazioni dei due caratteri sul medesimo individuo, l'individuo cioè che abbiamo etichettato come *individuo i*.

Chiamiamo **coverianza di x e y** il numero

$$\mathrm{Cov}\,[x,y] := \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$$

dove $\overline{x}$ e $\overline{y}$ sono le medie dei campioni x e y, rispettivamente.

Nel caso in cui né x né y siano campioni costanti (ipotesi lavorativa che sarà sempre sottointesa), definiamo **coefficiente di correlazione di x e y** il numero

$$\rho\,[x,y] := \frac{\mathrm{Cov}\,[x,y]}{\mathrm{Std}\,[x]\,\mathrm{Std}\,[y]} = \frac{\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\left(\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})^2\right)\left(\displaystyle\sum_{i=1}^{n}(y_i-\overline{y})^2\right)}}.$$

Osservazione 2.1.1. $\mathrm{Cov}\,[x,x] = \mathrm{Var}\,[x]$; $\rho\,[x,x] = 1$.

Si possono diostrare le seguenti proprietà:

1. $-1 \le \rho\,[x,y] \le 1$;
2. $\rho\,[x,y] = 1$ se e solo se esiste $a > 0$, $b \in \mathbb{R}$ tale che $y_i = ax_i + b \quad \forall i = 1, \dots, n$. In tal caso i campioni x e y si dicono *positivamente correlati*;
3. $\rho\,[x,y] = -1$ se e solo se esiste $a < 0$, $b \in \mathbb{R}$ tale che $y_i = ax_i + b \quad \forall i = 1, \dots, n$. In tal caso i campioni x e y si dicono *negativamente correlati*.

Se $\rho\,[x,y] = 0$ i campioni x e y si dicono *scorrelati*.

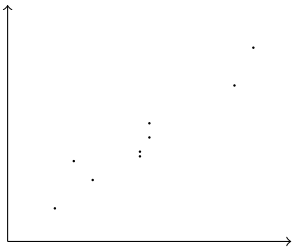

Figura 2.1: Campione bivariato

2.2. Retta di regressione

Supponiamo di avere un campione bivariato

$$x = (x_1, x_2, \dots x_n)\,, \qquad y = (y_1, y_2, \dots y_n)$$

dove x_i e y_i sono i dati relativi all'i–esimo individuo. Rappresentiamo i punti (x_i, y_i) sul piano cartesiano Oxy. Capita, molto spesso, di trovarsi a disposizioni *pressoché allineate* come illustrato nella figura 2.1 Si cerca allora una retta che in qualche senso *approssimi* i punti (x_i, y_i).

Supponiamo che $y = ax + b$ sia l'equazione della retta cercata. Per $x = x_i$ si ottiene il punto sulla retta (x_i, ax_i+b). Cerchiamo la retta (ovvero i parametri a e b) che minimizza la *somma degli errori quadratici*

$$S(a,b) := \sum_{i=1}^{n} (y_i - (ax_i + b))^2\,.$$

Si ha

$$\begin{aligned}
S(a,b) =& \sum_{i=1}^{n} (y_i - \overline{y} + \overline{y} - (ax_i - a\overline{x} + a\overline{x} + b))^2 = \\
=& \sum_{i=1}^{n} ((y_i - \overline{y}) - a\,(x_i - \overline{x}) + (\overline{y} - a\overline{x} - b))^2 = \\
=& \sum_{i=1}^{n} (y_i - \overline{y})^2 + a^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 + \\
& + n\,(\overline{y} - a\overline{x} - b)^2 - 2a \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y}) = \\
=&(n-1)\left(\operatorname{Var}[y] + a^2 \operatorname{Var}[x] - 2a \operatorname{Cov}[x,y]\right) + n\,(\overline{y} - a\overline{x} - b)^2\,.
\end{aligned}$$

L'incognita b compare solo nell'ultimo addendo, che è un quadrato. Quindi per ottenere il minimo basterà scegliere a che minimizza la funzione $f(a) := \operatorname{Var}[y] + a^2 \operatorname{Var}[x] -$

$2a\,\mathrm{Cov}\,[x,y]$ e poi scegliere $b=\overline{y}-a\overline{x}$. Si ha

$$f'(a)=2a\,\mathrm{Var}\,[x]-2\,\mathrm{Cov}\,[x,y]=0 \quad \text{se e solo se} \quad a=\frac{\mathrm{Cov}\,[x,y]}{\mathrm{Var}\,[x]}$$
$$f''(a)=2\,\mathrm{Var}\,[x]>0$$

Il minimo dello somma degli errori quadratici $S(a,b)$ si ottiene allora per

$$a=\frac{\mathrm{Cov}\,[x,y]}{\mathrm{Var}\,[x]}; \qquad b=\overline{y}-\frac{\mathrm{Cov}\,[x,y]}{\mathrm{Var}\,[x]}\overline{x};$$

il minimo dell'errore S vale

$$(n-1)\left(\mathrm{Var}\,[y]-\frac{(\mathrm{Cov}\,[x,y])^2}{\mathrm{Var}\,[x]}\right)=(n-1)\,\mathrm{Var}\,[y]\left(1-(\rho\,[x,y])^2\right)$$

e la retta ha equazione

$$y=\overline{y}+\frac{\mathrm{Cov}\,[x,y]}{\mathrm{Var}\,[x]}\,(x-\overline{x})\,.$$

Osservazione 2.2.1. La retta così determinata si chiama **retta di regressione del campione** y **sul campione** x. Osserviamo infine che il punto $(\overline{x},\overline{y})$ appartiene alla retta.

Esempio 2.2.1. Riconsideriamo l'esempio 1.5.1. Carichiamo in R la tabella dei dati.

```
> setwd("/home/laura/Documents/didattica/2012-13_elaborazioni_B194")

> X <-
+   read.table("/home/laura/Documents/didattica/2012-13_elaborazioni_B194/table2_noR5.csv",
+    header=TRUE, sep="\t", na.strings="NA", dec=".", strip.white=TRUE)
```

Tracciamo sul piano cartesiano i dati relativi ai caratteri porosità totale (in ascissa) e densità (in ordinata) e salviamo la figura in un file.

```
> scatterplot(Densi~Totpor, reg.line=FALSE, smooth=FALSE, spread=FALSE,
+   boxplots=FALSE, span=0.5, data=X)

> dev.print(png,
+   filename="/home/laura/Documents/didattica/2012-13_elaborazioni_B194/TotPorVSDensi.png",
+    width=500, height=500)
```

Sembrano *ragionevolmente allineati*. Calcoliamo il loro coefficiente di correlazione

```
> CorTotporDensi<- cor(X$Totpor, X$Densi)
> CorTotporDensi
[1] -0.8187597
```

Calcoliamo la retta di regressione del carattere Densità sul carattere Porosità Totale

```
> RegModel.Densi.Totpor <- lm(Densi~Totpor, data=X)

> summary(RegModel.Densi.Totpor)
```

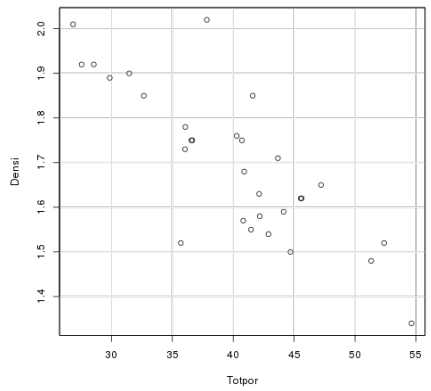

Figura 2.2: Porosità totale versus Densità

```
Call:
lm(formula = Densi ~ Totpor, data = X)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26174 -0.04070 -0.00072  0.05092  0.27972

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.476682   0.106138  23.335  < 2e-16 ***
Totpor      -0.019466   0.002618  -7.434 4.26e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09982 on 28 degrees of freedom
Multiple R-squared: 0.6637,Adjusted R-squared: 0.6517
F-statistic: 55.27 on 1 and 28 DF,  p-value: 4.264e-08
```

Intercept dice che l'ordinata all'origine (il coefficiente b) della retta di regressione è 2.476682 mentre il coefficiente angolare (cioè a) è -0.019466. Ridisegniamo i punti sul piano cartesiano, aggiungendo la retta di regressione (e salviamo l'immagine in un file).

```
> abline(lm(X$Densi ~ X$Totpor))
```

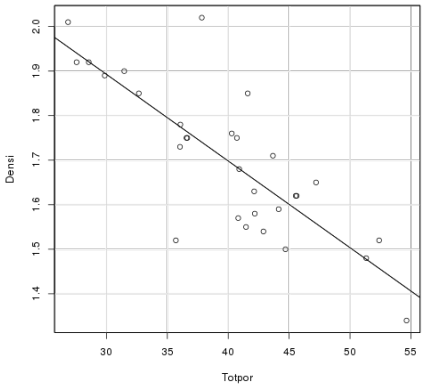

Figura 2.3: Retta di regressione lineare

3. Campioni multivariati. Principal Components Analysis

Lo scopo di questa analisi è il seguente: supponiamo di avere un campione multivariato. Supponiamo cioè di aver raccolto dati relativi a più caratteri, diciamo k caratteri, su una popolazione di n individui.

Riportiamo le informazioni raccolte come nella tabella dell'esempio 1.5.1. Ovvero

- Nella prima riga riportiamo i dati relativi al primo individuo, carattere per carattere

$$x_{11} \qquad x_{12} \qquad \dots \qquad x_{1k}$$

- Nella seconda riga riportiamo i dati relativi al secondo individuo, carattere per carattere

$$x_{21} \qquad x_{22} \qquad \dots \qquad x_{2k}$$

Procedendo di individuo in individuo otteniamo una matrice di n righe e k colonne:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} = (x_{ij})_{\substack{i=1,\dots,n \\ j=1,\dots,k}} \quad \in \mathbb{R}^{n\times k}$$

in cui il numero in posizione (i, j) (i-esima riga e j-esima colonna) è il dato rilevato sull'i-esimo indivduo relativamente al j-esimo carattere. Possiamo leggere la matrice colonna per colonna e rilevare le informazioni relative ad un singolo carattere. Infatti la prima colonna

$$X_1 := \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix}$$

contiene tutti i dati relativi al primo carattere, la seconda colonna

$$X_2 := \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix}$$

contiene tutti i dati relativi al secondo carattere e così via.

Per ogni $j = 1, \dots, k$ indichiamo, rispettivamente, con μ_j e σ_j la media e la deviazione standard del j-esimo carattere. Si ha

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \qquad \sigma_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_j)^2.$$

Possiamo anche calcolare la covarianza e il coefficiente di correlazione di due diversi caratteri. Più precisamente la covarianza del carattere j-esimo e del carattere ℓ-esimo è data da

$$\mathrm{Cov}\,[X_\ell, X_j] = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i\ell} - \mu_\ell)(x_{ij} - \mu_j).$$

Riportiamo varianze e covarianze in una matrice $k \times k$, detta **matrice di covarianza** del campione X:

$$C = \big(c_{\ell j}\big)_{\substack{\ell=1,\dots,k \\ j=1,\dots,k}} \in \mathbb{R}^{k\times k} \qquad c_{\ell j} := \mathrm{Cov}\,[X_\ell, X_j]\,, \qquad \ell, j = 1, \dots, k.$$

Poichè $\mathrm{Cov}\,[X_\ell, X_j] = \mathrm{Cov}\,[X_j, X_\ell]$ la matrice C è simmetrica. Inoltre gli elementi sulla diagonale principale sono le varianze dei caratteri in esame:

$$c_{jj} = \mathrm{Cov}\,[X_j, X_j] = \sigma_j^2 \quad \forall j = 1, \dots, k.$$

Supponiamo che i coefficienti di correlazione non siano prossimi a zero, indicando dunque che i caratteri in esame sono legati gli uni agli altri.

Cerchiamo di ridurre il numero di caratteri da osservare sostituendo i caratteri originari con delle loro combinazioni lineari, in modo che i nuovi caratteri siano a due a due scorrelati e la *variabilità* del campione sia concentrata in pochi caratteri. La procedura si compone di due passi. Il primo passo consiste nel rendere le variabili adimensionali (in modo che abbia senso sommarle) e *centrate* (cioè a media nulla).

Primo passo: Standardizzazione del campione
Per ogni $i = 1, \dots, n$ e ogni $j = 1, \dots, k$ pongo

$$y_{ij} := \frac{x_{ij} - \mu_j}{\sigma_j}.$$

Ovvero il dato relativo a ciascun carattere X_j è stato sostituito da

$$Y_1 = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n1} \end{pmatrix}, \qquad Y_2 = \begin{pmatrix} y_{12} \\ y_{22} \\ \vdots \\ y_{n2} \end{pmatrix}, \qquad \dots, \qquad Y_k = \begin{pmatrix} y_{1k} \\ y_{2k} \\ \vdots \\ y_{nk} \end{pmatrix}$$

In che senso i dati Y_j sono standardizzati? Calcoliamone media e varianza

$$\overline{Y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij} = \frac{1}{n}\sum_{i=1}^{n} \frac{x_{ij}-\mu_j}{\sigma_j} = \frac{1}{n\,\sigma_j}\left(\sum_{i=1}^{n} x_{ij} - \sum_{i=1}^{n}\mu_j\right) = \frac{1}{\sigma_j}\left(\mu_j - \mu_j\right) = 0$$

$$\begin{aligned}\mathrm{Var}\left[Y_j\right] &= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_{ij}-\overline{Y}_j\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n} y_{ij}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\frac{\left(x_{ij}-\mu_j\right)^2}{\sigma_j^2} = \\ &= \frac{1}{\sigma_j^2}\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}-\mu_j\right)^2 = \frac{1}{\sigma_j^2}\sigma_j^2 = 1.\end{aligned}$$

Secondo passo: Scorrelazione dei caratteri
Innanzitutto calcoliamo la matrice di covarianza $C = \left(c_{\ell j}\right)$ del campione standardizzato Y. Si ha

$$c_{\ell j} = \mathrm{Cov}\left[Y_\ell, Y_j\right] = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_{i\ell}-\overline{y}_\ell\right)\left(y_{ij}-\overline{y}_j\right) = \frac{1}{n-1}\sum_{i=1}^{n} y_{i\ell}y_{ij}$$

ovvero, in termini di matrici

$$C = \frac{1}{n-1}Y^t\,Y. \tag{3.1}$$

Osservazione 3.0.2. La formula (3.1) è vera tutte le volte che i campioni in esame hanno media nulla.

Se vogliamo calcolare i coefficienti di C in termini del campione X otteniamo anche

$$\begin{aligned}c_{\ell j} = \mathrm{Cov}\left[Y_\ell, Y_j\right] &= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_{i\ell}-\overline{y}_\ell\right)\left(y_{ij}-\overline{y}_j\right) = \\ &= \frac{1}{n-1}\sum_{i=1}^{n} y_{i\ell}y_{ij} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_{i\ell}-\mu_\ell}{\sigma_\ell}\frac{x_{ij}-\mu_j}{\sigma_j} = \\ &= \frac{1}{\sigma_\ell\sigma_j}\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i\ell}-\mu_\ell\right)\left(x_{ij}-\mu_j\right) = \rho\left[X_\ell, X_j\right].\end{aligned}$$

La matrice di covarianza del campione standardizzato Y è dunque diagonale e gli elementi diagonali $c_{\ell\ell}$ sono tutti uguali ad 1.

Sull'i-esimo indivduo abbiamo i dati standardizzati $y_{i1},\ y_{i2},\ \ldots,\ y_{ik}$. Vogliamo sostituirli con $z_{i1},\ z_{i2},\ \ldots,\ z_{ik}$,

$$\begin{aligned} z_{i1} &= \sum_{j=1}^{k} y_{ij} a_{j1} \\ z_{i2} &= \sum_{j=1}^{k} y_{ij} a_{j2} \\ &\vdots \\ z_{ik} &= \sum_{j=1}^{k} y_{ij} a_{jk} \end{aligned}$$

ovvero vogliamo sostituire la matrice Y con una matrice $Z = YA$ in modo che

- la matrice $A \in \mathbb{R}^{k\times k}$ rappresenti una rotazione nello spazio a k dimensioni, ovvero vogliamo che A sia una matrice ortogonale: $A^t = A^{-1}$;
- i nuovi campioni $Z_1,\ Z_2,\ \ldots,\ Z_k$ siano scorrelati, ovvero richiediamo

  $$\mathrm{Cov}\left[Z_\ell, Z_j\right] = 0 \quad \forall \ell, j = 1, \ldots k, \quad \ell \neq j.$$

  Equivalentemente, richiediamo che la matrice di covarianza del campione Z, C_Z, sia una matrice diagonale.

Per calcolare matrice C_Z osserviamo preliminarmente che anche $Z_1,\ Z_2,\ \ldots,\ Z_k$ hanno media nulla:

$$\overline{Z}_\ell = \frac{1}{n}\sum_{i=1}^{n} z_{i\ell} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} a_{j\ell} = \sum_{j=1}^{k} a_{j\ell}\frac{1}{n}\sum_{i=1}^{n} y_{ij} = \sum_{j=1}^{k} a_{j\ell}\overline{y}_j = 0$$

visto che $\overline{y}_j = 0$ per ogni $j = 1, \ldots, k$.

Dunque

$$\begin{aligned} C_Z &= \frac{1}{n-1} Z^t Z = \frac{1}{n-1}\left(YA\right)^t \left(YA\right) = \frac{1}{n-1} A^t Y^t Y A = \\ &= A^t \left(\frac{1}{n-1} Y^t Y\right) A = A^t C_Y A \end{aligned}$$

Esiste A matrice ortogonale in modo che C_Z sia una matrice diagonale? Sì, un importante risultato di algebra lineare, il *Teorema spettrale* dice che questa matrice A esiste. Più precisamente

Teorema 3.0.1 (Teorema spettrale)**.** *Data $C \in \mathbb{R}^{k\times k}$ matrice simmetrica esiste $A \in \mathbb{R}^{k\times k}$ matrice ortogonale tale che $A^t CA$ è una matrice diagonale*

$$A^t CA = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & \dots & 0 & \lambda_k \end{pmatrix}.$$

Le colonne A_1, A_2, …, A_k della matrice A sono gli autovettori di C e gli elementi diagonali $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ sono i rispettivi autovalori cioè $CA_j = \lambda_j A_j \quad \forall j = 1,\dots,k$. Inoltre λ_1 è il massimo della funzione $f(X) := \dfrac{X^t CX}{X^t X}$ e A_1 è un punto di massimo.

Gli diagonali $\lambda_1,\ \lambda_2,\ \dots,\ \lambda_k$ della matrice $C_Z = A^t C_Y A$ sono, per costruzione della matrice delle covarianze, le varianze dei *nuovi* campioni $Z_1,\ Z_2,\ \dots,\ Z_k$.

Esempio 3.0.2. Ritorniamo all'esempio tratto da [3]. Carichiamo la tabella a cui abbiamo tolto l'individuo R5. e visualizziamo in una *matrice di grafici*, salvando poi l'immagine

```
> setwd("/home/laura/Documents/didattica/2012-13_elaborazioni_B194")
> X <-  read.table("table2_noR5.csv", header = TRUE)
> plot(X)
> dev.copy(png,'bricks_pca_var.png');dev.off()
png
  3
X11cairo
       2
```

Calcoliamo la matrice dei coefficienti di correlazione (che abbiamo visto essere la matrice di covarianza del campione standardizzato), con i coefficienti approssimati alla tre cifre decimali.

```
> MatrixCorr <- round(cor(table2noR5), 3)
> MatrixCorr
        Totpor    PRA     PV  Densi TenStr CO2SBW FirTemp
Totpor   1.000 -0.116  0.411 -0.815 -0.461 -0.318  -0.398
PRA     -0.116  1.000 -0.268  0.017  0.024 -0.211  -0.258
PV       0.411 -0.268  1.000 -0.324 -0.162 -0.671  -0.624
Densi   -0.815  0.017 -0.324  1.000  0.467  0.217   0.277
TenStr  -0.461  0.024 -0.162  0.467  1.000  0.289   0.328
CO2SBW  -0.318 -0.211 -0.671  0.217  0.289  1.000   0.906
FirTemp -0.398 -0.258 -0.624  0.277  0.328  0.906   1.000
```

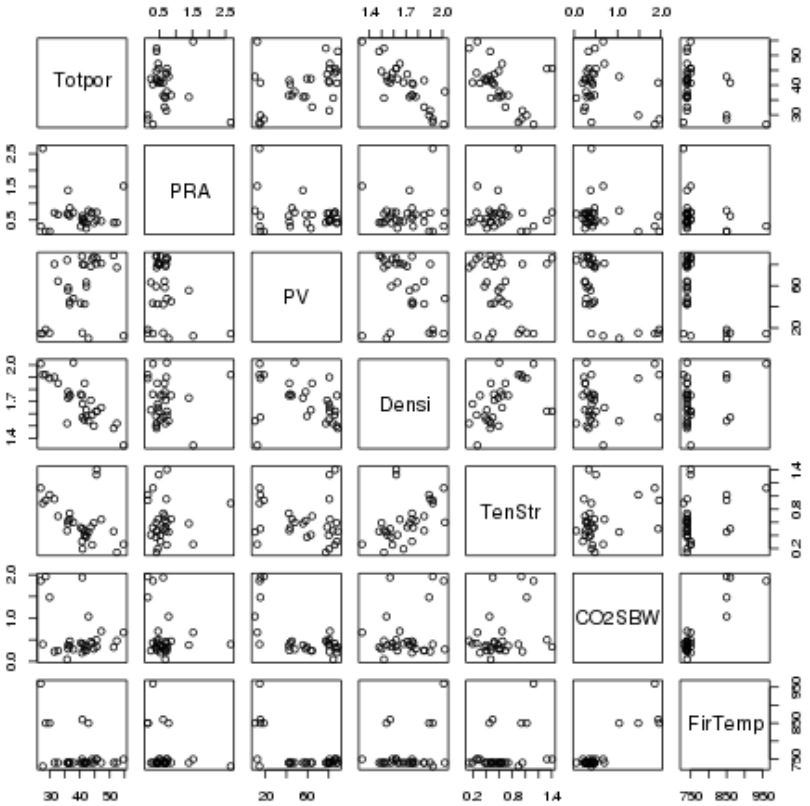

Figura 3.1: Plot dei caratteri, due a due

Visualizziamo i dati normalizzati (arrotondati a tre cifre decimali) e li salviamo in un file

```
>  Y <- round(scale(X[,c("CO2SBW","Densi","FirTemp","PRA","PV", "TenStr", "Totpor")]), 3)

> Y

      CO2SBW  Densi FirTemp    PRA     PV TenStr Totpor
 [1,] -0.383 -0.883  -0.443 -0.281  0.833 -0.684  0.216
 [2,]  0.225 -0.292  -0.443 -0.408  0.876  0.084  1.028
 [3,] -0.231  0.063  -0.443  0.071  0.780 -0.291  0.528
 [4,] -0.193 -1.060  -0.443 -0.501  0.737 -1.508  1.760
 [5,] -0.555 -1.178  -0.443 -0.524  1.098 -0.081  0.673
 [6,] -0.479 -1.297  -0.443 -0.501  1.141 -0.493  1.610
 [7,] -0.669  1.186  -0.443  0.115  0.855  1.067 -1.197
 [8,] -0.326 -0.114  -0.443 -0.426  0.848 -1.343  0.137
 [9,] -0.155 -0.469  -0.256 -0.356  0.862  2.250  0.792
```

```
[10,] -0.460 -0.469  -0.256  0.148  1.055  2.494  0.803
[11,] -0.307 -0.646  -0.256  0.140  1.038 -1.150  0.594
[12,] -0.726  0.300  -0.256 -0.249  1.113 -0.982  0.110
[13,] -1.011 -1.060  -0.443  0.048  0.987 -0.465 -0.598
[14,] -0.288  0.359  -0.443 -0.743 -0.475 -0.313  0.050
[15,] -0.421  0.300  -0.443 -0.079 -0.518  0.379 -0.475
[16,] -0.631 -0.410  -0.443 -0.861  0.231 -0.662  0.310
[17,] -0.574  1.896  -0.443  0.142 -0.317 -0.056 -0.297
[18,] -0.460 -0.705  -0.443 -0.532  0.095 -0.769  0.317
[19,] -0.612  0.891  -0.443 -0.451 -0.500 -0.462  0.235
[20,] -0.631  0.891  -0.443  0.002  0.271  0.242 -1.027
[21,] -0.555  0.477  -0.443  0.021  0.052  0.017 -0.546
[22,] -0.383  0.181  -0.443  1.527 -0.041 -0.116 -0.550
[23,] -0.212  0.300  -0.443  0.412 -0.414  0.100 -0.465
[24,]  0.871 -0.942   1.615  0.254 -1.668 -0.525  0.418
[25,]  2.431  1.837   3.672 -0.724 -1.507  1.603 -1.848
[26,]  2.621  1.305   1.615 -1.051 -1.367  1.010 -1.608
[27,]  1.708  1.127   1.615 -1.051 -1.485  1.273 -1.423
[28,]  0.168 -2.124  -0.256  1.794 -1.586 -1.115  2.077
[29,] -0.345  1.305  -0.630  4.149 -1.510  0.867 -1.749
[30,]  2.583 -0.765   1.802 -0.085 -1.485 -0.370  0.125
attr(,"scaled:center")
     CO2SBW       Densi     FirTemp         PRA          PV      TenStr
  0.5816667   1.6993333 763.6666667   0.6629000  56.7466667   0.6186000
     Totpor
 39.9333333
attr(,"scaled:scale")
    CO2SBW      Densi    FirTemp         PRA          PV     TenStr     Totpor
 0.5259152  0.1691548 53.4649955  0.4806106 27.9061201  0.3153048  7.0795326
```

```
> write.table(Y, "/home/laura/Documents/didattica/2012-13_elaborazioni_B194/normalizzate.csv",
+  sep="\t", col.names=TRUE, row.names=TRUE, quote=TRUE)
```

Infine facciamo calcolare la matrice A (la matrice Rotation) e stampare un sommario

```
> bricks.PC <- princomp(~CO2SBW+Densi+FirTemp+PRA+PV+TenStr+Totpor, cor=TRUE,
+   data=X)

> bricks.PC
Call:
princomp(formula = ~CO2SBW + Densi + FirTemp + PRA + PV + TenStr +
    Totpor, data = X, cor = TRUE)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7
1.8037550 1.2270089 1.0671061 0.8072493 0.4753775 0.3753498 0.2892731
```

```
 7  variables and  30 observations.

> unclass(loadings(bricks.PC))  # component loadings
            Comp.1     Comp.2      Comp.3      Comp.4     Comp.5     Comp.6
CO2SBW   0.44597862 -0.4327470  0.07954887 -0.08714546 -0.1416903  0.3163847
Densi    0.37599285  0.4539186 -0.24367883  0.36533155  0.3559843  0.5724149
FirTemp  0.46180279 -0.3933811 -0.01070518 -0.04944464 -0.3480988  0.1633589
PRA     -0.01780654  0.4429432  0.74323926 -0.24937193 -0.2964313  0.3160646
PV      -0.41158535  0.1139636 -0.52677278 -0.12259468 -0.5921070  0.4114405
TenStr   0.31740811  0.2746593 -0.30365291 -0.82533066  0.1765009 -0.1388565
Totpor  -0.41952730 -0.4090418  0.10990336 -0.31322074  0.5122611  0.5070450
             Comp.7
CO2SBW   0.692629300
Densi   -0.073192547
FirTemp -0.693953321
PRA     -0.033540667
PV       0.072227697
TenStr  -0.002205038
Totpor  -0.164285158

> round(bricks.PC$sd^2, 3)  # component variances
Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
 3.254  1.506  1.139  0.652  0.226  0.141  0.084
> summary(bricks.PC) # proportions of variance
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4    Comp.5
Standard deviation     1.8037550 1.2270089 1.0671061 0.80724932 0.4753775
Proportion of Variance 0.4647903 0.2150787 0.1626736 0.09309307 0.0322834
Cumulative Proportion  0.4647903 0.6798690 0.8425426 0.93563570 0.9679191
                           Comp.6     Comp.7
Standard deviation     0.37534975 0.28927306
Proportion of Variance 0.02012678 0.01195413
Cumulative Proportion  0.98804587 1.00000000

> screeplot(bricks.PC)

> dev.copy(png,'bricks_pca_var_comp.png');dev.off()
png
  3
X11cairo
       2
```

Vediamo come leggere questo output. Dalla prima riga del `summary` vediamo che le componenti principale sono numerate in ordine di deviazione standard decrescente. La prima componente principale ha la deviazione standard massima.

Dalla prima colonna della matrice `Rotation` abbiamo che la prima componente prin-

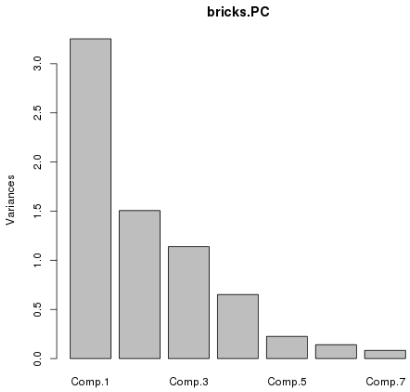

Figura 3.2: Varianza delle componenti principali

cipale Z_1, che qui è indicata con `Comp.1` è pari a

$$\begin{aligned} Z_1 = &- 0.41952730 \cdot \texttt{Totpor}_\texttt{s} - 0.01780654 \cdot \texttt{PRA}_\texttt{s} \\ &- 0.41158535 \cdot \texttt{PV}_\texttt{s} + 0.37599285 \cdot \texttt{Densi}_\texttt{s} \\ &+ 0.31740811 \cdot \texttt{TenStr}_\texttt{s} + 0.44597862 \cdot \texttt{CO2SBW}_\texttt{s} \\ &+ 0.46180279 \cdot \texttt{FirTemp}_\texttt{s} \end{aligned}$$

dove il pedice `s` indica che dobbiamo prendere il dato standardizzato e non nella sua forma originale. Possiamo ottenere la stessa informazione anche scrivendo

```
> Z1 <- bricks_pca$rotation[,1]
> Z1
```

da cui otteniamo

```
Totpor          PRA           PV       Densi      TenStr      CO2SBW
-0.41952730 -0.01780654 -0.41158535  0.37599285  0.31740811  0.44597862
    FirTemp
 0.46180279
```

Possiamo anche visualizzare (approssiamiamo a 3 cifre decimali) il valore della prima componente principale su ciascun individuo del campione (numerati da 1 a 30)

```
> round(predict(bricks_pca)[,1], 3)
 [1] -1.353 -0.972 -0.920 -2.200 -1.645 -2.198  0.430 -1.218 -0.330 -0.482
[11] -1.542 -1.141 -1.358 -0.110  0.255 -1.060  0.487 -1.081 -0.174  0.245
[21] -0.060 -0.124  0.204  1.120  5.388  3.982  3.562 -1.446  1.602  2.139
```

4. Analisi dei cluster

La parola *cluster* significa gruppo, agglomerato. L'analisi dei cluster consiste appunto nel raggruppare gli individui di una popolazione in base ad un qualche *criterio di vicinanza*.

Definiamo innanzitutto la nozione di **distanza tra due individui**.

4.1. Distanza tra individui

Consideriamo due individui della popolazione, li indichiamo con x e y. Su ciascuno di essi abbiamo i dati relativi a k caratteri e di averli *standardizzati*:

$$x = \big(x_1,\ x_2,\ \ldots,\ x_k\big) \qquad y = \big(y_1,\ y_2,\ \ldots,\ y_k\big) \qquad \text{dati standardizzati.}$$

Si possono definire varie nozioni di distanza tra i due individui x e y.

- **Distanza euclidea**
$$d_2(x,y) := \sqrt{\sum_{j=1}^{k} (x_j - y_j)^2}$$
- **Distanza Manhattan**
$$d_1(x,y) := \sum_{j=1}^{k} |x_j - y_j|$$
- **Distanza di Chebyshev**
$$d_\infty(x,y) := \max_{j=1,\ldots,k} |x_j - y_j|$$
- **p-distanza,** $p \in [1, +\infty)$
$$d_p(x,y) := \left(\sum_{j=1}^{k} |x_j - y_j|^p \right)^{1/p}$$
- **p-distanza pesata,** $p \in [1, +\infty)$**,** $w = \big(w_1, \ldots, w_k\big)$**,** $w_j \geq 0 \quad \forall j = 1, \ldots, k$
$$d_p(x,y) := \left(\sum_{j=1}^{k} w_j\, |x_j - y_j|^p \right)^{1/p}$$

Osservazione 4.1.1. La p distanza con $p = 2$ coincide con la distanza euclidea, mentre la p-distanza con $p = 1$ coincide con la distanza Manhattan.
La p-distanza pesata con peso $w = \big(1, \dots, 1\big)$ coincide con la p-distanza.

Supponiamo di avere scelto una nozione di distanza $\mathrm{dist}\,(\cdot, \cdot)$ tra gli individui della popolazione.

I metodi di *clustering* ovvero di raggruppamento degli individui della popolazione si suddividono in

- **Clustering partizionale**: si fissa a priori il numero k di gruppi che si vuole ottenere.
- **Clustering gerarchico**: si produce una rappresentazione gerarchica ad albero (o **dendrogramma**). I metodi gerarchici a loro volta si suddividono in due categorie fondamentali
  - **Metodi gerarchici aggregativi** Al passo iniziale ciascun individuo è un gruppo a sé stante, poi si procede per aggregazioni successive dei cluster, in base ad un *criterio di vicinanza* tra cluster.
  - **Metodi gerarchici disgiuntivi** Al passo iniziale c'è un unico cluster che contiene l'intera popolazione. Poi si procede per divisioni successive dei cluster, in base ad un *criterio di vicinanza* tra cluster.

4.2. Clustering partizionale

Ad ogni sottoinsieme G della popolazione si associa un costo $E(G)$. Si ripartisce la popolazione in k sottoinsiemi (i cluster) $G_1, \ \dots, \ G_k$ (questo vuol dire che ciascun individuo della popolazione appartiene ad uno e ad un solo sottoinsieme e nessun sottoinsieme è l'insieme vuoto) e ad ogni partizione $\mathcal{G} = \{G_1, \ \dots, \ G_k\}$ si associa il costo

$$\mathcal{E}(\mathcal{G}) := \sum_{s=1}^{k} E(G_s)$$

Osservazione 4.2.1. Data una popolazione di n individui è possibile calcolare il numero di partizioni possibili della popolazione in k cluster. Se non si tiene conto dell'ordine dei cluster, questo numero è

$$\frac{1}{k!} \sum_{j=0}^{k-1} (-1)^j \binom{k}{j} (k-j)^n$$

Un esempio di costo è il k-means clustering. Assegniamo a ciascun cluster G_s, $s = 1, \dots, k$, un costo $E(G_s) = \displaystyle\sum_{x \in G_s} d_2(x, \mu_s)$ dove $\mu_s := \dfrac{1}{|G_s|} \displaystyle\sum_{x \in G_s} x$ è il baricentro di G_s.

Esempio 4.2.1. Supponiamo di aver rilevato s caratteri su una popolazione di n individui. Riportiamo i dati in una matrice $X = \big(x_{ij}\big)_{\substack{i=1\dots,n \\ j=1,\dots s}} \in \mathbb{R}^{n \times s}$ e supponiamo di voler

suddividerla in due clusters secondo il criterio della media. Per ogni possibile sottoinsieme $G \subset \{1, \ldots, n\}$ il baricentro μ_G è un vettore di dimensione s, $\mu_G = \big(\mu_{G,1},\ \mu_{G,2}, \ldots\ \mu_{G,s}\big)$, dove $\mu_{G,j} = \dfrac{1}{|G|} \displaystyle\sum_{i \in G} x_{ij}$ per ogni $j = 1, \ldots, k$. Per ogni sottoinsieme G si calcolano dunque

$$E(G) = \sum_{i \in G} d_2(x_i, \mu_G) = \sum_{i \in G} \sqrt{\sum_{j=1}^{s} (x_{ij} - \mu_{G,j})^2}$$

Il costo della partizione $\mathcal{G}$ della popolazione $\{1,\ 2,\ \ldots,\ n\}$ nei cluster G e $H := \{1,\ 2,\ \ldots,\ n\} \setminus G$ è allora

$$\mathcal{E}(\mathcal{G}) := E(G) + E(H)$$

Ovviamente esistono dei software che effettuano questi calcoli. Torniamo al nostro campione tratto da [3]. Carichiamo la matrice dei dati e la standardizziamo, visualizzando la standardizzazione dove i caratteri sono ora in ordine alfabetico e chiediamo di dividere in 2 cluster secondo il criterio k-means

```
> setwd("/home/laura/Documents/didattica/2012-13_elaborazioni_B194")
```

```
> X <-
+   read.table("/home/laura/Documents/didattica/2012-13_elaborazioni_B194/table2_noR5.csv",
+    header=TRUE, sep="\t", na.strings="NA", dec=".", strip.white=TRUE)
```

```
> X
   Totpor   PRA   PV Densi TenStr CO2SBW FirTemp
1   41.46 0.528 80.0  1.55  0.403   0.38     740
2   47.21 0.467 81.2  1.65  0.645   0.70     740
3   43.67 0.697 78.5  1.71  0.527   0.46     740
4   52.39 0.422 77.3  1.52  0.143   0.48     740
5   44.70 0.411 87.4  1.50  0.593   0.29     740
6   51.33 0.422 88.6  1.48  0.463   0.33     740
7   31.46 0.718 80.6  1.90  0.955   0.23     740
8   40.90 0.458 80.4  1.68  0.195   0.41     740
9   45.54 0.492 80.8  1.62  1.328   0.50     750
10  45.62 0.734 86.2  1.62  1.405   0.34     750
11  44.14 0.730 85.7  1.59  0.256   0.42     750
12  40.71 0.543 87.8  1.75  0.309   0.20     750
13  35.70 0.686 84.3  1.52  0.472   0.05     740
14  40.29 0.306 43.5  1.76  0.520   0.43     740
15  36.57 0.625 42.3  1.75  0.738   0.36     740
16  42.13 0.249 63.2  1.63  0.410   0.25     740
17  37.83 0.731 47.9  2.02  0.601   0.28     740
18  42.18 0.407 59.4  1.58  0.376   0.34     740
19  41.60 0.446 42.8  1.85  0.473   0.26     740
20  32.66 0.664 64.3  1.85  0.695   0.25     740
21  36.07 0.673 58.2  1.78  0.624   0.29     740
```

```
22  36.04 1.397 55.6  1.73  0.582   0.38     740
23  36.64 0.861 45.2  1.75  0.650   0.47     740
24  42.89 0.785 10.2  1.54  0.453   1.04     850
25  26.85 0.315 14.7  2.01  1.124   1.86     960
26  28.55 0.158 18.6  1.92  0.937   1.96     850
27  29.86 0.158 15.3  1.89  1.020   1.48     850
28  54.64 1.525 12.5  1.34  0.267   0.67     750
29  27.55 2.657 14.6  1.92  0.892   0.40     730
30  40.82 0.622 15.3  1.57  0.502   1.94     860
```

```
> Y <- round(scale(X[,c("CO2SBW","Densi","FirTemp","PRA","PV",
+ "TenStr", "Totpor")]), 3)
```

```
> Y
      CO2SBW  Densi FirTemp    PRA     PV TenStr Totpor
 [1,] -0.383 -0.883  -0.443 -0.281  0.833 -0.684  0.216
 [2,]  0.225 -0.292  -0.443 -0.408  0.876  0.084  1.028
 [3,] -0.231  0.063  -0.443  0.071  0.780 -0.291  0.528
 [4,] -0.193 -1.060  -0.443 -0.501  0.737 -1.508  1.760
 [5,] -0.555 -1.178  -0.443 -0.524  1.098 -0.081  0.673
 [6,] -0.479 -1.297  -0.443 -0.501  1.141 -0.493  1.610
 [7,] -0.669  1.186  -0.443  0.115  0.855  1.067 -1.197
 [8,] -0.326 -0.114  -0.443 -0.426  0.848 -1.343  0.137
 [9,] -0.155 -0.469  -0.256 -0.356  0.862  2.250  0.792
[10,] -0.460 -0.469  -0.256  0.148  1.055  2.494  0.803
[11,] -0.307 -0.646  -0.256  0.140  1.038 -1.150  0.594
[12,] -0.726  0.300  -0.256 -0.249  1.113 -0.982  0.110
[13,] -1.011 -1.060  -0.443  0.048  0.987 -0.465 -0.598
[14,] -0.288  0.359  -0.443 -0.743 -0.475 -0.313  0.050
[15,] -0.421  0.300  -0.443 -0.079 -0.518  0.379 -0.475
[16,] -0.631 -0.410  -0.443 -0.861  0.231 -0.662  0.310
[17,] -0.574  1.896  -0.443  0.142 -0.317 -0.056 -0.297
[18,] -0.460 -0.705  -0.443 -0.532  0.095 -0.769  0.317
[19,] -0.612  0.891  -0.443 -0.451 -0.500 -0.462  0.235
[20,] -0.631  0.891  -0.443  0.002  0.271  0.242 -1.027
[21,] -0.555  0.477  -0.443  0.021  0.052  0.017 -0.546
[22,] -0.383  0.181  -0.443  1.527 -0.041 -0.116 -0.550
[23,] -0.212  0.300  -0.443  0.412 -0.414  0.100 -0.465
[24,]  0.871 -0.942   1.615  0.254 -1.668 -0.525  0.418
[25,]  2.431  1.837   3.672 -0.724 -1.507  1.603 -1.848
[26,]  2.621  1.305   1.615 -1.051 -1.367  1.010 -1.608
[27,]  1.708  1.127   1.615 -1.051 -1.485  1.273 -1.423
[28,]  0.168 -2.124  -0.256  1.794 -1.586 -1.115  2.077
[29,] -0.345  1.305  -0.630  4.149 -1.510  0.867 -1.749
[30,]  2.583 -0.765   1.802 -0.085 -1.485 -0.370  0.125
attr(,"scaled:center")
     CO2SBW       Densi     FirTemp         PRA          PV      TenStr
  0.5816667   1.6993333 763.6666667   0.6629000  56.7466667   0.6186000
     Totpor
 39.9333333
attr(,"scaled:scale")
    CO2SBW      Densi    FirTemp        PRA         PV     TenStr     Totpor
```

```
 0.5259152  0.1691548 53.4649955  0.4806106 27.9061201  0.3153048  7.0795326

> X2means.cluster <- KMeans(X, centers = 2, iter.max = 10, num.seeds = 10)

> X2means.cluster
K-means clustering with 2 clusters of sizes 25, 5

Cluster means:
   Totpor     PRA     PV Densi  TenStr CO2SBW FirTemp
1 41.1612 0.71396 65.132 1.682 0.58088 0.3668   741.6
2 33.7940 0.40760 14.820 1.786 0.80720 1.6560   874.0

Clustering vector:
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 1 1 2

Within cluster sum of squares by cluster:
[1] 13549.987  9580.919
 (between_SS / total_SS =  78.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"

> X2means.cluster$size # Cluster Sizes
[1] 25  5

> X2means.cluster$centers # Cluster Centroids
   Totpor     PRA     PV Densi  TenStr CO2SBW FirTemp
1 41.1612 0.71396 65.132 1.682 0.58088 0.3668   741.6
2 33.7940 0.40760 14.820 1.786 0.80720 1.6560   874.0

> X2means.cluster$withinss # Within Cluster Sum of Squares
[1] 13549.987  9580.919

> X2means.cluster$tot.withinss # Total Within Sum of Squares
[1] 23130.91

> X2means.cluster$betweenss # Between Cluster Sum of Squares
[1] 83821.46

> X2means.cluster
K-means clustering with 2 clusters of sizes 25, 5

Cluster means:
   Totpor     PRA     PV Densi  TenStr CO2SBW FirTemp
1 41.1612 0.71396 65.132 1.682 0.58088 0.3668   741.6
2 33.7940 0.40760 14.820 1.786 0.80720 1.6560   874.0

Clustering vector:
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 1 1 2

Within cluster sum of squares by cluster:
[1] 13549.987  9580.919
```

```
 (between_SS / total_SS =  78.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"
```

Ripetiamo con il campione normalizzato

```
> Y2means.cluster <- KMeans(Y, centers = 2, iter.max = 10, num.seeds = 10)

> Y2means.cluster
K-means clustering with 2 clusters of sizes 5, 25

Cluster means:
  new.x.CO2SBW new.x.Densi new.x.FirTemp new.x.PRA new.x.PV new.x.TenStr
1      2.04280     0.51240       2.06380  -0.53140 -1.50240       0.5982
2     -0.40856    -0.10232      -0.41308   0.10628  0.30044      -0.1196
  new.x.Totpor
1     -0.86720
2      0.17344

Clustering vector:
 [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 2 2 1

Within cluster sum of squares by cluster:
[1]  21.75768 107.18867
 (between_SS / total_SS =  36.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"
```

Come prima cosa ci viene detto che i due cluster contengono rispettivamente 5 e 25 individui. Vengono stampati i vettori baricentro (`Cluster means`) di ciascun cluster. Il `Clustering vector` ci dice quali individui vanno nel primo cluster e quali nel secondo e viene visualizzato il costo di ciascun cluster (`Within cluster sum of squares by cluster`

Poi in 3 cluster

```
> Y3means.cluster <- KMeans(Y, centers = 3, iter.max = 100, num.seeds = 10)

> Y3means.cluster
K-means clustering with 3 clusters of sizes 15, 5, 10

Cluster means:
  new.x.CO2SBW new.x.Densi new.x.FirTemp  new.x.PRA   new.x.PV new.x.TenStr
1   -0.3682667     -0.6896    -0.3806667 -0.1625333  0.6738667   -0.3143333
2    2.0428000      0.5124     2.0638000 -0.5314000 -1.5024000    0.5982000
3   -0.4690000      0.7786    -0.4617000  0.5095000 -0.2597000    0.1725000
  new.x.Totpor
1    0.6904667
```

```
2   -0.8672000
3   -0.6021000

Clustering vector:
 [1] 1 1 1 1 1 1 3 1 1 1 1 1 1 3 3 1 3 1 3 3 3 3 3 2 2 2 2 1 3 2

Within cluster sum of squares by cluster:
[1] 45.12115 21.75768 29.64811
 (between_SS / total_SS =  52.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"
```

Esercizio 4.2.1. La nostra popolazione è composta da 30 individui. In quanti possibili modi la possiamo suddividere in due sottoinsiemi? E in 3?

4.3. Clustering gerarchico

Come anticipato, nei metodi gerarchici si producono aggregazioni successive (metodi gerarchici aggregativi) partendo da una situazione iniziale in cui ogni individuo è un cluster a sé stante, o scissioni successive dei cluster (metodi gerarchici disgiuntivi) partendo da una situazione iniziale in cui l'intera popolazione è raccolta in un unico cluster.

L'aggregazione (o scissione) si basa su una nozione di **distanza tra cluster** che può essere definita in vari modi:

- **Distanza del nearest neighborhood** o **Single–link proximity**
  Dati due gruppi di individui G_1 e G_2, chiamo distanza di G_1 e G_2 il *minimo* di tutte le possibili distanza tra un individuo di G_1 e un individuo di G_2

$$D(G_1, G_2) := \min \{\mathrm{dist}\,(x, y) : x \in G_1,\ y \in G_2\}$$

- **Distanza del furthest neighborhood** o **Complete–link proximity**
  Dati due gruppi di individui G_1 e G_2, chiamo distanza di G_1 e G_2 il *massimo* di tutte le possibili distanza tra un individuo di G_1 e un individuo di G_2

$$D(G_1, G_2) := \max \{\mathrm{dist}\,(x, y) : x \in G_1,\ y \in G_2\}$$

- **Distanza media intragruppo** o **Average-link proximity**
  Dati due gruppi di individui G_1 e G_2, chiamo distanza di G_1 e G_2 la *media aritmetica* delle distanze tra ciascun individuo di G_1 e ciascun individuo di G_2

$$D(G_1, G_2) := \frac{1}{n_1 n_2} \sum_{\substack{x \in G_1 \\ y \in G_2}} \mathrm{dist}\,(x, y)$$

  dove n_1 è la numerosità di G_1 e n_2 è la numerosità di G_2.

- **Distanza media intergruppo** o **Average internal similarity**
  Dati due gruppi di individui G_1 e G_2, chiamo distanza di G_1 e G_2 la *media aritmetica* delle distanze tra ciascun individuo di $G_1 \cup G_2$ e ciascun individuo di $G_1 \cup G_2$

$$D(G_1, G_2) := \frac{1}{(n_1 + n_2)(n_1 + n_2 - 1)} \sum_{\substack{x,y \in G_1 \cup G_2 \\ x \neq y}} \text{dist}\,(x, y)$$

  dove n_1 è la numerosità di G_1 e n_2 è la numerosità di G_2.

- **Distanza del baricentro** o **Distanza tra centroidi**
  Dati due gruppi di individui G_1 e G_2, definisco il **baricentro** o **centroide** di ciascun gruppo:

$$\overline{g}_1 := \frac{1}{n_1} \sum_{x \in G_1} x, \qquad \overline{g}_2 := \frac{1}{n_2} \sum_{y \in G_2} y.$$

  dove n_1 è la numerosità di G_1 e n_2 è la numerosità di G_2.

  Chiamo distanza di G_1 e G_2 la distanza tra $\overline{g}_1$ e $\overline{g}_2$:

$$D(G_1, G_2) := \text{dist}\,(\overline{g}_1, \overline{g}_2)\,.$$

  dove n_1 è la numerosità di G_1 e n_2 è la numerosità di G_2.

Vediamo con un semplice esempio tratto da [1] come si procede

Supponiamo di avere una popolazione di 5 individui che indichiamo come a, b, c, d, e e di avere calcolato le loro reciproche distanze $\text{dist}\,(\cdot, \cdot)$ (ricordiamo che anche la distanza tra due individui va *scelta*) tra le distanza introdotte nella sezione 4.1. Supponiamo di aver calcolato le seguenti distanze

$$\begin{array}{llll} \text{dist}\,(a,b) = 9 & \text{dist}\,(a,c) = 3 & \text{dist}\,(a,d) = 6 & \text{dist}\,(a,e) = 11 \\ & \text{dist}\,(b,c) = 7 & \text{dist}\,(b,d) = 5 & \text{dist}\,(b,e) = 10 \\ & & \text{dist}\,(c,d) = 9 & \text{dist}\,(c,e) = 2 \\ & & & \text{dist}\,(d,e) = 8 \end{array}$$

Come distanza tra cluster scegliamo la distanza del nearest neighborhood:

$$D(G_1, G_2) := \min\{\text{dist}\,(x, y) : x \in G_1,\ y \in G_2\}$$

- **Passo 0** Ciascun punto è un cluster:

$$(a), \quad (b), \quad (c), \quad (d), \quad (e)$$

- **Passo 1** La distanza minima tra due punti è due ed è realizzata dagli individui c ed e che quindi riunisco in unico cluster. Abbiamo allora

$$(a), \quad (b), \quad (c,e), \quad (d) \qquad \text{unione avvenuta a distanza 2}$$

- **Passo 2** Calcoliamo la distanza tra i cluster ottenuti al passo precedente

$$
\begin{aligned}
&D\left((a),(b)\right) = \operatorname{dist}(a,b) = 9\\
&D\left((a),(c,e)\right) = \min\left\{\operatorname{dist}(a,c),\ \operatorname{dist}(a,e)\right\} = \min\left\{3,\ 11\right\} = 3\\
&D\left((a),(d)\right) = \operatorname{dist}(a,e) = 11\\
&D\left((b),(c,e)\right) = \min\left\{\operatorname{dist}(b,c),\ \operatorname{dist}(b,e)\right\} = \min\left\{7,\ 10\right\} = 7\\
&D\left((b),(d)\right) = \operatorname{dist}(b,d) = 5\\
&D\left((c,e),(d)\right) = \min\left\{\operatorname{dist}(d,c),\ \operatorname{dist}(d,e)\right\} = \min\left\{9,\ 8\right\} = 8
\end{aligned}
$$

  La distanza minima è 3 ed è realizzata dai cluster (a) e (c,e) che dunque unisco. Abbiamo allora

$$(a,c,e), \quad (b), \quad (d) \qquad \text{unione avvenuta a distanza } 3$$

- **Passo 3** Calcoliamo la distanza tra i cluster ottenuti al passo precedente

$$
\begin{aligned}
&D\left((a,c,e),(b)\right) = \min\left\{\operatorname{dist}(a,b),\ \operatorname{dist}(c,b),\ \operatorname{dist}(e,b)\right\} = \min\left\{9,\ 7,\ 10\right\} = 7\\
&D\left((a,c,e),(d)\right) = \min\left\{\operatorname{dist}(a,d),\ \operatorname{dist}(c,d),\ \operatorname{dist}(e,d)\right\} = \min\left\{6,\ 9,\ 8\right\} = 6\\
&D\left((b),(d)\right) = \operatorname{dist}(b,d) = 5
\end{aligned}
$$

  La distanza minima è 5 ed è realizzata dai cluster (b) e (d) che dunque unisco. Abbiamo allora

$$(a,c,e), \quad (b,d) \qquad \text{unione avvenuta a distanza } 5$$

- **Passo 4** Calcoliamo la distanza tra i cluster ottenuti al passo precedente

$$
\begin{aligned}
D\left((a,c,e),(b,d)\right) = \min\{&\operatorname{dist}(a,b), \operatorname{dist}(a,d), \operatorname{dist}(c,b), \operatorname{dist}(c,d),\\
&\operatorname{dist}(e,b), \operatorname{dist}(e,d)\} =\\
= \min\{&9,6,7,9,10,8\} = 6
\end{aligned}
$$

  Finiamo dunque con unico cluster che contiene l'intera popolazione e l'unione è avvenuta a distanza 6.

I risultati ottenuti si rappresentano in un dendrogramma (grafico ad albero)

Esercizio 4.3.1. Costruire il dendrogramma nel caso della distanza media intragruppo e nel caso della distanza del furthest neighborhood.

Esempio 4.3.1. Vediamo i dendrogrammi costruiti secondo la distanza euclidea per il campione normalizzato tratto da [3], secondo i criteri del nearest neighborhood, del furthest neighborhood e della media intragruppo.

```
> setwd("/home/laura/Documents/didattica/2012-13_elaborazioni_B194")

> X <-
+   read.table("/home/laura/Documents/didattica/2012-13_elaborazioni_B194/table2_noR5.csv",
```

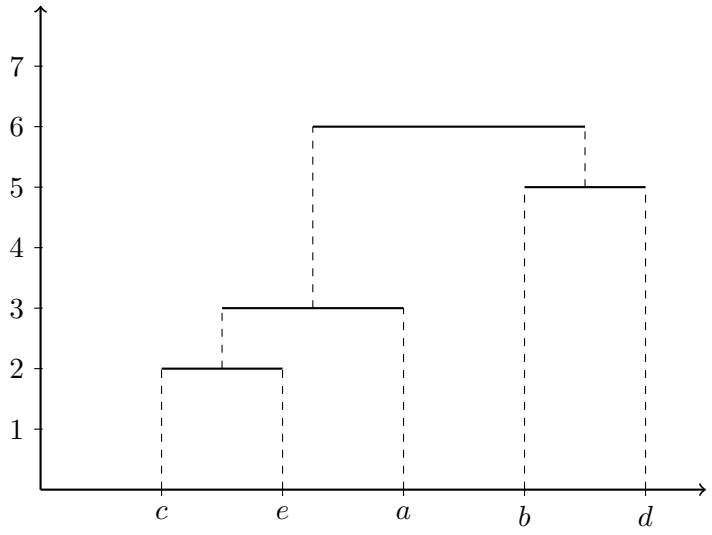

Figura 4.1: Dendrogramma con la distanza del nearest-neighborhood

```
+    header=TRUE, sep="\t", na.strings="NA", dec=".", strip.white=TRUE)

> Y <- scale(table2noR5[,c("CO2SBWn","Densin","FirTempn","PRAn","PVn",
+ "TenStrn", "Totpor")])

> Single_Euclidean <- hclust(dist(Y) , method= "single")

> plot(Single_Euclidean, main= "Cluster Dendrogram for Solution
+   Single_Euclidean", xlab= "Observation Number in Y", sub="Method=single;
+   Distance=euclidian")

> dev.copy(png,'hiera_eucl_single.png');dev.off()
png
  3
X11cairo
       2


> Complete_Euclidean <- hclust(dist(Y) , method= "complete")

> plot(Complete_Euclidean, main= "Cluster Dendrogram for Solution
+   Complete_Euclidean", xlab= "Observation Number in Y", sub="Method=complete;
+   Distance=euclidian")

> dev.copy(png,'hiera_eucl_complete.png');dev.off()
png
  3
X11cairo
```

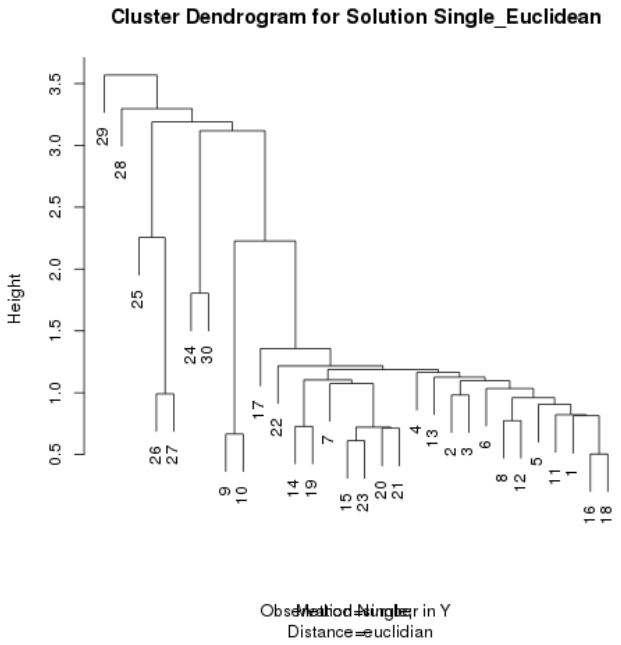

Figura 4.2: Distanza euclidea, single–link proximity

```
       2

> Average_Euclidean <- hclust(dist(Y) , method= "average")

> plot(Average_Euclidean, main= "Cluster Dendrogram for Solution
+   Average_Euclidean", xlab= "Observation Number in Y", sub="Method=average;
+   Distance=euclidian")

> dev.copy(png,'hiera_eucl_average.png');dev.off()
png
  3
X11cairo
       2
```

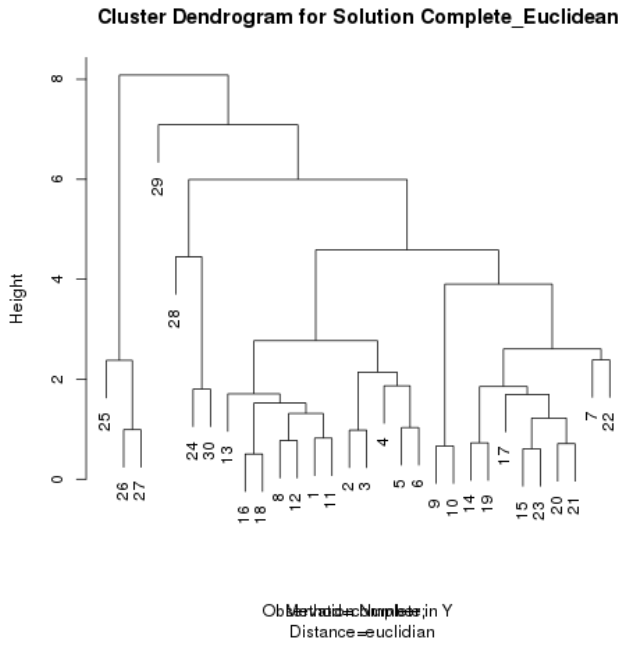

Figura 4.3: Distanza euclidea, complete–link proximity

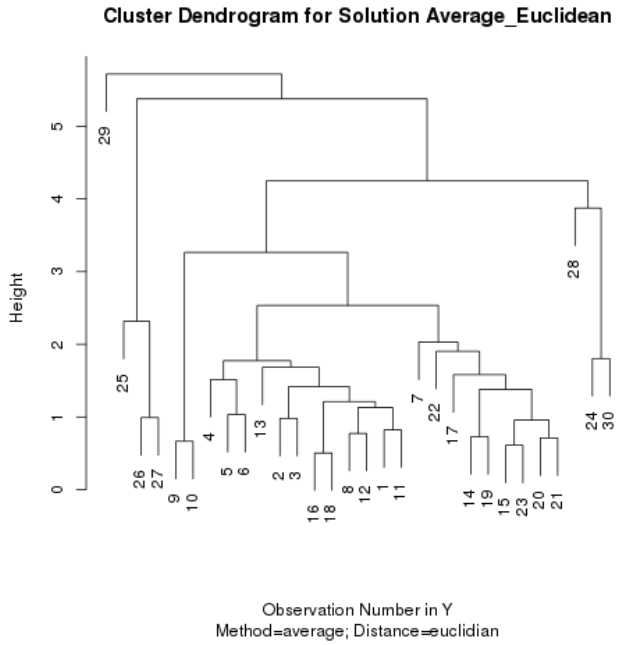

Figura 4.4: Distanza euclidea, average–link proximity

Bibliografia

[1] Luigi Barletti. Appunti del corso applicazioni di matematiche e statistica, a.a. 2007–08.

[2] Fabio Frascati. *Formulario di Statistica con R.* http://cran.r-project.org/doc/contrib/Frascati-FormularioStatisticaR.pdf, 2008.

[3] Antonia Morpoulou and Kyriaki Polikreti. Principal component analysis in monument conservation: Three application examples. *Journal of Cultural Heritage*, 10:73–81, 2009.

[4] John Verzani. *simpleR.* http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf, 2001.