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 dato, 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 Comune 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*

$$x = (x_1,\ \ldots,\ x_n)$$

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, $i = 1, \ldots, n$ assunti dal carattere sugli individui della popolazione. Suddividiamo l'intervallo $[a, b)$ in N parti uguali (N sarà suggerito dall'esperienza). Otteniamo N intervalli

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

Chiamo ciascuno di questi intervalli **classe di modalità**, se esso contiene almeno una osservazione.

1.2 Frequenza assoluta e frequenza relativa

Consideriamo un campione $x = (x_1,\ \ldots,\ x_n)$ 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

$$z_1,\ \ldots,\ z_k, \qquad 1 \leq k \leq n.$$

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

$$n_j := \#\left\{i \in \{1, \ldots, n\} \colon x_i = z_j\right\}$$

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

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

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

$$n_j := \#\left\{i \in \{1, \ldots, n\} \colon x_i \in I_j\right\}.$$

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

1.3 Moda e valori modali

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

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

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

Possiamo visualizzare con degli istogrammi, vedi Figura 1.3

1.4 Mediana

D'ora innanzi consideriamo solo caratteri numerici.
Sia dunque $x = (x_1, \ldots, x_n)$ 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 maggiore-uguale di m dati e minore-uguale 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 maggiore-uguale di $m - 1$ dati e minore-uguale di m dati. Il dato $x_{(m+1)}$ è maggiore-uguale di m dati e minore-uguale 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

$$x = (x_1, \ \ldots, \ x_n) \, .$$

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

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

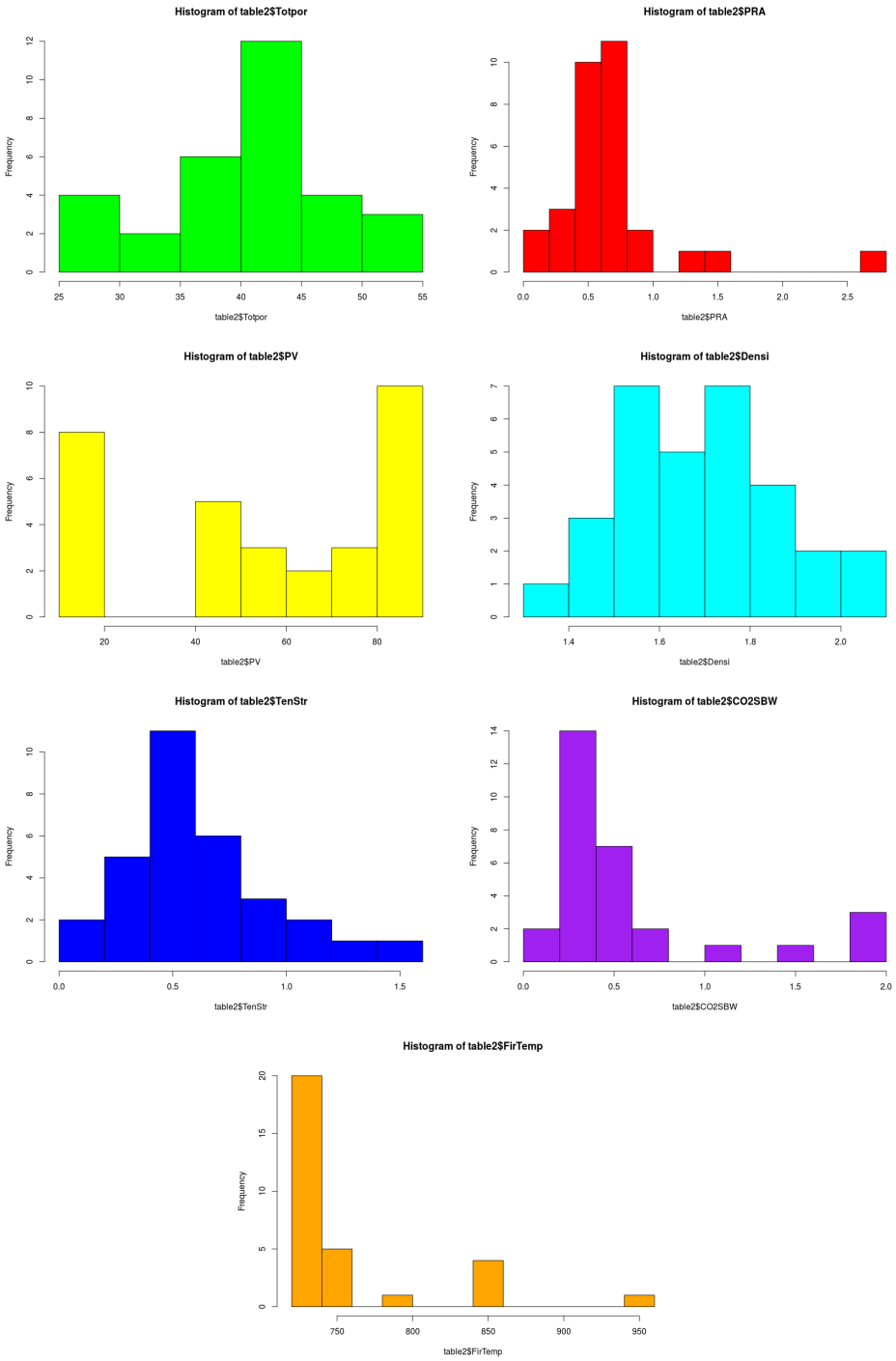

Figura 1.1: Alcuni istogrammi dall'Esempio 1.5.1

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

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \ldots x_n) = \frac{1}{n}(N_1 z_1 + N_2 z_2 + \ldots N_k z_k) =$$

$$= p_1 z_1 + p_2 z_2 + \ldots p_k z_k = \sum_{j=1}^{k} p_j z_j.$$

Chiamo **varianza campionaria** di x il numero non-negativo

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

Osserviamo che la media è un valore centrale attorno al quale si dispongono i dati $x_1, \ldots, x_n$ 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

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

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.

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

Esempio 1.5.1. Nella tabella che segue, tratta da [2], riportiamo alcuni dati relativi a campioni di laterizio e che useremo per fare alcuni esempi relativi alle nozioni introdotte mediante il software R `http://cran.r-project.org/`. Per una introduzione si rimanda ai manuali [3] e [1].

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

```
> library(readr)
> table2 <- read_delim("~/Documents/didattica/2017-18_analisi_reale/alcuni_appunti/table2.
+     "\t", escape_double = FALSE, trim_ws = TRUE)
Parsed with column specification:
cols(
  Code = col_character(),
  Totpor = col_double(),
  PRA = col_double(),
  PV = col_double(),
  Densi = col_double(),
  TenStr = col_double(),
  CO2SBW = col_character(),
  FirTemp = col_integer()
```

```
)
> View(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 con attenzione perché su un individuo non è stato rilevato)

Il commando `summary` indica l numero di dati mancanti, ci dà gli indicatori di centralità ma non quelli di dispersione

```
> summary(table2)
     Code              Totpor           PRA              PV           Densi          TenStr          CO2SBW          FirTemp
 Length:31          Min.   :26.85   Min.   :0.1580   Min.   :10.20   Min.   :1.340   Min.   :0.1430   Min.   :0.0500   Min.   :730.0
 Class :character   1st Qu.:36.05   1st Qu.:0.4220   1st Qu.:30.45   1st Qu.:1.560   1st Qu.:0.4065   1st Qu.:0.2900   1st Qu.:740.0
 Mode  :character   Median :40.90   Median :0.6220   Median :59.40   Median :1.680   Median :0.5270   Median :0.3900   Median :740.0
                    Mean   :40.12   Mean   :0.6733   Mean   :55.33   Mean   :1.693   Mean   :0.6092   Mean   :0.5817   Mean   :764.8
                    3rd Qu.:44.42   3rd Qu.:0.7305   3rd Qu.:80.70   3rd Qu.:1.815   3rd Qu.:0.7165   3rd Qu.:0.4950   3rd Qu.:750.0
                    Max.   :54.64   Max.   :2.6570   Max.   :88.60   Max.   :2.020   Max.   :1.4050   Max.   :1.9600   Max.   :960.0
                                                                                     NA's   :1
```

Richiediamo anche varianza campionaria e deviazione standard.

```
> 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
```

```
> medianaCO2SBW <- median(na.omit(table2$CO2SBW));
> meanCO2SBW <- mean(na.omit(table2$CO2SBW));
> VarCO2SBW <- var(na.omit(table2$CO2SBW));
> StdCO2SBW <- sd(na.omit(table2$CO2SBW))
> medianaCO2SBW; meanCO2SBW; VarCO2SBW;  StdCO2SBW
[1] 0.39
[1] 0.5816667
[1] 0.2765868
[1] 0.5259152
> 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
```

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

$$x = (x_1,\ \ldots,\ x_n)\,, \qquad y = (y_1,\ \ldots,\ y_n)\,.$$

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

Chiamiamo **covarianza 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 sottintesa), definiamo **coefficiente di correlazione di x e y** il numero

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

Osservazione 2.1.1. $\mathrm{Cov}\,(x,x) = s_x^2$; $\rho\,[x,x] = 1$.

Osservando che $\rho\,[x,y]$ non è altro che il rapporto tra $\langle x-(\overline{x},\ldots,\overline{x}), y-(\overline{y},\ldots,\overline{y})\rangle$ (prodotto scalare) e $\|x-(\overline{x},\ldots,\overline{x})\|\|y-(\overline{y},\ldots,\overline{y})\|$ (prodotto delle norme) si dimostrano facilmente 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 $y_i = ax_i + b \quad \forall i = 1,\ldots,n$. 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 $y_i = ax_i + b \quad \forall i = 1,\ldots,n$. 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*.

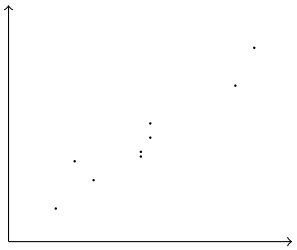

Figura 2.1: Campione bivariato *pressoché lineare*

2.2 Retta di regressione

Supponiamo di avere un campione bivariato

$$x = (x_1,\ \ldots,\ x_n)\,, \qquad y = (y_1,\ \ldots,\ y_n)$$

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 nella direzione y*

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

Si ha

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

L'incognita b compare solo nell'ultimo addendo, che è un quadrato. Quindi per ottenere il minimo basterà scegliere a che minimizza la funzione $f(a) := s_y^2 + a^2 s_x^2 - 2a\mathrm{Cov}\left(x,y\right)$ e poi scegliere $b = \overline{y} - a\overline{x}$. Si ha

$$\begin{aligned}
& f'(a) = 2as_x^2 - 2\mathrm{Cov}\left(x,y\right) = 0 \ \text{ se e solo se } \ a = \frac{\mathrm{Cov}\left(x,y\right)}{s_x^2} \\
& f''(a) = 2s_x^2 > 0
\end{aligned}$$

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

$$a = \frac{\mathrm{Cov}\,(x,y)}{s_x^2}; \qquad b = \overline{y} - \frac{\mathrm{Cov}\,(x,y)}{s_x^2}\overline{x};$$

il minimo dell'errore S vale

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

e la retta ha equazione

$$y = \overline{y} + \frac{\mathrm{Cov}\,(x,y)}{s_x^2}\,(x - \overline{x})\,.$$

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.

```
> library(readr)
> table2 <- read_delim("~/Documents/didattica/2017-18_analisi_reale/alcuni_appunti/table2.csv",
+     "\t", escape_double = FALSE, trim_ws = TRUE)
Parsed with column specification:
cols(
  Code = col_character(),
  Totpor = col_double(),
  PRA = col_double(),
  PV = col_double(),
  Densi = col_double(),
  TenStr = col_double(),
  CO2SBW = col_character(),
  FirTemp = col_integer()
)
```

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

```
> library(car)
> scatterplot(Densi~Totpor, lm=TRUE, smooth=FALSE, spread=FALSE, boxplots=TRUE, span=0.5, data= table2)
```

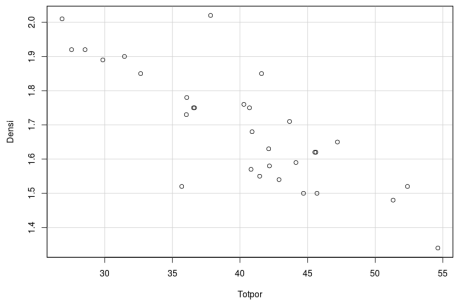

Figura 2.2: Porosità totale versus Densità

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

```
> CorTotporDensi<- cor(table2$Totpor, table2$Densi)
> CorTotporDensi
[1] -0.8187597
```

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

```
> RegModel.Densi.Totpor <- lm(Densi~Totpor, data=table2)
> summary(RegModel.Densi.Totpor)

Call:
lm(formula = Densi ~ Totpor, data = table2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.260377 -0.054570 -0.001898  0.045213  0.281783

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.486995   0.104930   23.70  < 2e-16 ***
Totpor      -0.019793   0.002577   -7.68 1.81e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09934 on 29 degrees of freedom
Multiple R-squared:  0.6704,Adjusted R-squared:  0.659
F-statistic: 58.98 on 1 and 29 DF,  p-value: 1.814e-08
```

Intercept dice che l'ordinata all'origine (il coefficiente b) della retta di regressione è 2.486995 mentre il coefficiente angolare (cioè a) è -0.019793. Ridisegniamo i punti sul piano cartesiano, aggiungendo la retta di regressione (e salviamo l'immagine in un file).

```
> abline(lm(Densi ~ Totpor, data=table2), col="red")
```

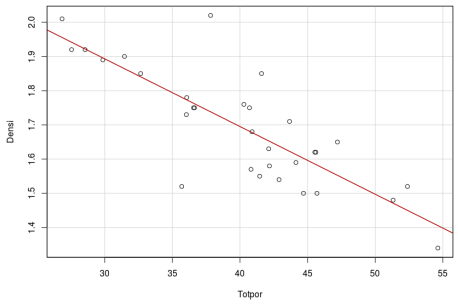

Figura 2.3: Retta di regressione lineare

Parte II

Statistica inferenziale

3. Campioni statistici

3.1 Introduzione

Scopo della statistica inferenziale è lo stabilire metodi rigorosi per ottenere – con un calcolabile *grado di certezza* proprietà generali di una popolazione a partire da una raccolta di dati sulla popolazione stessa.

Possiamo sintetizzare il modello matematico che applichiamo come segue

- Se rileviamo un carattere su una popolazione di n individui, consideriamo ciascun dato rilevato come il valore assunto da $X_1,\ X_2,\ \ldots,\ X_n$ variabili aleatorie aventi tutte la stessa distribuzione μ e che (molto spesso) si possono supporre indipendenti.

- La distribuzione μ è (parzialmente) incognita; si cercano informazioni su μ a partire dai dati rilevati. Le informazioni ricavate sulla distribuzione μ sono di natura probabilistica. Per esempio, non riusciremo ad ottenere informazioni del tipo *il valore atteso della distribuzione μ è* 50 ma informazioni del tipo *il valore atteso della distribuzione μ è compresa tra 49.8 e 50.2 con probabilità del 90%.*

Comunemente si suppone di conoscere il *tipo* della distribuzione μ, ovvero si suppone di sapere se è gaussiana, esponenziale o binomiale o altro, ma di non conoscere i parametri che la caratterizzano.

Definizione 3.1.1 (Campione statistico)**.** Una famiglia di variabili aleatorie

$$X_1,\ \ldots,\ X_n$$

si dice un *campione statistico di numerosità n* se le v.a. $X_1,\ \ldots,\ X_n$ sono indipendenti ed identicamente distribuite.
Se f è la comune densità delle v.a. $X_1,\ \ldots,\ X_n$, allora la v.a. vettoriale
$X := (X_1,\ \ldots,\ X_n)$ ha densità congiunta

$$g_{X_1,\ \ldots,\ X_n}(x_1,\ \ldots,\ x_n) = f(x_1)f(x_2)\ldots f(x_n).$$

La comune distribuzione delle X_i si dice *distribuzione campionaria di* $X_1,\ \ldots,\ X_n$.

Osservazione 3.1.1. Poiché le v.a. $X_1,\ \ldots,\ X_n$ seguono la stessa distribuzione, esse hanno anche lo stesso valore atteso e la stessa varianza (se queste quantità esistono).

Definizione 3.1.2 (Statistica)**.** Sia $X_1,\ \ldots,\ X_n$ un campione statistico. Sia $f\colon \mathbb{R}^n \to \mathbb{R}$ una funzione misurabile secondo Borel. Allora la v.a. $Y := f(X_1,\ \ldots,\ X_n)$ si dice una statistica del campione $X_1,\ \ldots,\ X_n$.

3.2 Media campionaria e varianza campionaria

Definizione 3.2.1. Sia $X_1, \ldots, X_n$ un campione statistico. Chiamiamo **media campionaria** di $X_1, \ldots, X_n$ la statistica

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i,$$

chiamiamo **varianza campionaria** di $X_1, \ldots, X_n$ la statistica

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

Proposizione 3.2.1. *Sia $X_1, \ldots, X_n$ un campione statistico di numerosità n con valore atteso μ e varianza σ^2 finiti. Siano $\overline{X}$ e S^2 la media campionaria e la varianza campionaria. Allora*

$$\mathbb{E}\left[\overline{X}\right] = \mu, \qquad \mathrm{Var}\left[\overline{X}\right] = \frac{\sigma^2}{n}, \qquad \mathbb{E}\left[S^2\right] = \sigma^2.$$

Dimostrazione.

$$\mathbb{E}\left[\overline{X}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] = \frac{1}{n}n\mu = \mu$$

$$\mathrm{Var}\left[\overline{X}\right] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

Per calcolare la media di S^2 osserviamo preliminarmente che

$$\begin{aligned} S^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}\overline{X}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - 2\overline{X}n\overline{X} + n\overline{X}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right). \end{aligned}$$

Dunque

$$\begin{aligned} (n-1)\mathbb{E}\left[S^2\right] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right] = \mathbb{E}\left[\sum_{i=1}^{n}\left(X_i - \mu + \mu\right)^2 - n\left(\overline{X} - \mu + \mu\right)^2\right] \\ &= \sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \mu + \mu\right)^2\right] - n\mathbb{E}\left[\left(\overline{X} - \mu + \mu\right)^2\right] \\ &= \sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \mu\right)^2 + \mu^2 + 2\mu(X_i - \mu)\right] \\ &\qquad - n\left(\mathbb{E}\left[\left(\overline{X} - \mu\right)^2\right] + \mu^2 - 2\mu\mathbb{E}\left[\overline{X} - \mu\right]\right) \\ &= \sum_{i=1}^{n}\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right) = (n-1)\,\sigma^2 \end{aligned}$$

e quindi $\mathbb{E}\left[S^2\right] = \sigma^2$. □