Elaborazioni Matematiche di Dati Sperimentali

Laura Poggiolini

B194 – a.a. 2012–13

Indice

I Statistica descrittiva **v**

1 Popolazioni, individui e caratteri. Indicatori sintetici di campioni monovariati **1**
1.1 Campione statistico, modalità e classi modali 2
1.2 Frequenza assoluta e frequenza relativa 2
1.3 Moda e valori modali 3
1.4 Mediana 3
1.5 Media e varianza campionaria. Scarto quadratico medio (o deviazione standard) 4

2 Campioni bivariati: covarianza, coefficiente di correlazione e retta di regressione **9**
2.1 Covarianza e coefficiente di correlazione 9
2.2 Retta di regressione 10

3 Campioni multivariati. Principal Components Analysis **15**

4 Analisi dei cluster **25**
4.1 Distanza tra individui 25
4.2 Clustering partizionale 26
4.3 Clustering gerarchico 31

II Statistica inferenziale **37**

5 Campioni statistici **39**
5.1 Introduzione 39
5.2 Media campionaria e varianza campionaria 40
5.2.1 La disuguaglianza di Chebyshev e la legge (debole) dei grandi numeri 41
5.2.2 La distribuzione gaussiana $\mathcal{N}\left(\mu, \sigma^2\right)$ e il teorema del limite centrale 43
5.3 Alcune distribuzioni legate alla distribuzione gaussiana 46
5.3.1 Distribuzione di Pearson (o χ^2) con n gradi di libertà, χ^2_n 46
5.3.2 Distribuzione t di Student con n gradi di libertà, $t(n)$ 51

6 Intervalli di confidenza **53**
6.1 Stima per intervalli della media di campioni gaussiani 54
6.1.1 Campione gaussiano di cui è nota la varianza 54
6.1.2 Campione gaussiano di cui non è nota la varianza 55
6.2 Stima per intervalli della varianza di campioni gaussiani 57

7 Test d'ipotesi **61**
7.1 Test d'ipotesi per la media di campioni gaussiani 67
7.1.1 Campione gaussiano di cui è nota la varianza 67
7.1.2 Campione gaussiano di cui non è nota la varianza 71
7.2 Test d'ipotesi per l'uguaglianza di medie di campioni gaussiani 73
7.2.1 Primo caso: le varianze σ_X^2 e σ_Y^2 sono note 73
7.2.2 Secondo caso: le varianze σ_X^2 e σ_Y^2 sono ignote ma uguali 74
7.2.3 Terzo caso: le varianze σ_X^2 e σ_Y^2 sono ignote e diverse 76
7.3 Test d'ipotesi per la varianza di campioni gaussiani 76
7.4 Test d'ipotesi per la media di campioni bernoulliani 78
7.5 Test del χ^2 80
7.5.1 Test di normalità 81

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*

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

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**.

1.2. Frequenza assoluta e frequenza relativa

Consideriamo un campione $x = (x_1,\ x_2,\ \ldots,\ x_n)$ 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

$$z_1,\ z_2,\ \ldots,\ z_k, \qquad 1 \leq k \leq n.$$

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

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

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 più 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 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

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

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

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

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

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

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

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

Osserviamo che la media è un valore centrale attorno al quale si dispongono i dati $x_1,\ x_2, \ldots, x_n$ 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

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

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.

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

Esempio 1.5.1. Nella tabella che segue, tratta da [**?**], 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 [**?**] e [**?**].

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

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

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

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

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

Si possono dimostrare 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*.

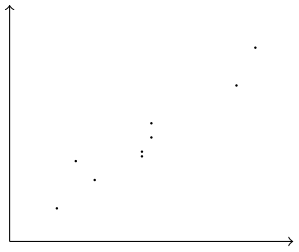

Figura 2.1: Campione bivariato

2.2. Retta di regressione

Supponiamo di avere un campione bivariato

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

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*

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

Si ha

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

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)
```

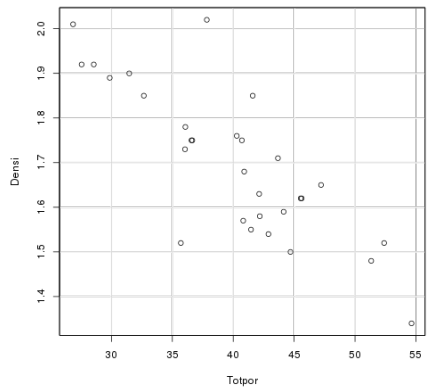

Figura 2.2: Porosità totale versus Densità

```
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))
```

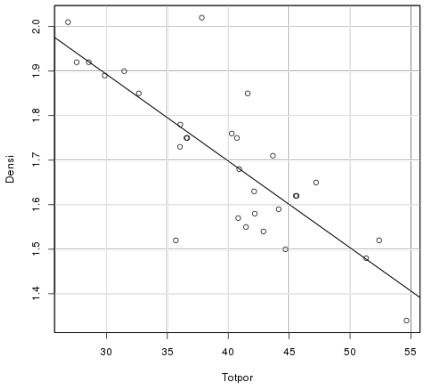

Figura 2.3: Retta di regressione lineare

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 individuo 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

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

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

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

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

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

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 simmetrica e gli elementi diagonali $c_{\ell\ell}$ sono tutti uguali ad 1.

Sull'i-esimo individuo abbiamo i dati standardizzati $y_{i1},\ y_{i2},\ \ldots,\ y_{ik}$. Vogliamo sos-

tituirli 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

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

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 [**?**]. 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
```

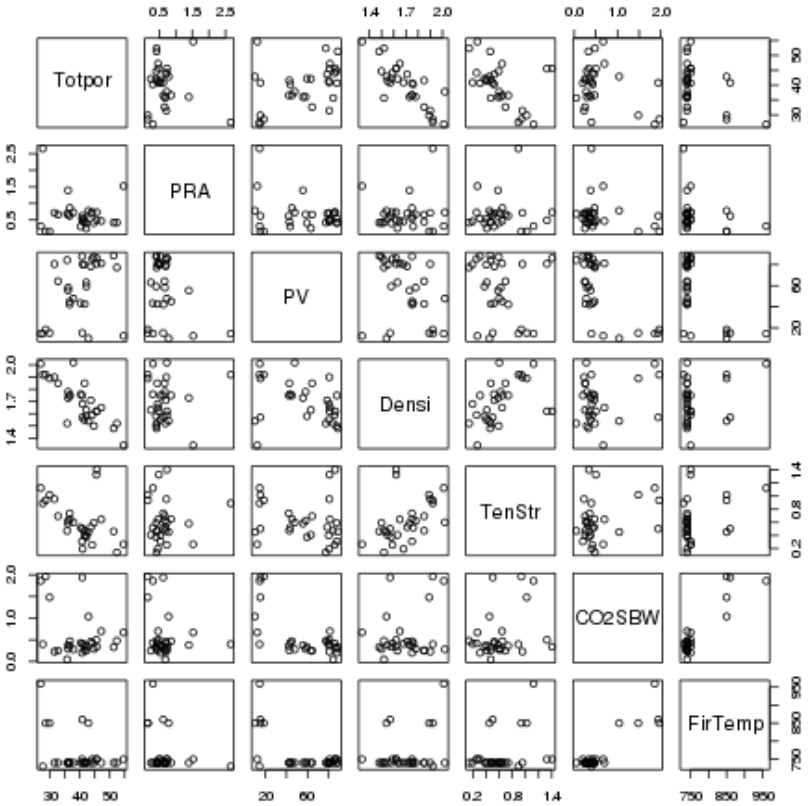

Figura 3.1: Plot dei caratteri, due a due

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-

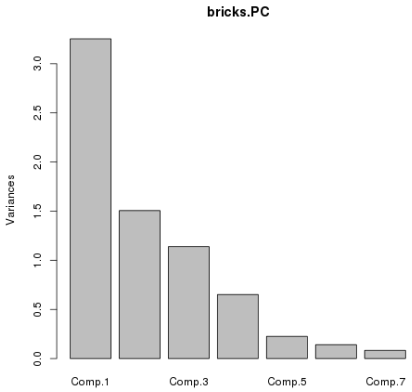

Figura 3.2: Varianza delle componenti principali

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 (approssimiamo 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,** $p \in [1, +\infty)$**,** $w = \big(w_1,\ w_2,\ \ldots,\ w_k\big)$**,** $w_j \geq 0 \quad \forall j = 1, \ldots, k$
$$d_p(x,y) := \left(\sum_{j=1}^{k} w_j\, |x_j - y_j|^p \right)^{1/p}$$

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 [**?**]. 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

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

- **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

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

- **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

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

  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 [**?**] 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 [**?**], 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",
```

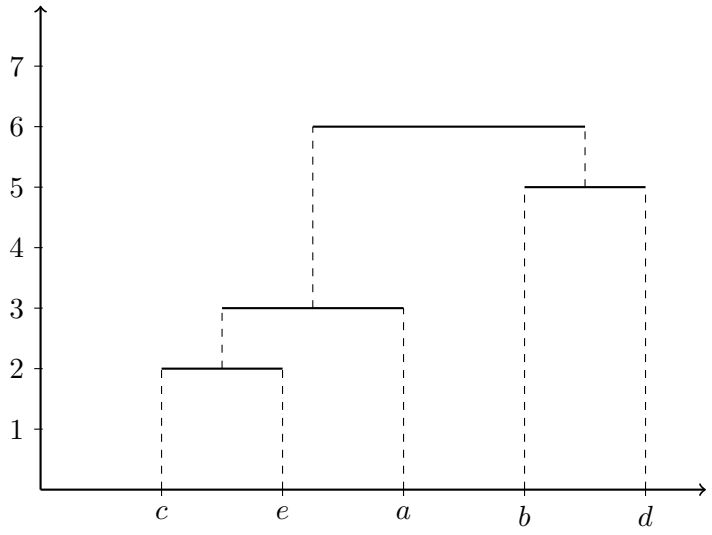

Figura 4.1: Dendrogramma con la distanza del nearest-neighborhood

```
+    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
```

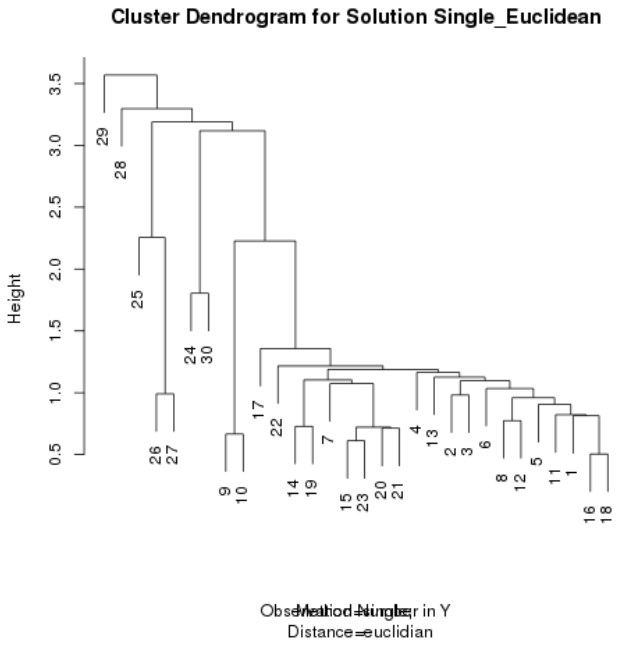

Figura 4.2: Distanza euclidea, single–link proximity

```
       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
```

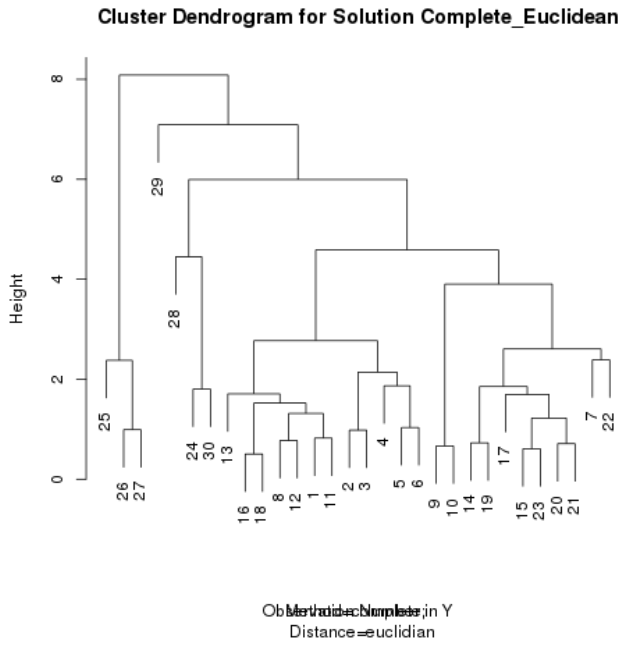

Figura 4.3: Distanza euclidea, complete–link proximity

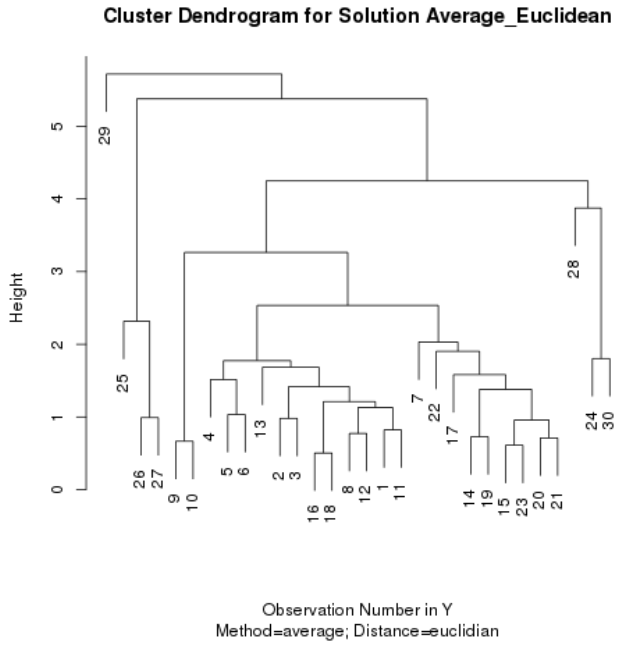

Figura 4.4: Distanza euclidea, average–link proximity

Parte II

Statistica inferenziale

5. Campioni statistici

5.1. 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 $\mathcal{D}$ e che (molto spesso) si possono supporre indipendenti.

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

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

Definizione 5.1.1 (Campione statistico)**.** Una famiglia di variabili aleatorie

$$X_1,\ X_2,\ \ldots,\ X_n$$

si dice un *campione statistico di numerosità n* se le v.a. $X_1,\ X_2,\ \ldots,\ X_n$ sono indipendenti ed identicamente distribuite.
Se f è la comune densità delle v.a. $X_1,\ X_2,\ \ldots,\ X_n$, allora la v.a. vettoriale
$X := (X_1,\ X_2,\ \ldots,\ X_n)$ ha densità congiunta

$$g_{X_1,\ X_2,\ \ldots,\ X_n}(x_1,\ x_2,\ \ldots,\ x_n) = f(x_1)f(x_2)\ldots f(x_n).$$

La comune distribuzione delle X_i si dice *distribuzione campionaria di $X_1,\ X_2,\ \ldots,\ X_n$.*

Osservazione 5.1.1. Poiché le v.a. $X_1,\ X_2,\ \ldots,\ X_n$ seguono la stessa distribuzione, esse hanno anche la stessa media e la stessa varianza (se queste quantità esistono).

Definizione 5.1.2 (Statistica)**.** Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico. Una funzione (non dipendente da parametri) di $X_1,\ X_2,\ \ldots,\ X_n$ si dice una statistica.

Osservazione 5.1.2. Chiariamo cosa si intende per statistica: $3X_1 - 2X_2$ è una statistica; $\max\{X_1,\ X_2,\ \ldots,\ X_n\}$ è una statistica. $X_1 - \mu\ \mu \in \mathbb{R}$ *non* è una statistica.

5.2. Media campionaria e varianza campionaria

Definizione 5.2.1. Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico. Chiamiamo **media campionaria** di $X_1,\ X_2,\ \ldots,\ X_n$ la statistica

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

chiamiamo **varianza campionaria** di $X_1,\ X_2,\ \ldots,\ X_n$ la statistica

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

Proposizione 5.2.1. *Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico di numerosità n con media μ e varianza σ^2 finite. 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

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

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

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

5.2.1. La disuguaglianza di Chebyshev e la legge (debole) dei grandi numeri

Enunciamo alcuni importanti risultati asintotici che giustificano l'uso della media campionaria $\overline{X}$ come stima della media μ del campione.

Teorema 5.2.1 (Disuguaglianza di Chebyshev)**.** *Se X è una variabile aleatoria con media μ e varianza non superiore a σ^2, allora*

$$\mathbb{P}\left(|X-\mu| \geq t\right) \leq \frac{\sigma^2}{t^2} \qquad \forall t > 0.$$

Dimostrazione. Consideriamo la v.a. $Y := \begin{cases} 0 & \text{se } |X-\mu| < t, \\ t & \text{se } |X-\mu| \geq t. \end{cases}$
Sicuramente $0 \leq Y \leq |X-\mu|$, quindi $Y^2 \leq (X-\mu)^2$ e dunque

$$\mathbb{E}\left[Y^2\right] \leq \mathbb{E}\left[(X-\mu)^2\right] = \mathrm{Var}\left[X\right].$$

D'altra parte

$$\mathbb{E}\left[Y^2\right] = 0\,\mathbb{P}\left(|X-\mu| < t\right) + t^2\,\mathbb{P}\left(|X-\mu| \geq t\right) = t^2\,\mathbb{P}\left(|X-\mu| \geq t\right),$$

da cui la tesi. □

Osservazione 5.2.1. La disuguaglianza di Chebyshev può anche essere formulata nel seguente modo: Se X è una variabile aleatoria con media μ e varianza σ^2 finite, allora

$$\mathbb{P}\left(|X-\mu| > \eta\,\sigma\right) \leq \frac{1}{\eta^2} \qquad \forall \eta > 0.$$

Ovvero: la probabilità che X disti dalla sua media μ più di una frazione η della deviazione standard σ è inferiore a $\frac{1}{\eta^2}$.

Esempio 5.2.1. Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico di numerosità n. Supponiamo di conoscere la varianza $\sigma^2 = 4$ del campione e che la media μ sia ignota. Quanto deve essere grande n per poter affermare che

$$\mathbb{P}\left(\left|\overline{X}-\mu\right| > 1\right) \leq \frac{1}{10}?$$

Sappiamo che

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| > 1\right) \leq \frac{\sigma^2}{n\,1^2} = \frac{4}{n}.$$

è allora sufficiente richiedere $\dfrac{4}{n} \leq \dfrac{1}{10}$ cioè $n \geq 40$.

Dalla disuguaglianza di Chebyshev segue facilmente il seguente

Teorema 5.2.2 (Legge debole dei grandi numeri)**.** *Sia $\{X_i\}_{i=1}^{\infty}$ una successione di v.a. indipendenti, identicamente distribuite, con media μ e varianza σ^2 finite.*

Per ogni $n \in \mathbb{N}$ sia $\overline{X}_n := \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$. Allora

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| > t\right) = 0 \qquad \forall t > 0.$$

Dimostrazione. Poiché $\mathbb{E}\left[\overline{X}_n\right] = \mu$ e $\mathrm{Var}\left[\overline{X}_n\right] = \dfrac{\sigma^2}{n}$, per la disuguaglianza di Chebyshev si ha

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| > t\right) \leq \frac{\sigma^2}{nt^2} \qquad \forall n \in \mathbb{N}.$$

La tesi segue passando al limite. □

La legge debole dei grandi numeri ci *autorizza* a usare il valore di $\overline{X}_n$ come sostituto della media μ della distribuzione e la disuguaglianza di Chebyshev ci dice con precisione quanto è *probabilisticamente accettabile* questa sostituzione.

Esempio 5.2.2. Ho una monetina che potrebbe essere truccata. Voglio scoprire, con un'approssimazione di ± 0.05 e con un grado di certezza del 90% quanto vale la probabilità di ottenere testa in un singolo lancio. Posso formalizzare ogni singolo lancio della monetina con una variabile aleatoria di Bernoulli di parametro p dove p è la probabilità (incognita) di ottenere testa in un singolo lancio. Se lancio la monetina n volte ho allora un campione statistico $X_1,\ X_2,\ \ldots,\ X_n$ che segue la distribuzione $B(p)$. Sia $\overline{X}_n$ la media campionaria di questo campione. Allora

$$\mathbb{E}\left[\overline{X}_n\right] = p, \qquad \mathrm{Var}\left[\overline{X}_n\right] = \frac{p(1-p)}{n}.$$

Per la disuguaglianza di Chebyshev

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| \geq 0.05\right) \leq \frac{p(1-p)}{n(0.05)^2} \leq \frac{400}{4n} = \frac{100}{n}$$

Voglio

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| \leq 0.05\right) \geq \frac{90}{100}$$

cioè

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| \geq 0.05\right) \leq 1 - \frac{90}{100} = \frac{1}{10}$$

Basta allora avere $\frac{100}{n} \leq \frac{1}{10}$ cioè $n \geq 1000$. Dunque: tiro la monetina 1000 volte registrando il risultato ad ogni i-esimo lancio ($x_i = 1$) o croce ($x_i = 0$) vedendo questo numero come il valore assunto da una v.a. bernoulliana X_i di parametro p.

Calcolo $\overline{x} = \frac{1}{1000} \sum_{i=1}^{1000} x_i$ e lo vedo come il valore assunto dalla v.a. $\overline{X}$. La probabilità che il valore $\overline{x}$ differisca da p per meno di 0.05 è maggiore-uguale del 90%.

Più in generale

Esempio 5.2.3. Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico di numerosità n, bernoulliano di parametro (incognito) $p \in [0,1]$. Dunque

$$\mathbb{E}[X_i] = p \qquad \mathrm{Var}[X_i] = p(1-p)$$
$$\mathbb{E}\left[\overline{X}\right] = p \qquad \mathrm{Var}\left[\overline{X}\right] = \frac{p(1-p)}{n}$$

Allora, per la disuguaglianza di Chebyshev

$$\mathbb{P}\left(\left|\overline{X} - p\right| > t\right) \leq \frac{p(1-p)}{n\,t^2} \leq \frac{1}{4n\,t^2} \qquad \forall t > 0. \tag{5.1}$$

poiché $p(1-p) \leq \frac{1}{4} \quad \forall p \in [0,1]$.

5.2.2. La distribuzione gaussiana $\mathcal{N}\left(\mu, \sigma^2\right)$ e il teorema del limite centrale

Ricordiamo che la distribuzione gaussiana di parametri $\mu \in \mathbb{R}$ e $\sigma^2 > 0$, $\mathcal{N}\left(\mu, \sigma^2\right)$, è la distribuzione assolutamente continua associata alla densità

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}.$$

Se una v.a. X segue la distribuzione $\mathcal{N}\left(\mu, \sigma^2\right)$, allora

$$\mathbb{E}[X] = \mu, \qquad \mathrm{Var}[X] = \sigma^2.$$

Inoltre $f(x) > 0$ per ogni $x \in \mathbb{R}$, quindi la funzione di ripartizione $F_X(x) := \mathbb{P}(X \leq x)$ è strettamente monotona crescente. Dunque, per ogni $\alpha \in (0,1)$ esiste uno ed un solo $x = x_\alpha \in \mathbb{R}$ tale $F_X(x_\alpha) = \alpha$. x_α si dice **quantile** di X di livello α. Inoltre, se $\mu = 0$, la densità è una funzione pari, e dunque $F_X(t) + F_X(-t) = 1$ per ogni $t \in \mathbb{R}$; in particolare $x_{1-\alpha} = -x_\alpha$.

Nel caso in cui $\mu = 0$, $\sigma^2 = 1$, la distribuzione $\mathcal{N}((,0),1)$ si dice *distribuzione gaussiana standard*, la funzione di ripartizione associata si indica con la lettera Φ,

$$\Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-t^2}{2}\right) \mathrm{d}t, \qquad x \in \mathbb{R}.$$

e per ogni $\alpha \in (0,1)$ il quantile di livello α si indica z_α. Dunque

$$\Phi(x) + \Phi(-x) = 1 \quad \forall x \in \mathbb{R}, \qquad z_{1-\alpha} = -z_\alpha \quad \forall \alpha \in (0,1).$$

Si possono inoltre dimostrare le seguenti proprietà

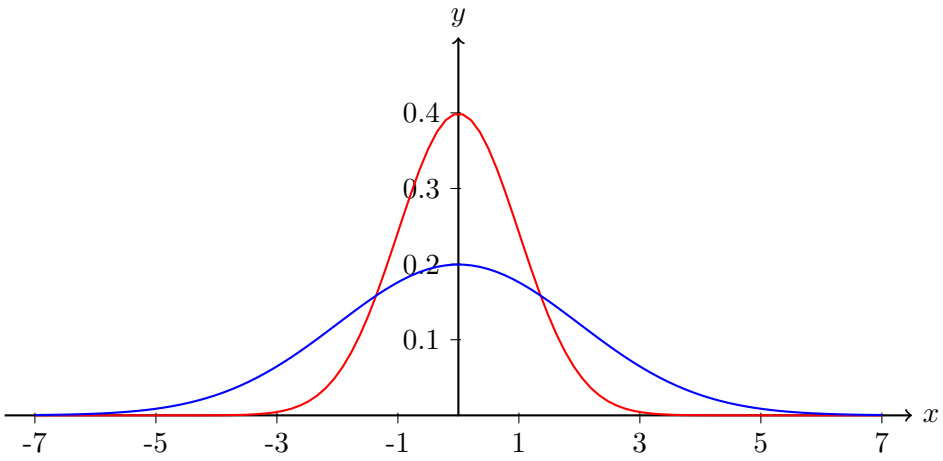

Figura 5.1: Densità associate alle distribuzioni $\mathcal{N}((,0),1)$ (in rosso) e $\mathcal{N}((,0),4)$ (in blu)

Proprietà 5.2.1. 1. Se X è una v.a. gaussiana di media μ e varianza σ^2: $X \sim \mathcal{N}(\mu, \sigma^2)$ e α, β sono due numeri reali, $\alpha \neq 0$, allora la v.a. $\alpha X + \beta$ è gaussiana di media $\alpha\mu + \beta$ e varianza $\alpha^2\sigma^2$: $\alpha X + \beta \sim \mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2)$. In particolare $Y := \dfrac{X - \mu}{\sigma}$ è una v.a. gaussiana standard: $Y \sim \mathcal{N}((,0),1)$.

2. Siano $X_1, X_2, \ldots, X_n$ v.a. indipendenti. Supponiamo che ciascuna v.a. X_i sia gaussiana di media μ_i e varianza σ_i^2: $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ $\forall i = 1, \ldots, n$. Allora la v.a. $X_1 + X_2 + \ldots + X_n$ è gaussiana di media pari alla somma delle medie e varianza pari alla somma delle varianze:

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Teorema 5.2.3 (Teorema del limite centrale)**.** *Sia $\{X_i\}_{i=1}^{\infty}$ una successione di v.a. indipendenti, identicamente distribuite, con media μ e varianza σ^2 finite. Sia $\Phi(t)$ la funzione di ripartizione associata alla distribuzione gaussiana standard $\mathcal{N}(0,1)$.*

Per ogni $n \in N$ sia $\overline{X}_n$ la media campionaria di $X_1, X_2, \ldots, X_n$ e sia $\overline{Z}_n$ la sua standardizzazione:

$$\overline{Z}_n := \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

Allora

$$\lim_{n\to\infty} \mathbb{P}\left(\overline{Z}_n \leq t\right) = \Phi(t) \qquad \forall t \in \mathbb{R}$$

Osservazione 5.2.2. Una formulazione equivalente della tesi del teorema del limite centrale è

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leq t\right) = \Phi(t) \qquad \forall t \in \mathbb{R}$$

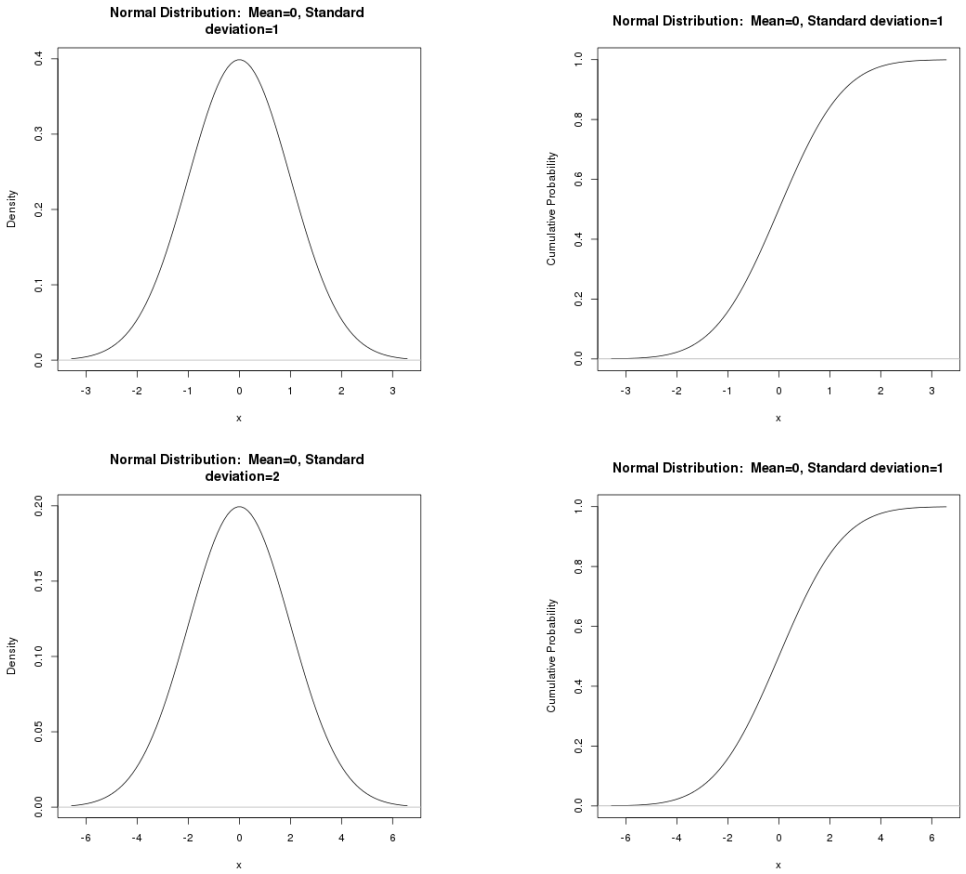

Figura 5.2: $\mathcal{N}((,0),1)$ e $\mathcal{N}((,0),4)$, densità e funzione di ripartizione

Esempio 5.2.4. Supponiamo di avere un campione statistico di numerosità 25 e deviazione standard 8. Qual è la probabilità che la media campionaria differisca dalla media del campione per più di 4?

Devo calcolare

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| > 4\right)$$

dove $\mu = \mathbb{E}[X_i] \quad \forall i = 1, \ldots, n$ e dunque è anche $\mu = \mathbb{E}\left[\overline{X}\right]$. Applicando la disuguaglianza di Chebyshev otteniamo

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| > 4\right) \leq \frac{\mathrm{Var}\left[\overline{X}\right]}{4^2} = \frac{64}{25 \cdot 16} = \frac{4}{25} = 0.16$$

Proviamo ad applicare il teorema del limite centrale. Indico con $\overline{Z}$ la standardizzazione

della media campionaria. Si ha

$$\begin{aligned}
&\mathbb{P}\left(\left|\overline{X}-\mu\right|>4\right)=\mathbb{P}\left(\frac{\left|\overline{X}-\mu\right|}{\frac{\sigma}{\sqrt{n}}}>\frac{4}{\frac{\sigma}{\sqrt{n}}}\right)=\mathbb{P}\left(\left|\overline{Z}\right|>\frac{4}{\frac{8}{\sqrt{25}}}\right)= \\
&=\mathbb{P}\left(\left|\overline{Z}\right|>\frac{5}{2}\right)=\mathbb{P}\left(\overline{Z}>\frac{5}{2}\right)+\mathbb{P}\left(\overline{Z}<\frac{-5}{2}\right) \\
&=1-\Phi\left(2.5\right)+\Phi\left(-2.5\right)=2\left(1-\Phi\left(2.5\right)\right) \\
&=2\left(1-\Phi\left(2.5\right)\right)\simeq 2\left(1-0.9938\right)=0.0124
\end{aligned}$$

Perché questa stima *sembra* tanto migliore di quella ottenuta con la disuguaglianza di Chebyshev? Perché non abbiamo un'indicazione sul significato di quei $\simeq$. In altre parole, il teorema del limite centrale è appunto un teorema di passaggio al limite e non fornisce una stima dell'errore che si compie sostituendo $\mathbb{P}\left(Z_n \leq t\right)$ con $\Phi(t)$. A tal proposito vale il seguente

Teorema 5.2.4 (Teorema di Berry–Esseen)**.** *Sia $\{X_i\}_{i=1}^{\infty}$ una successione di v.a. indipendenti, identicamente distribuite, con media $\mu = 0$, varianza σ^2 e momento terzo $\gamma := \mathbb{E}\left[|X_i|^3\right]$ finiti. Sia $\Phi(t)$ la funzione di ripartizione associata alla distribuzione gaussiana standard $\mathcal{N}\left((,0),1\right)$.*

Sia $C := \dfrac{0.8\gamma}{\sigma^3}$. Allora

$$\left|\mathbb{P}\left(\frac{\overline{X}_n}{\frac{\sigma}{\sqrt{n}}}\leq t\right)-\Phi(t)\right|\leq\frac{C}{\sqrt{n}}\qquad\forall t\in\mathbb{R}$$

Dal Teorema di Berry–Esseen, teorema 5.2.4, otteniamo dunque

$$\left|\mathbb{P}\left(\overline{Z}_n\leq t\right)-\Phi(t)\right|\leq\frac{C}{\sqrt{n}}\qquad\forall t\in\mathbb{R}$$

5.3. Alcune distribuzioni legate alla distribuzione gaussiana

5.3.1. Distribuzione di Pearson (o χ^2) con n gradi di libertà, χ^2_n

Si chiama così la distribuzione associata alla densità

$$f(x):=\begin{cases}\dfrac{1}{\Gamma\left(\frac{n}{2}\right)}\left(\dfrac{1}{2}\right)^{\frac{n}{2}}x^{\frac{n}{2}-1}\exp\left(\frac{-x}{2}\right) & x>0,\\ 0 & x\leq 0,\end{cases}$$

dove $\Gamma(a):=\displaystyle\int_0^{+\infty}x^{a-1}\mathrm{e}^{-x}\,\mathrm{d}x$, $a>0$

Osservazione 5.3.1. Si può dimostrare che $\forall a > 0$ si ha $\Gamma(a+1) = a\Gamma(a)$ e che $\Gamma(1) = 1$, $\Gamma\left(\dfrac{1}{2}\right) = \sqrt{\pi}$. Quindi

$\Gamma(2) = 1{\cdot}1$, $\quad\Gamma(3) = 2\,\Gamma(2) = 2{\cdot}1 = 2!\quad\ldots\quad\Gamma(n) = (n-1)!$ per ogni intero positivo n

mentre

$$\Gamma\left(\frac{3}{2}\right) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{1}{2}\sqrt{\pi}, \qquad \Gamma\left(\frac{5}{2}\right) = \frac{3}{2}\Gamma\left(\frac{3}{2}\right) = \frac{3\cdot 1}{2\cdot 2}\sqrt{\pi} = \frac{3!!}{2^2}\sqrt{\pi},$$
$$\ldots \qquad \Gamma\left(\frac{2k+1}{2}\right) = \frac{(2k-1)!!}{2^k}\sqrt{\pi} \quad \text{per ogni intero non-negativo } k.$$

Si possono calcolare media e varianza di una v.a. che segua una distribuzione di Pearson:

Proprietà 5.3.1. Se X è una v.a. con distribuzione χ^2 a n gradi di libertà, $X \sim \chi^2_n$, allora

$$\mathbb{E}[X] = n, \qquad \mathrm{Var}[X] = 2n.$$

Proprietà 5.3.2. Se X e Y sono due variabili di Pearson indipendenti, $X \sim \chi^2_n, Y \sim \chi^2_k$, allora la v.a. $X + Y$ segue la distribuzione di Pearson a $n + k$ gradi di libertà:

$$X + Y \sim \chi^2_{n+k}.$$

Il seguente teorema dà un legame tra la distribuzione gaussiana e le distribuzioni χ^2:

Teorema 5.3.1. *Se $X_1,\ X_2,\ \ldots,\ X_n$ sono v.a. indipendenti e gaussiane, con X_i di media μ_i e varianza σ_i^2, $\forall i = 1, \ldots, n$, allora la v.a. $\displaystyle\sum_{i=1}^{n}\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2$ segue la distribuzione di Pearson a n gradi di libertà, χ^2_n.*

Corollario 5.3.2. *Se $X_1,\ X_2,\ \ldots,\ X_n$ è un campione statistico gaussiano, con media μ e varianza σ^2, allora la v.a. $\chi^2 := \displaystyle\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$ segue una distribuzione χ^2 con n gradi di libertà.*

Esempio 5.3.1. Si vuole localizzare un oggetto puntiforme, misurandone le tre coordinate cartesiane rispetto ad un prefissato sistema di riferimento. L'errore sperimentale, misurato in millimetri per ciascuna delle tre coordinate è una v.a. gaussiana di media 0 e deviazione standard 2.

Supponendo che i tre errori siano v.a. indipendenti, calcolare la probabilità che la distanza tra la posizione misurata e la posizione reale sia inferiore a 1.2 mm.

Soluzione. Indico con X_1, X_2, X_3, gli errori commessi nella misurazione delle tre coordinate. Per il Teorema di Pitagora la distanza tra le due posizioni è

$$D = \sqrt{X_1^2 + X_2^2 + X_3^2}$$

Vogliamo calcolare $\mathbb{P}(D < 1.2) = \mathbb{P}\left(D^2 < 1.44\right) = \mathbb{P}\left(X_1^2 + X_2^2 + X_3^2 < 1.44\right)$.

Pongo $Z_i := \dfrac{X_i}{\sigma} = \dfrac{X_i}{2}$, $i = 1, 2, 3$, da cui $X_i^2 = 4Z_i^2$ e dunque

$$\begin{aligned}\mathbb{P}(D < 1.2) &= \mathbb{P}\left(X_1^2 + X_2^2 + X_3^2 < 1.44\right) = \mathbb{P}\left(4\left(Z_1^2 + Z_2^2 + Z_3^2\right) < 1.44\right) \\ &= \mathbb{P}\left(Z_1^2 + Z_2^2 + Z_3^2 < .36\right).\end{aligned}$$

Basterà dunque controllare (vedi ultima riga del listato a seguire) il valore della funzione di ripartizione delle v.a. di distribuzione χ^2_3 nel punto 0.36 che è (circa) 0.052.

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

> .x <- seq(0.015, 17.73, length.out=100)

> plot(.x, dchisq(.x, df=3), xlab="x", ylab="Density",
 main=paste("ChiSquared Distribution:  Degrees of freedom=3"), type="l")

> abline(h=0, col="gray")

> remove(.x)

> dev.copy(png,'densitachiquadro3.png');dev.off()
png
  3
X11cairo
       2

> .x <- seq(0.015, 17.73, length.out=100)

> plot(.x, pchisq(.x, df=3), xlab="x", ylab="Cumulative Probability",
 main=paste("ChiSquared Distribution:  Degrees of freedom=3"), type="l")

> abline(h=0, col="gray")

> remove(.x)

> pchisq(c(0.36), df=3, lower.tail=TRUE)
[1] 0.05162424

> .x <- seq(0.015, 17.73, length.out=100)

> plot(.x, dchisq(.x, df=3), xlab="x", ylab="Density",
 main=paste("ChiSquared   Distribution:  Degrees of freedom=3"),type="l")

> abline(h=0, col="gray")

> remove(.x)

> dev.copy(png,'densitachiquadro3.png');dev.off()
png
  3
```

```
X11cairo
       2

> .x <- seq(0.015, 17.73, length.out=100)

> plot(.x, pchisq(.x, df=3), xlab="x", ylab="Cumulative Probability",
 main=paste("ChiSquared Distribution:  Degrees of freedom=3"),type="l")

> abline(h=0, col="gray")

> remove(.x)

> dev.copy(png,'ripartizionechiquadro3.png');dev.off()
png
  3
X11cairo
       2

> pchisq(c(0.36), df=3, lower.tail=TRUE)
[1] 0.05162424
```

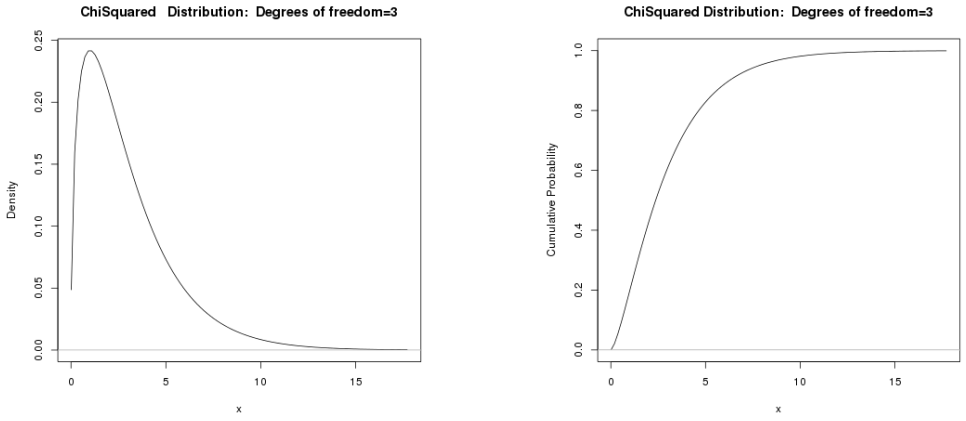

Figura 5.3: χ^2_3, densità e funzione di ripartizione

A titolo di confronto, visualizziamo anche densità e funzione di ripartizione delle distribuzioni χ^2_{10} e χ^2_{100}. Il seguente teorema raccoglie alcune importanti proprietà dei campioni statistici gaussiani e delle loro media e varianza campionarie.

Teorema 5.3.3. *Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico gaussiano di numerosità n, media μ e varianza σ^2.*

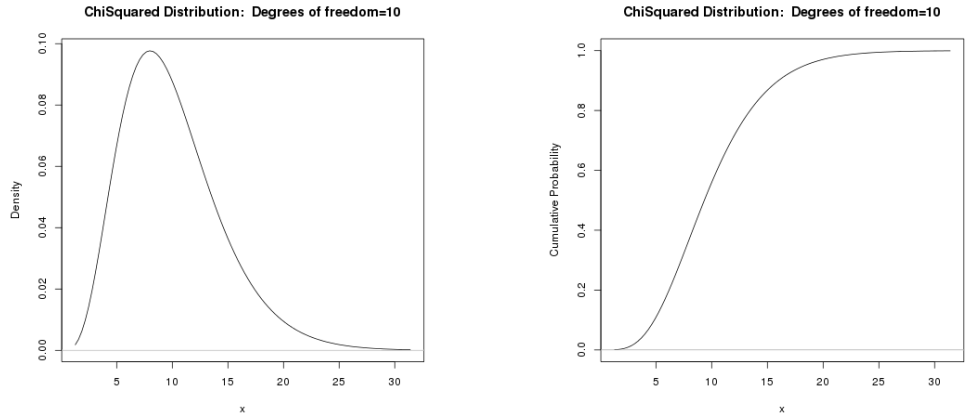

Figura 5.4: χ^2_{10}, densità e funzione di ripartizione

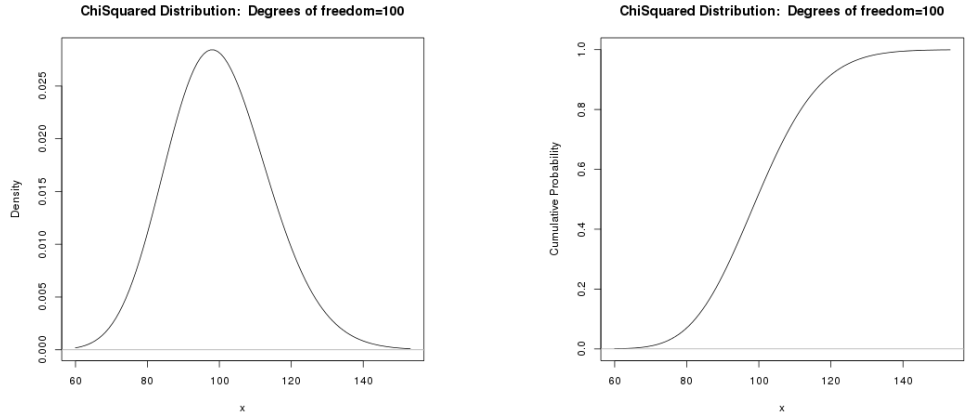

Figura 5.5: χ^2_{100}, densità e funzione di ripartizione

Allora, la media campionaria $\overline{X} := \dfrac{1}{n}\sum_{i=1}^{n} X_i$ *e la varianza campionaria* $S^2 := \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ *sono v.a. indipendenti.*

Sia $Z_1,\ Z_2,\ \ldots Z_n$ *la standardizzazione del campione statistico* $X_1,\ X_2,\ \ldots,\ X_n$ *i.e.*

$$Z_i := \frac{X_i - \mu}{\sigma} \qquad \forall i = 1, \ldots, n.$$

e sia $\overline{Z}$ *la media campionaria del campione normalizzato* $Z_1,\ Z_2,\ \ldots,\ Z_n$.

Allora $\overline{Z} = \dfrac{\overline{X} - \mu}{\sigma}$ *e la v.a.* $\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2$ *segue una distribuzione* χ^2 *con* $n-1$ *gradi di libertà.*

Corollario 5.3.4. *Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico gaussiano di numerosità n, media μ e varianza σ^2 e sia S^2 la sua varianza campionaria. Allora la v.a. $V := (n-1)\dfrac{S^2}{\sigma^2}$ segue una distribuzione χ^2 con $n-1$ gradi di libertà.*

Dimostrazione. Si ha infatti

$$V = (n-1)\frac{S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left((\mu+\sigma Z_i) - (\mu+\sigma\overline{Z})\right)^2 = \sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2$$

□

5.3.2. Distribuzione t di Student con n gradi di libertà, $t(n)$

Si chiama così la distribuzione associata alla densità

$$\tau_n(x) = \frac{\Gamma\left(\dfrac{n+1}{2}\right)}{\Gamma\left(\dfrac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{\frac{-(n+1)}{2}} \qquad x \in \mathbb{R}.$$

Proprietà 5.3.3. Se X è una v.a. con distribuzione t di Student a n gradi di libertà, allora

$$\mathbb{E}[X] = 0, \qquad \mathrm{Var}[X] = \begin{cases} \dfrac{n}{n-2} & \text{se } n \geq 3, \\ +\infty & \text{se } n = 1,\ 2. \end{cases}$$

Osservazione 5.3.2. Il quantile di livello $\alpha \in (0,1)$ associato alla distribuzione $t(n)$ si indica $t_{n,\alpha}$. Poiché la densità τ_n è una funzione pari, se $X \sim t(n)$, allora $F_X(x) + F_X(-x) = 1$. Dunque per i quantili della distribuzione $t(n)$ si ha $t_{n,\alpha} = -t_{n,1-\alpha}$ per ogni $\alpha \in (0,1)$.

Teorema 5.3.5. *w Se Z è una v.a. gaussiana standard, $Z \sim \mathcal{N}(0,1)$, se Y segue la distribuzione χ^2 con n gradi di libertà, $Y \sim \chi^2_n$ e se Z e Y sono indipendenti, allora la v.a. $T := \dfrac{Z\sqrt{n}}{\sqrt{Y}}$ segue la distribuzione t di Student a n gradi di libertà:*

$$T := \frac{Z\sqrt{n}}{\sqrt{Y}} \sim t(n).$$

Corollario 5.3.6. *Se $X_1,\ X_2,\ \ldots,\ X_n$ è un campione statistico gaussiano di numerosità n, media μ e varianza σ^2, allora*

$$T := \frac{\left(\overline{X} - \mu\right)\sqrt{n}}{S}$$

segue la distribuzione t di Student con $n-1$ gradi di libertà:

$$T \sim t(n-1).$$

Dimostrazione. Basta applicare il teorema 5.3.5 con $Z = \dfrac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ e $Y = (n-1)\dfrac{S^2}{\sigma^2}$. □

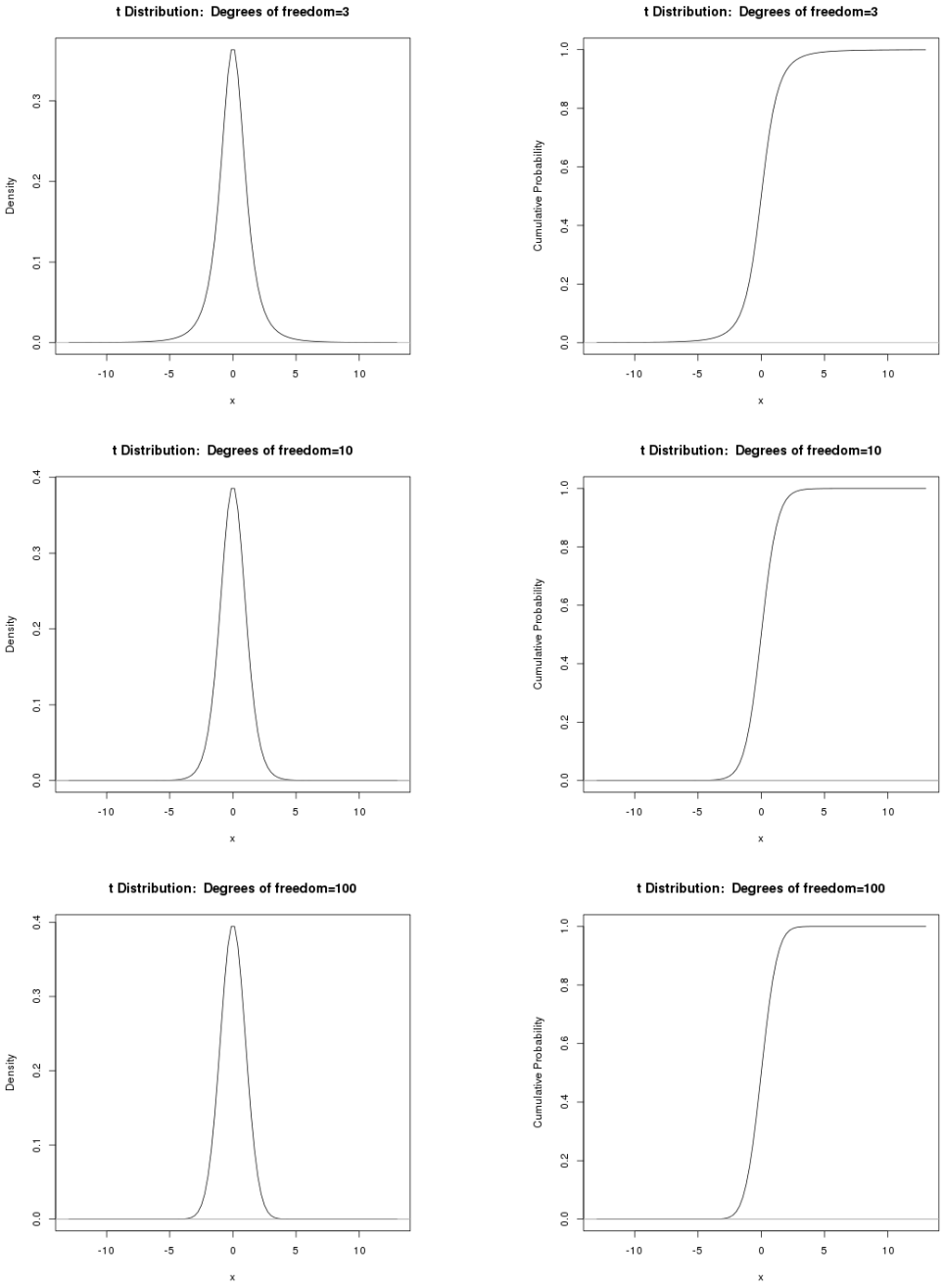

Figura 5.6: $t(3)$, $t(10)$, $t(100)$, densità e funzione di ripartizione

6. Intervalli di confidenza

La media campionaria e la varianza campionaria ci offrono una stima dei parametri media e varianza del campione statistico in esame. Abbiamo però bisogno di sapere *quanto ci si possa fidare di questa stima* ovvero quale sia la probabilità che il *vero* valore del parametro incognito non sia *troppo distante* dalla stima trovata.

Diamo perciò la seguente definizione:

Definizione 6.0.1 (Intervallo di confidenza)**.** Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico e sia θ un parametro (ignoto) che caratterizza la distribuzione del campione.
Siano $L_i = l_i(X_1,\ X_2,\ \ldots,\ X_n)$ e $L_s = l_s(X_1,\ X_2,\ \ldots,\ X_n)$ due statistiche del campione e sia $\alpha \in (0,1)$. Dico che l'intervallo (L_i, L_s) è un *intervallo di confidenza* (o di fiducia) di livello $1-\alpha$ se $\mathbb{P}(\theta \in (L_i, L_s)) \geq 1-\alpha$, ovvero che (L_i, L_s) è un intervallo di confidenza (o di fiducia) di errore α se $\mathbb{P}(\theta \notin (L_i, L_s)) \leq \alpha$.

Dico che la semiretta $(L_i, +\infty)$ è un *intervallo di confidenza unilaterale superiore* di livello $1-\alpha$ se $\mathbb{P}(\theta > L_i) \geq 1-\alpha$

Dico che la semiretta $(-\infty, L_s)$ è un *intervallo di confidenza unilaterale inferiore* di livello $1-\alpha$ se $\mathbb{P}(\theta < L_s) \geq 1-\alpha$

Osservazione 6.0.3.

1. La scelta dei nomi delle due statistiche non è casuale: L_i sta per limitazione inferiore mentre L_s sta per limitazione superiore.
2. Di solito si è interessati a *piccoli* valori di α, più precisamente a $\alpha \in (10^{-2}, 10^{-1})$.
3. La disuguaglianza di Chebyshev ci ha fornito un intervallo di confidenza per la media μ del campione nel caso in cui la varianza σ^2 sia nota

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| \geq t\right) \leq \frac{\sigma^2}{t^2} \qquad \forall t > 0$$

ovvero

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| < t\right) \geq 1 - \frac{\sigma^2}{t^2} \qquad \forall t > 0$$

cioè

$$\mathbb{P}\left(\overline{X} - t < \mu < \overline{X} + t\right) \geq 1 - \frac{\sigma^2}{t^2} \qquad \forall t > 0.$$

Fissato $\alpha \in (0,1)$ scelgo $t = \dfrac{\sigma}{\sqrt{\alpha}}$. La disuguaglianza di Chebyshev si legge allora

$$\mathbb{P}\left(\overline{X} - \frac{\sigma}{\sqrt{\alpha}} < \mu < \overline{X} + \frac{\sigma}{\sqrt{\alpha}}\right) \geq 1 - \alpha \qquad \forall \alpha \in (0,1).$$

Dunque l'intervallo $\left(\overline{X} - \frac{\sigma}{\sqrt{\alpha}}, \overline{X} + \frac{\sigma}{\sqrt{\alpha}}\right)$ è un intervallo di confidenza di livello $1-\alpha$ per la media μ del campione.

6.1. Stima per intervalli della media di campioni gaussiani

6.1.1. Campione gaussiano di cui è nota la varianza

Intervallo bilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ incognita e varianza σ^2 nota.

Sia Z una v.a. gaussiana standard e sia $\alpha \in (0,1)$. Calcolo $\mathbb{P}\left(|Z| \le z_{1-\frac{\alpha}{2}}\right)$:

$$\begin{aligned}
&\mathbb{P}\left(|Z| \le z_{1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \le Z \le z_{1-\frac{\alpha}{2}}\right) \\
&= \mathbb{P}\left(Z \le z_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(Z \le -z_{1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(Z \le z_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(Z \le z_{\frac{\alpha}{2}}\right) \qquad (6.1)\\
&= \Phi\left(z_{1-\frac{\alpha}{2}}\right) - \Phi\left(z_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha.
\end{aligned}$$

Sappiamo che $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ e che dunque $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$. Applichiamo quindi la disuguaglianza (6.1) a $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$. Si ha:

$$\begin{aligned}
1-\alpha &= \mathbb{P}\left(\frac{\left|\overline{X}-\mu\right|}{\frac{\sigma}{\sqrt{n}}} \le z_{1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \le \frac{\mu-\overline{X}}{\frac{\sigma}{\sqrt{n}}} \le z_{1-\frac{\alpha}{2}}\right) \\
&= \mathbb{P}\left(\frac{-\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \le \mu - \overline{X} \le \frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right) \qquad (6.2)\\
&= \mathbb{P}\left(\overline{X} - \frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \le \mu \le \overline{X} + \frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right)
\end{aligned}$$

L'intervallo

$$\left(\overline{X} - \frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}}, \overline{X} + \frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right)$$

è dunque un intervallo di confidenza di livello $1-\alpha$ per la media μ del campione.

Osservazione 6.1.1 (Dimensionamento del campione)**.** Fissato il livello di confidenza $1-\alpha$, supponiamo di voler controllare l'ampiezza dell'intervallo di confidenza $L_s - L_i$. Nel caso in esame l'ampiezza dell'intervallo di confidenza è $\frac{2\,\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}}$. Se fissiamo una limitazione superiore $2\,\delta$ per l'ampiezza di tale intervallo, deve dunque essere

$$\frac{2\,\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \le 2\,\delta$$

ovvero

$$n \ge \left(\frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\delta}\right)^2.$$

Intervallo unilaterale superiore

Sia $Z \sim \mathcal{N}(0,1)$. Sappiamo che

$$\mathbb{P}(Z \leq t) = 1 - \alpha \quad \text{se e solo se} \quad t = z_{1-\alpha}.$$

Abbiamo dunque

$$1 - \alpha = \mathbb{P}\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\alpha}\right) = \mathbb{P}\left(\overline{X} - \mu \leq \frac{\sigma\, z_{1-\alpha}}{\sqrt{n}}\right) = \mathbb{P}\left(\mu \geq \overline{X} - \frac{\sigma\, z_{1-\alpha}}{\sqrt{n}}\right).$$

Quindi la semiretta

$$\left(\overline{X} - \frac{\sigma\, z_{1-\alpha}}{\sqrt{n}}, +\infty\right)$$

è un intervallo di confidenza unilaterale superiore di livello $1 - \alpha$.

Intervallo unilaterale inferiore

Sia $Z \sim \mathcal{N}(0,1)$. Sappiamo che

$$\mathbb{P}(Z \geq t) = 1 - \alpha \quad \text{se e solo se} \quad \mathbb{P}(Z \leq t) = \alpha \quad \text{se e solo se} \quad t = z_{\alpha}.$$

Abbiamo dunque

$$1 - \alpha = \mathbb{P}\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \geq z_{\alpha}\right) = \mathbb{P}\left(\overline{X} - \mu \geq \frac{\sigma\, z_{\alpha}}{\sqrt{n}}\right) = \mathbb{P}\left(\mu \leq \overline{X} - \frac{\sigma\, z_{\alpha}}{\sqrt{n}}\right).$$

Quindi la semiretta

$$\left(-\infty, \overline{X} - \frac{\sigma\, z_{\alpha}}{\sqrt{n}}\right) = \left(-\infty, \overline{X} + \frac{\sigma\, z_{1-\alpha}}{\sqrt{n}}\right)$$

è un intervallo di confidenza unilaterale inferiore di livello $1 - \alpha$.

6.1.2. Campione gaussiano di cui non è nota la varianza

Intervallo bilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ varianza σ^2, entrambe incognite.

Sappiamo che la v.a. $T := \dfrac{(\overline{X} - \mu)\sqrt{n}}{S}$ segue la distribuzione t di Student con $n - 1$ gradi di libertà:

$$T \sim t(n-1).$$

Sia $t_{n-1,1-\frac{\alpha}{2}}$ il relativo quantile di livello $1 - \frac{\alpha}{2}$:

$$\mathbb{P}\left(T \leq t_{n-1,1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}.$$

Calcolo $\mathbb{P}\left(|T| \leq t_{n-1,1-\frac{\alpha}{2}}\right)$:

$$\begin{aligned}\mathbb{P}\left(|T| \leq t_{n-1,1-\frac{\alpha}{2}}\right) &= \mathbb{P}\left(-t_{n-1,1-\frac{\alpha}{2}} \leq T \leq t_{n-1,1-\frac{\alpha}{2}}\right)\\ &= \mathbb{P}\left(T \leq t_{n-1,1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(T \leq -t_{n-1,1-\frac{\alpha}{2}}\right)\\ &= \mathbb{P}\left(T \leq t_{n-1,1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(T \leq t_{n-1,\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha.\end{aligned}$$

Abbiamo dunque

$$\begin{aligned}1-\alpha = \mathbb{P}\left(|T| \leq t_{n-1,1-\frac{\alpha}{2}}\right) &= \mathbb{P}\left(\frac{\left|\overline{X}-\mu\right|\sqrt{n}}{S} \leq t_{n-1,1-\frac{\alpha}{2}}\right)\\ &= \mathbb{P}\left(\left|\overline{X}-\mu\right| \leq \frac{S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right)\\ &= \mathbb{P}\left(\frac{-S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} \leq \mu - \overline{X} \leq \frac{S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right)\\ &= \mathbb{P}\left(\overline{X} - \frac{S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} \leq \mu \leq \overline{X} + \frac{S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right)\end{aligned}$$

L'intervallo

$$\left(\overline{X} - \frac{S\, t_{n-1,1-\frac{\alpha}{2}}}{\sqrt{n}}, \overline{X} + \frac{S\, t_{n-1,1-\frac{\alpha}{2}}}{\sqrt{n}}\right)$$

è dunque un intervallo di confidenza di livello $1-\alpha$ per la media μ del campione.

Intervallo unilaterale superiore

Sappiamo che

$$\mathbb{P}\left(T \leq t\right) = 1-\alpha \quad \text{ se e solo se } \quad t = t_{n-1,1-\alpha}.$$

Abbiamo dunque

$$\begin{aligned}1-\alpha = \mathbb{P}\left(\frac{(\overline{X}-\mu)\sqrt{n}}{S} \leq t_{n-1,1-\alpha}\right) &= \mathbb{P}\left(\overline{X} - \mu \leq \frac{S\, t_{n-1,1-\alpha}}{\sqrt{n}}\right)\\ &= \mathbb{P}\left(\mu \geq \overline{X} - \frac{S\, t_{n-1,1-\alpha}}{\sqrt{n}}\right).\end{aligned}$$

Quindi la semiretta

$$\left(\overline{X} - \frac{S\, t_{n-1,1-\alpha}}{\sqrt{n}}, +\infty\right)$$

è un intervallo di confidenza unilaterale superiore di livello $1-\alpha$.

Intervallo unilaterale inferiore

Sappiamo che

$$\mathbb{P}(T \geq t) = 1 - \alpha \quad \text{se e solo se} \quad \mathbb{P}(T \leq t) = \alpha \quad \text{se e solo se} \quad t = t_{n-1,\alpha}.$$

Abbiamo dunque

$$1 - \alpha = \mathbb{P}\left(\frac{(\overline{X} - \mu)\sqrt{n}}{S} \geq t_{n-1,\alpha}\right) = \mathbb{P}\left(\overline{X} - \mu \geq \frac{S\, t_{n-1,\alpha}}{\sqrt{n}}\right) = \mathbb{P}\left(\mu \leq \overline{X} - \frac{S\, t_{n-1,\alpha}}{\sqrt{n}}\right).$$

Quindi la semiretta

$$\left(-\infty, \overline{X} - \frac{S\, t_{n-1,\alpha}}{\sqrt{n}}\right) = \left(-\infty, \overline{X} + \frac{S\, t_{n-1,1-\alpha}}{\sqrt{n}}\right)$$

è un intervallo di confidenza unilaterale inferiore di livello $1 - \alpha$.

6.2. Stima per intervalli della varianza di campioni gaussiani

Intervallo bilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ (incognita o nota)e varianza σ^2 incognita.

Sappiamo che la v.a. $V := (n-1)\dfrac{S^2}{\sigma^2}$ segue la distribuzione χ^2 a $n-1$ gradi di libertà. Per ogni $\alpha \in (0,1)$ indico con $\chi^2_{n-1,\alpha}$ il quantile di livello α della v.a. V:

$$F_V\left(\chi^2_{n-1,\alpha}\right) = \alpha \qquad \forall \alpha \in (0,1).$$

Osservazione 6.2.1. $\chi^2_{n-1,\alpha} > 0$ per ogni $\alpha \in (0,1)$.

Calcolo $\mathbb{P}\left(\chi^2_{n-1,\frac{\alpha}{2}} < V < \chi^2_{n-1,1-\frac{\alpha}{2}}\right)$:

$$\mathbb{P}\left(\chi^2_{n-1,\frac{\alpha}{2}} < V < \chi^2_{n-1,1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(V < \chi^2_{n-1,1-\frac{\alpha}{2}}\right) - \\ - \mathbb{P}\left(V < \chi^2_{n-1,\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.$$

Dunque

$$1 - \alpha = \mathbb{P}\left(\chi^2_{n-1,\frac{\alpha}{2}} < (n-1)\frac{S^2}{\sigma^2} < \chi^2_{n-1,1-\frac{\alpha}{2}}\right)$$
$$= \mathbb{P}\left(\frac{1}{\chi^2_{n-1,1-\frac{\alpha}{2}}} < \frac{\sigma^2}{(n-1)S^2} < \frac{1}{\chi^2_{n-1,\frac{\alpha}{2}}}\right) = \mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right)$$

Quindi l'intervallo

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right)$$

è un intervallo di confidenza di livello $1 - \alpha$ per la varianza σ^2 del campione.

Intervallo unilaterale superiore

Sappiamo che

$$\mathbb{P}(V \leq t) = 1 - \alpha \quad \text{se e solo se} \quad t = \chi^2_{n-1,1-\alpha}.$$

Dunque

$$1 - \alpha = \mathbb{P}\left((n-1)\frac{S^2}{\sigma^2} < \chi^2_{n-1,1-\alpha}\right) = \mathbb{P}\left(\sigma^2 > (n-1)\frac{S^2}{\chi^2_{n-1,1-\alpha}}\right).$$

Quindi la semiretta

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha}}, +\infty\right)$$

è un intervallo di confidenza di livello $1-\alpha$ per la varianza σ^2 del campione.

Intervallo unilaterale inferiore

Sappiamo che

$$\mathbb{P}(V \geq t) = 1 - \alpha \quad \text{se e solo se} \quad \mathbb{P}(V \leq t) = \alpha \quad \text{se e solo se} \quad t = \chi^2_{n-1,\alpha}.$$

Dunque

$$1 - \alpha = \mathbb{P}\left((n-1)\frac{S^2}{\sigma^2} > \chi^2_{n-1,\alpha}\right) = \mathbb{P}\left(\sigma^2 \leq (n-1)\frac{S^2}{\chi^2_{n-1,\alpha}}\right).$$

Quindi l'intervallo

$$\left(0, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha}}\right)$$

è un intervallo di confidenza di livello $1-\alpha$ per la varianza σ^2 del campione.

Esempio 6.2.1. Calcoliamo gli intervalli di confidenza per il carattere CO2SBW dei dati tratti da [**?**], nell'ipotesi che si tratti della realizzazione di v.a. normali.

```
> 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

> numSummary(X[,c("CO2SBW", "Densi", "FirTemp", "PRA", "PV", "TenStr", "Totpor")],
statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
               mean         sd      0%       25%      50%       75%    100%  n
CO2SBW    0.5816667  0.5259152   0.050   0.29000   0.3900   0.49500   1.960 30
Densi     1.6993333  0.1691548   1.340   1.57250   1.6950   1.83250   2.020 30
FirTemp 763.6666667 53.4649955 730.000 740.00000 740.0000 750.00000 960.000 30
PRA       0.6629000  0.4806106   0.158   0.42200   0.5825   0.72700   2.657 30
PV       56.7466667 27.9061201  10.200  42.42500  61.3000  80.75000  88.600 30
TenStr    0.6186000  0.3153048   0.143   0.42075   0.5545   0.72725   1.405 30
Totpor   39.9333333  7.0795326  26.850  36.04750  40.8600  44.02250  54.640 30

> ## definisco la funzione che calcola l'intervallo bilaterale con varianza nota

> bilat.norm = function(x,sigma,conf) { n = length(x); xbar=mean(x);
+ alpha = 1 - conf;
+ zstar = qnorm(1-alpha/2);
+ SE = sigma/sqrt(n);
+ xbar + c(-zstar*SE,zstar*SE)
+ }

> bilat.norm(X[,c("CO2SBW")],1,.9) ## supponiamo deviazione standard = 1
[1] 0.2813589 0.8819745
```

```
> bilat.norm(X[,c("CO2SBW")],2,.9) ## supponiamo deviazione standard = 2
[1] -0.01894896  1.18228229

> bilat.norm(X[,c("CO2SBW")],1,.95) ## supponiamo deviazione standard = 1
[1] 0.2238278 0.9395055

> bilat.norm(X[,c("CO2SBW")],2,.95) ## supponiamo deviazione standard = 2
[1] -0.134011  1.297344

> ## definisco la funzione che calcola l'intervallo bilaterale con varianza ignota

> bilat.stud = function(x,conf) { n = length(x); m = n-1; xbar=mean(x);
+ alpha = 1 - conf;
+ zstar = qt(1-alpha/2, m, lower.tail=TRUE);
+ SE = sd(x)/sqrt(n);
+ xbar + c(-zstar*SE,zstar*SE)
+ }

> bilat.stud(X[,c("CO2SBW")],.9)
[1] 0.4185190 0.7448144

> bilat.stud(X[,c("CO2SBW")],.95)
[1] 0.3852867 0.7780466

> ## definisco la funzione che calcola l'intervallo bilaterale per la varianza

> bilat.chi = function(x,conf) { n = length(x); m = n-1;
+ alpha = 1 - conf;
+ zsup = qchisq(alpha/2, m, lower.tail=TRUE);
+ zinf = qchisq(1 - alpha/2, m, lower.tail=TRUE);
+ SE = sd(x)*sd(x)*m;
+ c(SE/zinf,SE/zsup)
+ }

> bilat.chi(X[,c("CO2SBW")],.9)
[1] 0.1884772 0.4529507

> bilat.chi(X[,c("CO2SBW")],.95)
[1] 0.175429 0.499843
```

7. Test d'ipotesi

Un tipico problema che ci si può trovare ad affrontare è il seguente:

Faccio una certa ipotesi (che indico con H_0 e che chiamo **ipotesi nulla**). In base ai dati che ho a disposizione devo decidere se accettare o rifiutare la verità di questa ipotesi.

Si potranno verificare quattro situazioni alternative:

1. L'ipotesi è vera e l'accetto $\rightarrow$ bene
2. L'ipotesi è vera ma in base ai dati la rifiuto $\rightarrow$ in questo caso si dice che si commette **errore di prima specie**
3. L'ipotesi è falsa ma in base ai dati la accetto $\rightarrow$ in questo caso si dice che si commette **errore di seconda specie**
4. L'ipotesi è falsa e la rifiuto $\rightarrow$ bene

Per chiarirsi le idee vediamo prima un esempio.

Esempio 7.0.2. Ho una moneta. Voglio verificare se è bilanciata o meno. La lancio n volte. Pongo $X_i = \begin{cases} 1 & \text{se all'}i\text{-esimo lancio esce testa,} \\ 0 & \text{se all'}i\text{-esimo lancio esce croce.} \end{cases}$, $i = 1, \ldots, n$. Ho un campione statistico bernoulliano di numerosità n e parametro $p \in [0, 1]$ incognito, dove p è la probabilità che esca testa in un singolo lancio.

L'ipotesi nulla che dobbiamo testare è

$$H_0) \qquad p = 0.5.$$

Facciamo dunque n lanci. Otteniamo k teste ed $n - k$ croci:

$$x_1,\ x_2,\ \ldots,\ x_n \qquad \text{dove} \quad x_i = \begin{cases} 1 & \text{se all'}i\text{-esimo lancio esce testa,} \\ 0 & \text{se all'}i\text{-esimo lancio esce croce.} \end{cases}$$

e dunque $\overline{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i = \dfrac{k}{n}$.

Stabilisco una distanza massima ε tra $\overline{x}$ e 0.5 entro la quale accettare l'ipotesi $p = 0.5$ e oltre la quale rifiutarla. Ovvero: accetto H_0 se $|\overline{x} - 0.5| < \varepsilon$ e la rifiuto se $|\overline{x} - 0.5| \geq \varepsilon$. cioè se $\left| \displaystyle\sum_{i=1}^{n} x_i - \dfrac{n}{2} \right| \geq n\varepsilon$. Quanto vale la probabilità di commettere errore di prima specie, ovvero di rifiutarla quando esse invece è vera?

Commetto errore di prima specie con probabilità

$$\alpha := \mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - \frac{n}{2}\right| \geq n\varepsilon\right).$$

Poiché le v.a. X_i sono indipendenti e seguono tutte la distribuzione di Bernoulli di parametro p, la v.a. $Y := \sum_{i=1}^{n} X_i$ è una v.a. binomiale di parametri n e p. Se l'ipotesi H_0 è vera, allora $p = 0.5$ dunque : $Y \sim B(n, 0.5)$ e

$$\alpha := \mathbb{P}\left(\left|Y - \frac{n}{2}\right| \geq n\varepsilon\right) = \mathbb{P}\left(Y \geq \frac{n}{2} + n\varepsilon\right) + \mathbb{P}\left(Y \leq \frac{n}{2} - n\varepsilon\right)$$

Vediamo alcuni casi

- $n = 50$, $\varepsilon = 0.1$

$$\alpha = \mathbb{P}\left(Y \geq 25 + 5\right) + \mathbb{P}\left(Y \leq 25 - 5\right) = 1 - F_Y(29) + F_Y(20).$$

```
> 1 - pbinom(c(29), size=50, prob=0.5, lower.tail=TRUE)
+ pbinom(c(20), size=50, prob=0.5, lower.tail=TRUE)
[1] 0.2026388
```

- $n = 100$, $\varepsilon = 0.1$

$$\alpha = \mathbb{P}\left(Y \geq 50 + 10\right) + \mathbb{P}\left(Y \leq 50 - 10\right) = 1 - F_Y(59) + F_Y(40).$$

```
> 1 - pbinom(c(59), size=100, prob=0.5, lower.tail=TRUE)
+ pbinom(c(40), size=100, prob=0.5, lower.tail=TRUE)
[1] 0.05688793
```

- $n = 200$, $\varepsilon = 0.1$

$$\alpha = \mathbb{P}\left(Y \geq 100 + 20\right) + \mathbb{P}\left(Y \leq 100 - 20\right) = 1 - F_Y(119) + F_Y(80).$$

```
> 1 - pbinom(c(119), size=200, prob=0.5, lower.tail=TRUE)
+ pbinom(c(80), size=200, prob=0.5, lower.tail=TRUE)
[1] 0.005685156
```

- $n = 300$, $\varepsilon = 0.1$

$$\alpha = \mathbb{P}\left(Y \geq 150 + 30\right) + \mathbb{P}\left(Y \leq 150 - 30\right) = 1 - F_Y(179) + F_Y(120).$$

```
> 1 - pbinom(c(179), size=300, prob=0.5, lower.tail=TRUE)
+ pbinom(c(120), size=300, prob=0.5, lower.tail=TRUE)
[1] 0.00063422
```

- $n = 50$, $\varepsilon = 0.05$

$$\alpha = \mathbb{P}\left(Y \geq 25 + 2.5\right) + \mathbb{P}\left(Y \leq 25 - 2.5\right) = 1 - F_Y(27) + F_Y(22).$$

```
> 1 - pbinom(c(27), size=50, prob=0.5, lower.tail=TRUE)
+ pbinom(c(22), size=50, prob=0.5, lower.tail=TRUE)
[1] 0.4798877
```

- $n = 100$, $\varepsilon = 0.05$

$$\alpha = \mathbb{P}\left(Y \geq 50 + 5\right) + \mathbb{P}\left(Y \leq 50 - 5\right) = 1 - F_Y(54) + F_Y(45).$$

```
> 1 - pbinom(c(54), size=100, prob=0.5, lower.tail=TRUE)
+ pbinom(c(45), size=100, prob=0.5, lower.tail=TRUE)
[1] 0.3682016
```

- $n = 200$, $\varepsilon = 0.05$

$$\alpha = \mathbb{P}\left(Y \geq 100 + 10\right) + \mathbb{P}\left(Y \leq 100 - 10\right) = 1 - F_Y(109) + F_Y(90).$$

```
> 1 - pbinom(c(109), size=200, prob=0.5, lower.tail=TRUE)
+ pbinom(c(90), size=200, prob=0.5, lower.tail=TRUE)
[1] 0.178964
```

- $n = 300$, $\varepsilon = 0.05$

$$\alpha = \mathbb{P}\left(Y \geq 150 + 15\right) + \mathbb{P}\left(Y \leq 150 - 15\right) = 1 - F_Y(164) + F_Y(135).$$

```
> 1 - pbinom(c(164), size=300, prob=0.5, lower.tail=TRUE)
+ pbinom(c(135), size=300, prob=0.5, lower.tail=TRUE)
[1] 0.0939037
```

- $n = 400$, $\varepsilon = 0.05$

$$\alpha = \mathbb{P}\left(Y \geq 200 + 20\right) + \mathbb{P}\left(Y \leq 200 - 20\right) = 1 - F_Y(219) + F_Y(180).$$

```
> 1 - pbinom(c(219), size=400, prob=0.5, lower.tail=TRUE)
+ pbinom(c(180), size=400, prob=0.5, lower.tail=TRUE)
[1] 0.05104022
```

- $n = 500$, $\varepsilon = 0.05$

$$\alpha = \mathbb{P}\left(Y \geq 250 + 25\right) + \mathbb{P}\left(Y \leq 250 - 25\right) = 1 - F_Y(264) + F_Y(225).$$

```
> 1 - pbinom(c(274), size=500, prob=0.5, lower.tail=TRUE)
+ pbinom(c(225), size=500, prob=0.5, lower.tail=TRUE)
[1] 0.02832616
```

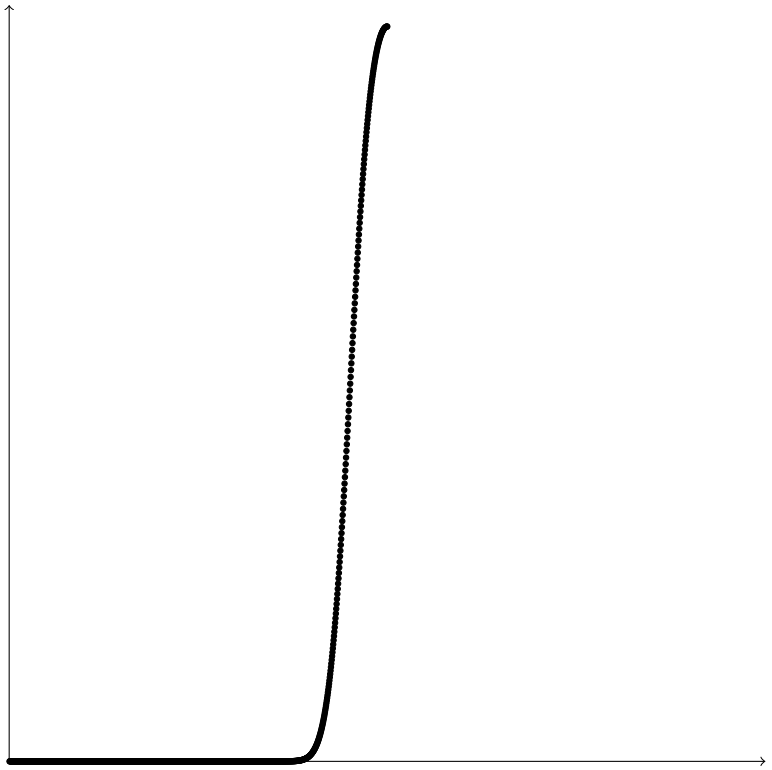

Figura 7.1: $\beta(p)$

Solitamente si vuole controllare (nel senso di tenere bassa, inferiore a 0.1 o a 0.05) la probabilità α di commettere errore di prima specie. Tale probabilità viene detta *livello di significatività* del test. Fissato il livello di significatività α, la numerosità n e la soglia di tolleranza ε andranno scelti di conseguenza come visto negli esempi precedenti.

Inoltre, fissato α, ci chiediamo quanto valga la probabilità di commettere errore di seconda specie, ovvero di accettare H_0 quand'essa invece è falsa.

Se H_0 è falsa, allora la probabilità di ottenere testa non è 0.5 ma assume un valore $p \neq 0.5$ (ignoto) e dunque $Y \sim B(n, p)$ e io accetto H_0 con probabilità

$$\beta(p) := \mathbb{P}_p\left(\left|Y - \frac{n}{2}\right| < n\varepsilon\right) = \mathbb{P}_p\left(Y < \frac{n}{2} + n\varepsilon\right) - \mathbb{P}_p\left(Y \leq \frac{n}{2} - n\varepsilon\right)$$

Si calcola $\beta(p)$ per vari valori di p. La funzione $\beta(p)$ è detta **curva operativa caratteristica (OC)** mentre $1 - \beta(p)$ cioè la probabilità di rifiutare H_0 quand'essa in effetti è falsa e il parametro incognito vale p, è detta **potenza del test**.

Esempio 7.0.3. Consideriamo la solita moneta e stavolta vogliamo vedere se è più probabile ottenere testa che ottenere croce. Vogliamo cioè testare l'ipotesi nulla

$$H_0) \qquad p \leq 0.5$$

Un test di questo tipo è detto *test unilaterale*.

Stabilisco una tolleranza massima ε entro la quale accettare l'ipotesi $p \leq 0.5$ e oltre la quale rifiutarla. Ovvero: accetto H_0 se $\overline{x} < 0.5 + \varepsilon$ e la rifiuto se $\overline{x} \geq 0.5 + \varepsilon$ cioè se $\sum_{i=1}^{n} x_i \geq \frac{n}{2} + n\varepsilon$. Quanto vale la probabilità di commettere errore di prima specie, ovvero di rifiutarla quando essa invece è vera?

Commetto errore di prima specie con probabilità

$$\alpha := \mathbb{P}\left(Y \geq \frac{n}{2} + n\varepsilon\right).$$

Se H_0 è vera, allora $Y \sim B(n, p)$ per qualche $p \leq 0.5$. Indico F_Y^p la sua funzione di ripartizione Vediamo alcuni casi

- $n = 50$, $\varepsilon = 0.1$

$$\alpha = 1 - \mathbb{P}\left(Y < 25 + 5\right) = 1 - F_Y^p(29) \leq \sup_{p \in [0, 0.5)} \left\{1 - F_Y^p(29)\right\} = 1 - F_Y^{0.5}(29)$$

```
> 1 - pbinom(c(29), size=50, prob=0.5, lower.tail=TRUE)
```

- $n = 100$, $\varepsilon = 0.1$

$$\alpha = 1 - \mathbb{P}\left(Y < 50 + 10\right) = 1 - F_Y^p(59) \leq \sup_{p \in [0, 0.5)} \left\{1 - F_Y^p(59)\right\} = 1 - F_Y^{0.5}(59).$$

```
> 1 - pbinom(c(59), size=100, prob=0.5, lower.tail=TRUE)
[1] 0.02844397
```

- $n = 200$, $\varepsilon = 0.1$

$$\alpha = 1 - \mathbb{P}\left(Y < 100 + 20\right) = 1 - F_Y^p(119) \leq \sup_{p \in [0, 0.5)} \left\{1 - F_Y^p(119)\right\} = 1 - F_Y^{0.5}(119).$$

```
> 1 - pbinom(c(119), size=200, prob=0.5, lower.tail=TRUE)
[1] 0.002842578
```

- $n = 50$, $\varepsilon = 0.05$

$$\alpha = 1 - \mathbb{P}\left(Y < 25 + 2.5\right) = 1 - F_Y^p(27) \leq \sup_{p \in [0, 0.5)} \left\{1 - F_Y^p(27)\right\} = 1 - F_Y^{0.5}(27)$$

```
> 1 - pbinom(c(27), size=50, prob=0.5, lower.tail=TRUE)
[1] 0.2399438
```

- $n = 100$, $\varepsilon = 0.05$

$$\alpha = 1 - \mathbb{P}\left(Y < 50 + 5\right) = 1 - F_Y^p(54) \leq \sup_{p \in [0, 0.5)} \left\{1 - F_Y^p(54)\right\} = 1 - F_Y^{0.5}(54)$$

```
> 1 - pbinom(c(55), size=100, prob=0.5, lower.tail=TRUE)
[1] 0.1841008
```

- $n = 200$, $\varepsilon = 0.05$

  $$\alpha = 1 - \mathbb{P}\left(Y < 100 + 10\right) = 1 - F_Y^p(109) \leq \sup_{p \in [0,0.5)} \left\{1 - F_Y^p(109)\right\} = 1 - F_Y^{0.5}(109).$$

```
> 1 - pbinom(c(109), size=200, prob=0.5, lower.tail=TRUE)
[1] 0.08948202
```

- $n = 300$, $\varepsilon = 0.05$

  $$\alpha = 1 - \mathbb{P}\left(Y < 150 + 15\right) = 1 - F_Y^p(164) \leq \sup_{p \in [0,0.5)} \left\{1 - F_Y^p(164)\right\} = 1 - F_Y^{0.5}(145).$$

```
> 1 - pbinom(c(164), size=300, prob=0.5, lower.tail=TRUE)
[1] 0.04695185
```

In generale dunque un test d'ipotesi ha la seguente struttura:

1. Si ha un campione statistico $X_1,\ X_2,\ \ldots,\ X_n$, $X_i \sim \mathcal{D}(\theta)$, dove θ è un parametro reale.

2. Si formula un'ipotesi (che si chiama *ipotesi nulla* e si indica con H_0), solitamente nella forma

   $$H_0) \qquad \theta \in \Theta_0$$

   dove Θ_0 è un sottoinsieme di $\mathbb{R}$.

3. Si formula una regola di decisione per l'accettazione o il rifiuto di H_0. La regola di decisione è di questo tipo: si sceglie una statistica che fornisce una stima del parametro θ e un sottoinsieme $A \subset \mathbb{R}$, detto *regione di accettazione*. Dopodiché

   - se $Y \in A$, allora si accetta H_0;
   - se $Y \notin A$, allora si rifiuta H_0.

   L'insieme $A^c := \mathbb{R} \setminus A$ è detto *regione di rifiuto*.

Come già detto, è solitamente richiesto di *limitare* la probabilità di commettere errore di prima specie, cioè di limitare la probabilità di rifiutare l'ipotesi nulla quando essa è vera. Vediamo come questo sia possibile nel caso di campioni gaussiani.

7.1. Test d'ipotesi per la media di campioni gaussiani

7.1.1. Campione gaussiano di cui è nota la varianza

Test bilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ incognita e varianza σ^2 nota. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \mu = \mu_0.$$

H_0 è vera se e solo se $\mathbb{E}\left[\overline{X}\right] = 0$ dunque accetto l'ipotesi nulla H_0 se la media campionaria si discosta da μ_0 per meno di un valore soglia ε ovvero se $|\overline{x} - \mu_0| < \varepsilon$ e la rifiuto altrimenti.

Il livello di di significatività (cioè la probabilità di commettere un errore di prima specie) è allora

$$\alpha = \mathbb{P}_{\mu_0}\left(\left|\overline{X} - \mu_0\right| \geq \varepsilon\right)$$

dove il pedice μ_0 indica che H_0 è vera, cioè che $\mu = \mu_0$.

Se H_0 è vera, $\overline{X} \sim \mathcal{N}\left(\mu_0, \dfrac{\sigma^2}{n}\right)$ e $Z := \dfrac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$. Dunque

$$\begin{aligned}
\alpha &= \mathbb{P}_{\mu_0}\left(\left|\overline{X} - \mu_0\right| \geq \varepsilon\right) = \mathbb{P}_{\mu_0}\left(\frac{\left|\overline{X} - \mu_0\right|}{\frac{\sigma}{\sqrt{n}}} \geq \frac{\varepsilon}{\frac{\sigma}{\sqrt{n}}}\right) = \mathbb{P}\left(|Z| \geq \frac{\varepsilon\sqrt{n}}{\sigma}\right) \\
&= \mathbb{P}\left(Z \geq \frac{\varepsilon\sqrt{n}}{\sigma}\right) + \mathbb{P}\left(Z \leq \frac{-\varepsilon\sqrt{n}}{\sigma}\right) = 1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) + \Phi\left(\frac{-\varepsilon\sqrt{n}}{\sigma}\right) \\
&= 2\left(1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)\right)
\end{aligned}$$

Se voglio fissare a priori α, deve essere allora $\Phi\left(\dfrac{\varepsilon\sqrt{n}}{\sigma}\right) = 1 - \dfrac{\alpha}{2}$ cioè deve essere $\dfrac{\varepsilon\sqrt{n}}{\sigma} = z_{1-\frac{\alpha}{2}}$ e dunque devo scegliere

$$\varepsilon = \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}.$$

Presi i dati $x_1,\ x_2,\ \ldots,\ x_n$, sia dunque $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ la loro media. Accetto H_0 se

$$|\overline{x} - \mu_0| < \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$$

e la rifiuto altrimenti.

Calcoliamo la *curva operativa caratteristica*. Se H_0 è falsa, $\mu \neq \mu_0$, commetto errore di seconda specie con probabilità

$$\begin{aligned}\beta(\mu) &= \mathbb{P}_\mu\left(\left|\overline{X}-\mu_0\right| < \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right) = \mathbb{P}_\mu\left(\mu_0 - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} < \overline{X} < \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right) \\ &= \mathbb{P}_\mu\left(\frac{\mu_0-\mu}{\frac{\sigma}{\sqrt{n}}} - z_{1-\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < \frac{\mu_0-\mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}}\right) \qquad (7.1) \\ &= \Phi\left(\frac{\mu_0-\mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}}\right) - \Phi\left(\frac{\mu_0-\mu}{\frac{\sigma}{\sqrt{n}}} + z_{\frac{\alpha}{2}}\right).\end{aligned}$$

Distinguiamo due casi

1. $\mu > \mu_0$
   In questo caso $\dfrac{\mu_0-\mu}{\frac{\sigma}{\sqrt{n}}} < 0$ dunque $\dfrac{\mu_0-\mu}{\frac{\sigma}{\sqrt{n}}} + z_{\frac{\alpha}{2}} < z_{\frac{\alpha}{2}}$ e quindi
   $$0 < \Phi\left(\frac{\mu_0-\mu}{\frac{\sigma}{\sqrt{n}}} + z_{\frac{\alpha}{2}}\right) < \frac{\alpha}{2}$$
   e la possiamo considerare una quantità trascurabile.
   Supponiamo di voler fissare (oltre ad α) anche $\beta(\mu) = \hat{\beta}$. Con la semplificazione fatta dalla (7.1) otteniamo $\hat{\beta} \geq \Phi\left(\dfrac{\mu_0-\mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}}\right)$. L'unica quantità che possiamo trattare è la numerosità n. Risolvendo l'equazione rispetto a n otteniamo
   $$z_{\hat{\beta}} \geq \frac{\mu_0-\mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}}$$
   e dunque
   $$\frac{\mu_0-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\hat{\beta}} + z_{\frac{\alpha}{2}},$$
   cioè
   $$n \geq \left(\frac{\sigma}{\mu_0-\mu}\right)^2 \left(z_{\hat{\beta}} + z_{\frac{\alpha}{2}}\right)^2$$

2. $\mu < \mu_0$
   In questo caso $\dfrac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}} < 0$ e scriviamo la (7.1) nella forma
   $$\begin{aligned}\beta(\mu) &= \Phi\left(\frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}} - z_{\frac{\alpha}{2}}\right) - \Phi\left(\frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}} - z_{1-\frac{\alpha}{2}}\right) \\ &= \Phi\left(\frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}}\right) - \Phi\left(\frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}} + z_{\frac{\alpha}{2}}\right).\end{aligned}$$

Si ha $\frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}} + z_{\frac{\alpha}{2}} < z_{\frac{\alpha}{2}}$ e dunque

$$0 < \Phi\left(\frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}} + z_{\frac{\alpha}{2}}\right) < \frac{\alpha}{2}$$

e la possiamo considerare una quantità trascurabile.

Supponiamo di voler fissare (oltre ad α) anche $\beta(\mu) = \hat{\beta}$. Con la semplificazione fatta possiamo considerare l'equazione $\hat{\beta} \geq \Phi\left(\frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}}\right)$ e ritroviamo la disuguaglianza trovata nel caso precedente:

$$n \geq \left(\frac{\sigma}{\mu_0-\mu}\right)^2 \left(z_{\hat{\beta}} + z_{\frac{\alpha}{2}}\right)^2$$

Test unilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ incognita e varianza σ^2 nota. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \mu \leq \mu_0.$$

Accetto l'ipotesi nulla H_0 se la media campionaria è inferiore a $\mu_0+\varepsilon$ cioè se $\overline{x} < \mu_0+\varepsilon$.

La probabilità di commettere un errore di prima specie è allora

$$\mathbb{P}_{\mu\leq\mu_0}\left(\overline{X} \geq \mu_0+\varepsilon\right).$$

dove il pedice indica che la media del campione è $\mu \leq \mu_0$.

Poiché $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ e $Z := \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$, si ha

$$\mathbb{P}_{\mu\leq\mu_o}\left(\overline{X} \geq \mu_0+\varepsilon\right) = \mathbb{P}_{\mu\leq\mu_0}\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \geq \frac{\mu_0-\mu+\varepsilon}{\frac{\sigma}{\sqrt{n}}}\right) = \mathbb{P}\left(Z \geq \frac{(\mu_0-\mu+\varepsilon)\sqrt{n}}{\sigma}\right)$$
$$= 1 - \mathbb{P}\left(Z \leq \frac{(\mu_0-\mu+\varepsilon)\sqrt{n}}{\sigma}\right) = 1 - \Phi\left(\frac{(\mu_0-\mu+\varepsilon)\sqrt{n}}{\sigma}\right) \leq 1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right).$$

Se voglio limitare superiormente $\mathbb{P}_{\mu\leq\mu_0}\left(\overline{X} > \mu_0+\varepsilon\right)$, cioè se voglio

$$\mathbb{P}_{\mu\leq\mu_0}\left(\overline{X} > \mu_0+\varepsilon\right) \leq \alpha \qquad \forall \mu \leq \mu_0$$

scelgo ε in modo da avere $1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) = \alpha$ cioè $\frac{\varepsilon\sqrt{n}}{\sigma} = z_{1-\alpha}$ e dunque scelgo

$$\varepsilon = \frac{\sigma}{\sqrt{n}} z_{1-\alpha}.$$

Presi i dati $x_1,\ x_2,\ \ldots,\ x_n$, sia dunque $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ la loro media. Accetto H_0 se

$$\overline{x} < \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$$

e la rifiuto altrimenti.

Test unilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ incognita e varianza σ^2 nota. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \mu \geq \mu_0.$$

Accetto l'ipotesi nulla H_0 se la media campionaria è superiore a $\mu_0 - \varepsilon$ cioè se $\overline{x} > \mu_0 - \varepsilon$. La probabilità di commettere un errore di prima specie è allora

$$\mathbb{P}_{\mu \geq \mu_0}\left(\overline{X} \leq \mu_0 - \varepsilon\right).$$

Poiché $\overline{X} \sim \mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$, e $Z := \dfrac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$, si ha

$$\mathbb{P}_{\mu \geq \mu_0}\left(\overline{X} \leq \mu_0 - \varepsilon\right) = \mathbb{P}_{\mu \geq \mu_0}\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\mu_0 - \mu - \varepsilon}{\frac{\sigma}{\sqrt{n}}}\right) = \mathbb{P}\left(Z \leq \frac{(\mu_0 - \mu - \varepsilon)\sqrt{n}}{\sigma}\right)$$
$$= \Phi\left(\frac{(\mu_0 - \mu - \varepsilon)\sqrt{n}}{\sigma}\right) \leq \Phi\left(\frac{-\varepsilon\sqrt{n}}{\sigma}\right) = 1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right).$$

Se voglio limitare superiormente $\mathbb{P}_{\mu \geq \mu_0}\left(\overline{X} \leq \mu_0 - \varepsilon\right)$ cioè se voglio

$$\mathbb{P}_{\mu \geq \mu_0}\left(\overline{X} \leq \mu_0 - \varepsilon\right) \leq \alpha \qquad \forall \mu \geq \mu_0$$

scelgo ε in modo da avere $\Phi\left(\dfrac{\varepsilon\sqrt{n}}{\sigma}\right) = 1 - \alpha$ cioè $\dfrac{\varepsilon\sqrt{n}}{\sigma} = z_{1-\alpha}$ e dunque scelgo

$$\varepsilon = \frac{\sigma}{\sqrt{n}} z_{1-\alpha}.$$

Presi i dati $x_1,\ x_2,\ \ldots,\ x_n$, sia dunque $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ la loro media. Accetto H_0 se

$$\overline{x} > \mu_0 - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$$

e la rifiuto altrimenti.

7.1.2. Campione gaussiano di cui non è nota la varianza

Test bilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ e varianza σ^2 entrambe ignote. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \mu = \mu_0.$$

Sappiamo che, se $\mu = \mu_0$, allora $T := \dfrac{(\overline{X} - \mu_0)\sqrt{n}}{S} \sim t(n-1)$.

Inoltre H_0 è vera se e solo se $\mathbb{E}\left[\overline{X}\right] = \mu_0$ ovvero, per l'indipendenza di $\overline{X}$ e S^2, se e solo se $\mathbb{E}\left[T\right] = 0$. Dunque accetto l'ipotesi nulla H_0 se $|T| \leq \varepsilon$.

Il livello di di significatività (cioè la probabilità di commettere un errore di prima specie) è allora

$$\alpha = \mathbb{P}\left(|T| \geq \varepsilon\right).$$

Si ha quindi

$$\begin{aligned}\alpha &= \mathbb{P}\left(|T| \geq \varepsilon\right) = \mathbb{P}\left(T \geq \varepsilon\right) + \mathbb{P}\left(T \leq -\varepsilon\right) \\ &= 1 - F_T\left(\varepsilon\right) + F_T\left(-\varepsilon\right) = 2\left(1 - F_T\left(\varepsilon\right)\right)\end{aligned}$$

Se voglio fissare a priori α, deve essere allora $F_T\left(\varepsilon\right) = 1 - \dfrac{\alpha}{2}$ dunque devo scegliere

$$\varepsilon = t_{n-1,1-\frac{\alpha}{2}}.$$

Presi i dati $x_1,\ x_2,\ \ldots,\ x_n$, sia dunque $t = \dfrac{(\overline{x} - \mu_0)\sqrt{n}}{\sqrt{s^2}}$ (dove $\overline{x}$ e s indicano la media e la deviazione campionaria del campione, rispettivamente). Accetto H_0 se

$$|t| \leq t_{n-1,1-\frac{\alpha}{2}}$$

e la rifiuto altrimenti, ovvero accetto H_0 se

$$\mu_0 - \frac{t_{n-1,1-\frac{\alpha}{2}}\, s}{\sqrt{n}} \leq \overline{x} \leq \mu_0 + \frac{t_{n-1,1-\frac{\alpha}{2}}\, s}{\sqrt{n}}$$

e la rifiuto altrimenti.

Test unilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ e varianza σ^2 entrambe incognite. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \mu \leq \mu_0.$$

Diamo la seguente regola di accettazione: accettiamo H_0 se $\dfrac{(\overline{X} - \mu_0)\sqrt{n}}{S} \leq \varepsilon$.

La probabilità di commettere un errore di prima specie è allora

$$\mathbb{P}_{\mu \leq \mu_0}\left(\frac{(\overline{X} - \mu_0)\sqrt{n}}{S} > \varepsilon\right).$$

Se H_0 è vera, allora $\mu \leq \mu_0$ e dunque

$$\frac{(\overline{X} - \mu_0)\sqrt{n}}{S} \leq \frac{(\overline{X} - \mu)\sqrt{n}}{S} =: T \sim t(n-1).$$

Di conseguenza

$$\left\{ \frac{(\overline{X} - \mu_0)\sqrt{n}}{S} > \varepsilon \right\} \subset \left\{ \frac{(\overline{X} - \mu)\sqrt{n}}{S} > \varepsilon \right\}$$

Dunque, per ogni $\mu \leq \mu_0$ si ha

$$\mathbb{P}_{\mu \leq \mu_0} \left(\frac{(\overline{X} - \mu_0)\sqrt{n}}{S} > \varepsilon \right) \leq \mathbb{P}_{\mu \leq \mu_0} \left(\frac{(\overline{X} - \mu)\sqrt{n}}{S} > \varepsilon \right) = \mathbb{P}\left(T > \varepsilon\right) = 1 - F_T(\varepsilon)$$

Se vogliamo stabilire il livello di significatività α dovremmo scegliere ε in modo che

$$1 - F_T(\varepsilon) = \alpha$$

cioè $\varepsilon = t_{n-1,1-\alpha}$.

Presi i dati $x_1,\ x_2,\ \ldots,\ x_n$, sia dunque $t_0 = \dfrac{(\overline{x} - \mu_0)\sqrt{n}}{s}$. Accetto H_0 se

$$t_0 \leq t_{n-1,1-\alpha}$$

ovvero se

$$\overline{x} \leq \mu_0 + \frac{t_{n-1,1-\alpha}\, s}{\sqrt{n}}$$

e la rifiuto altrimenti.

Test unilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ e varianza σ^2 entrambe incognite. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \mu \geq \mu_0.$$

Diamo la seguente regola di accettazione: accettiamo H_0 se $\dfrac{(\overline{X} - \mu_0)\sqrt{n}}{S} \geq -\varepsilon$.

La probabilità di commettere un errore di prima specie è allora

$$\mathbb{P}_{\mu \geq \mu_0} \left(\frac{(\overline{X} - \mu_0)\sqrt{n}}{S} < -\varepsilon \right).$$

Se H_0 è vera, allora $\mu \geq \mu_0$ e dunque

$$\frac{(\overline{X} - \mu_0)\sqrt{n}}{S} \geq \frac{(\overline{X} - \mu)\sqrt{n}}{S} =: T \sim t(n-1).$$

Di conseguenza

$$\left\{ \frac{(\overline{X} - \mu_0)\sqrt{n}}{S} < -\varepsilon \right\} \subset \left\{ \frac{(\overline{X} - \mu)\sqrt{n}}{S} < -\varepsilon \right\}$$

Dunque

$$\mathbb{P}_{\mu\geq\mu_0}\left(\frac{(\overline{X}-\mu_0)\sqrt{n}}{S}<-\varepsilon\right)\leq\mathbb{P}_{\mu\geq\mu_0}\left(\frac{(\overline{X}-\mu)\sqrt{n}}{S}<-\varepsilon\right)\qquad\forall\mu\geq\mu_0$$

e quindi, per ogni $\mu\geq\mu_0$ si ha

$$\begin{aligned}\mathbb{P}_{\mu\geq\mu_0}\left(\frac{(\overline{X}-\mu_0)\sqrt{n}}{S}<-\varepsilon\right)&=\mathbb{P}_{\mu\geq\mu_0}\left(\frac{(\overline{X}-\mu)\sqrt{n}}{S}<-\varepsilon\right)\\&=\mathbb{P}\left(T<-\varepsilon\right)=F_T(-\varepsilon)=1-F_T(\varepsilon).\end{aligned}$$

Se vogliamo stabilire il livello di significatività α dovremmo scegliere ε in modo che

$$1-F_T(\varepsilon)=\alpha$$

cioè $\varepsilon=t_{n-1,1-\alpha}$.

Presi i dati $x_1,\ x_2,\ \ldots,\ x_n$, sia dunque $t_0=\dfrac{(\overline{x}-\mu_0)\sqrt{n}}{s}$. Accetto H_0 se

$$t_0\geq-t_{n-1,1-\alpha}$$

e la rifiuto altrimenti, ovvero accetto H_0 se

$$\overline{x}\geq\mu_0-\frac{t_{n-1,1-\alpha}\,s}{\sqrt{n}}$$

e la rifiuto altrimenti.

7.2. Test d'ipotesi per l'uguaglianza di medie di campioni gaussiani

Supponiamo di avere due campioni, entrambi gaussiani

$$\begin{aligned}&X\colon X_1,\ X_2,\ \ldots,\ X_n\qquad&&X_i\sim\mathcal{N}\left(\mu_X,\sigma_X^2\right),\\&Y\colon Y_1,\ Y_2,\ \ldots,\ Y_k\qquad&&Y_j\sim\mathcal{N}\left(\mu_Y,\sigma_Y^2\right).\end{aligned}$$

Vogliamo testare l'ipotesi

$$H_0)\qquad\mu_X=\mu_Y$$

Osserviamo che $\mu_x=\mu_Y$ se e solo se $\mathbb{E}\left[\overline{X}-\overline{Y}\right]=0$.

Per limitare la probabilità di commettere errore di prima specie, distinguiamo tre diversi casi

7.2.1. Primo caso: le varianze σ_X^2 e σ_Y^2 sono note

Considero la v.a. $W:=\overline{X}-\overline{Y}$. Per le proprietà dei campioni gaussiani

$$W\sim\mathcal{N}\left(\mu_X-\mu_Y,\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{k}\right).$$

Dunque H_0 è vera se e solo se $W\sim\mathcal{N}\left(0,\dfrac{\sigma_X^2}{n}+\dfrac{\sigma_Y^2}{k}\right)$. Dunque stabilisco il seguente criterio di accettazione:

$$\text{Accetto } H_0 \text{ se e solo se } |w| = |\overline{x} - \overline{y}| < \varepsilon.$$

La probabilità di commettere errore di prima specie vale allora

$$\alpha = \mathbb{P}_{\mu_X=\mu_Y}(|W| \geq \varepsilon) = \mathbb{P}_{\mu_X=\mu_Y}\left(\frac{|W|}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{k}}} \geq \frac{\varepsilon}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{k}}}\right)$$

D'altra parte, se H_0 è vera, allora $Z := \dfrac{W}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{k}}} \sim \mathcal{N}(0,1)$, e dunque dovremo scegliere $\dfrac{\varepsilon}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{k}}} = z_{1-\frac{\alpha}{2}}$ ovvero

$$\varepsilon = z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{k}}.$$

Dunque accettiamo l'ipotesi H_0 se

$$|\overline{x} - \overline{y}| < z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{k}}$$

e la rifiutiamo altrimenti.

Osservazione 7.2.1. Se $\sigma_X^2 = \sigma_Y^2 = \sigma_0^2$ e $k = n$, allora $\varepsilon = z_{1-\frac{\alpha}{2}}\sigma_0\sqrt{\dfrac{2}{n}}$.

7.2.2. Secondo caso: le varianze σ_X^2 e σ_Y^2 sono ignote ma uguali

Consideriamo le due varianze campionarie

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad S_Y^2 = \frac{1}{k-1}\sum_{j=1}^{k}(Y_j - \overline{Y})^2.$$

Sia σ^2 il comune valore di σ_X^2 e σ_Y^2. Sappiamo che

$$V_X := \frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}, \quad V_Y := \frac{(k-1)S_Y^2}{\sigma^2} \sim \chi^2_{k-1}$$

dunque, per la Proprietà 5.3.2, $V_X + V_Y \sim \chi^2_{n-1+k-1} = \chi^2_{n+k-2}$

D'altra parte

$$V_X + V_Y = \frac{(n-1)S_X^2 + (k-1)S_Y^2}{\sigma^2} = \frac{n+k-2}{\sigma^2}\frac{(n-1)S_X^2 + (k-1)S_Y^2}{n+k-2}.$$

Consideriamo la statistica:

$$\overline{S}^2 := \frac{(n-1)S_X^2 + (k-1)S_Y^2}{n+k-2}.$$

Si ha

$$V_X + V_Y = \frac{(n+k-2)\overline{S}^2}{\sigma^2}. \tag{7.2}$$

Inoltre $\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{k}\right)\right)$, quindi

$$Z := \frac{\overline{X} - \overline{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{k}}} \sim \mathcal{N}(0,1) \quad \text{se e solo se} \quad \mu_X = \mu_Y.$$

Sia

$$T := \frac{Z\sqrt{n+k-2}}{\sqrt{V_X + V_Y}}.$$

Poiché i due campioni sono gaussiani e indipendenti le v.a. $\overline{X}$, S_X^2, $\overline{Y}$ e S_Y^2 sono indipendenti, quindi Z e $\overline{S}^2$ sono indipendenti, quindi $\mathbb{E}\left[\overline{X} - \overline{Y}\right] = 0$ se e solo se $\mathbb{E}[T] = 0$. Come criterio di accettazioneper l'ipotesi nulla H_0 scelgo dunque $|t| < \varepsilon$.

Inoltre, se $\mathbb{E}\left[\overline{X} - \overline{Y}\right] = 0$, allora per il Teorema 5.3.5

$$T \sim t(n+k-2).$$

Sostituendo l'espressione per Z e quella per $V_X + V_Y$ data nell' equazione (7.2) si ha

$$T = \frac{\overline{X}}{\overline{S}\sqrt{\frac{1}{n} + \frac{1}{k}}} \sim t(n+k-2)$$

Osserviamo anche che

$$\mathbb{E}\left[\overline{S}^2\right] = \frac{(n-1)\mathbb{E}\left[S_X^2\right] + (k-1)\mathbb{E}\left[S_Y^2\right]}{n+k-2} = \frac{(n-1)\sigma^2 + (k-1)\sigma^2}{n+k-2} = \sigma^2$$

e dunque possiamo usare $\overline{S}^2$ per stimare la varianza σ^2.

La probabilità di commettere errore di prima specie è allora

$$\alpha = \mathbb{P}(|T| \geq \varepsilon)$$

Fissato il livello di significatività α, devo dunque scegliere $\varepsilon = t_{n+k-2,1-\frac{\alpha}{2}}$.

Siano $x\colon x_1,\ x_2,\ \ldots,\ x_n$ e $y\colon y_1,\ y_2,\ \ldots,\ y_k$ i dati, $\overline{x}$ e $\overline{y}$ le rispettive medie, s_x^2 e s_y^2 le rispettive varianze e sia $\overline{s}^2 := \dfrac{(n-1)s_x^2 + (k-1)s_y^2}{n+k-2}$: accetto l'ipotesi nulla se

$$\frac{|\overline{x} - \overline{y}|}{\overline{s}\sqrt{\frac{1}{n} + \frac{1}{k}}} < t_{n+k-2,1-\frac{\alpha}{2}},$$

cioè se

$$|\overline{x} - \overline{y}| < t_{n+k-2,1-\frac{\alpha}{2}}\,\overline{s}\,\sqrt{\frac{1}{n} + \frac{1}{k}}$$

e la rifiuto altrimenti.

7.2.3. Terzo caso: le varianze σ_X^2 e σ_Y^2 sono ignote e diverse

Si può dimostrare che se n e k sono sufficientemente grandi e se H_0 è vera, allora la distribuzione della statistica

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{k}}}$$

è *approssimativamente* gaussiana standard. Dunque, per ottenere approssimativamente un livello di significatività α si accetta l'ipotesi nulla $H_0)$ $\quad \mu_X = \mu_Y$ quando

$$\frac{|\overline{x} - \overline{y}|}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{k}}} < z_{1-\frac{\alpha}{2}}$$

e la si rifiuta altrimenti.

Il problema di individuare un test che dia un livello di significatività α prescritto è ancora un problema aperto ed è noto come *problema di Behrens-Fisher*.

7.3. Test d'ipotesi per la varianza di campioni gaussiani

Test bilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ (nota o incognita) e varianza σ^2 incognita. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \sigma^2 = \sigma_0^2.$$

H_0 è vera se e solo se $\mathbb{E}\left[S^2\right] = \sigma_0^2$ se e solo se $\mathbb{E}\left[\frac{S^2}{\sigma_0^2}\right] = 1$.

Sappiamo che la v.a. $V := \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Dunque accetto H_0 se $1 - \varepsilon_1 < \dfrac{s^2}{\sigma_0^2} < 1 + \varepsilon_2$, ε_1, ε_2 positivi, cioè se e solo se

$$(n-1)(1-\varepsilon_1) < \frac{(n-1)s^2}{\sigma^2} < (n-1)(1+\varepsilon_2).$$

Devo scegliere ε_1 e e_2 in modo da ottenere il livello di significatività α desiderato:

$$\begin{aligned}\alpha &= \mathbb{P}\left(\frac{S^2}{\sigma_0^2} > 1+\varepsilon_2\right) + \mathbb{P}\left(\frac{S^2}{\sigma_0^2} < 1-\varepsilon_1\right) \\ &= \mathbb{P}\left(\frac{(n-1)S^2}{\sigma_0^2} > (n-1)(1+\varepsilon_2)\right) + \mathbb{P}\left(\frac{(n-1)S^2}{\sigma_0^2} < (n-1)(1-\varepsilon_1)\right).\end{aligned}$$

Una possibile scelta è allora

$$\mathbb{P}\left(\frac{(n-1)S^2}{\sigma_0^2} > (n-1)(1+\varepsilon_2)\right) = \frac{\alpha}{2} \qquad \text{cioè} \qquad (n-1)(1+\varepsilon_2) = \chi^2_{n-1,1-\frac{\alpha}{2}}$$
$$\mathbb{P}\left(\frac{(n-1)S^2}{\sigma_0^2} < (n-1)(1-\varepsilon_1)\right) = \frac{\alpha}{2} \qquad \text{cioè} \qquad (n-1)(1-\varepsilon_1) = \chi^2_{n-1,\frac{\alpha}{2}}.$$

Dunque accetto H_0 se

$$\chi^2_{n-1,\frac{\alpha}{2}} < \frac{(n-1)s^2}{\sigma_0^2} < \chi^2_{n-1,1-\frac{\alpha}{2}}$$

cioè se

$$\frac{\sigma_0^2}{n-1}\chi^2_{n-1,\frac{\alpha}{2}} < s^2 < \frac{\sigma_0^2}{n-1}\chi^2_{n-1,1-\frac{\alpha}{2}}$$

e la rifiuto altrimenti.

Test unilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ (nota o incognita) e varianza σ^2 incognita. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \sigma^2 \leq \sigma_0^2.$$

Accetto l'ipotesi nulla se $\dfrac{s^2}{\sigma_0^2} \leq 1+\varepsilon$.

Se la varianza è $\sigma^2 \leq \sigma_0^2$, allora $V := \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ e la probabilità di commettere errore di prima specie è

$$\begin{aligned}
&\mathbb{P}_{\sigma^2\leq\sigma_0^2}\left(\frac{S^2}{\sigma_0^2} > 1+\varepsilon\right) = \mathbb{P}_{\sigma^2\leq\sigma_0^2}\left(\frac{(n-1)S^2}{\sigma^2} > \frac{\sigma_0^2}{\sigma^2}(n-1)(1+\varepsilon)\right)\\
&= \mathbb{P}\left(V > \frac{\sigma_0^2}{\sigma^2}(n-1)(1+\varepsilon)\right) = 1 - F_V\left(\frac{\sigma_0^2}{\sigma^2}(n-1)(1+\varepsilon)\right)\\
&\leq 1 - F_V\left((n-1)(1+\varepsilon)\right).
\end{aligned}$$

Posso allora limitare superiormente con α la probabilità di commettere errore di prima specie imponendo

$$1 - F_V\left((n-1)(1+\varepsilon)\right) = \alpha$$

cioè scegliendo ε in modo che

$$(n-1)(1+\varepsilon) = \chi^2_{n-1,1-\alpha}.$$

Dunque accetto l'ipotesi nulla H_0 se

$$\frac{(n-1)s^2}{\sigma_0^2} < \chi^2_{n-1,1-\alpha}$$

cioè se

$$s^2 < \frac{\sigma_0^2}{n-1}\chi^2_{n-1,1-\alpha}$$

e la rifiuto altrimenti.

Test unilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ (nota o incognita) e varianza σ^2 incognita. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \sigma^2 \geq \sigma_0^2.$$

Accetto l'ipotesi nulla se $\dfrac{s^2}{\sigma_0^2} \geq 1 - \varepsilon$.

Se la varianza è $\sigma^2 \geq \sigma_0^2$, allora $V := \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ e la probabilità di commettere errore di prima specie è

$$\begin{aligned}\mathbb{P}_{\sigma^2}\left(\frac{S^2}{\sigma_0^2} < 1-\varepsilon\right) &= \mathbb{P}_{\sigma^2}\left(\frac{(n-1)S^2}{\sigma^2} < \frac{\sigma_0^2}{\sigma^2}(n-1)(1-\varepsilon)\right)\\ &= F_V\left(\frac{\sigma_0^2}{\sigma^2}(n-1)(1-\varepsilon)\right) \leq F_V\left((n-1)(1-\varepsilon)\right).\end{aligned}$$

Posso allora limitare superiormente con α la probabilità di commettere errore di prima specie imponendo

$$F_V\left((n-1)(1-\varepsilon)\right) = \alpha$$

cioè scegliendo ε in modo che

$$(n-1)(1-\varepsilon) = \chi^2_{n-1,\alpha}.$$

Dunque accetto l'ipotesi nulla H_0 se

$$\frac{(n-1)s^2}{\sigma_0^2} > \chi^2_{n-1,\alpha}$$

cioè se

$$s^2 > \frac{\sigma_0^2}{n-1}\chi^2_{n-1,\alpha}$$

e la rifiuto altrimenti.

7.4. Test d'ipotesi per la media di campioni bernoulliani

Abbiamo già trattato questo argomento nell' esempio introduttivo.

Test bilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ è un campione di Bernoulli di parametro p incognito. Sappiamo che $\mathbb{E}[X_i] = p$ e che, vedi (5.1)

$$\mathbb{P}\left(\left|\overline{X} - p\right| > t\right) \leq \frac{1}{4n\,t^2} \qquad \forall t > 0.$$

dunque usiamo $\overline{x}$ come stima di p.

Testiamo l'ipotesi nulla

$$H_0) \qquad p = p_0.$$

Stabiliamo il criterio di accettazione: accetto H_0 se $\overline{x} - p_0 < \varepsilon$ e la rifiuto altrimenti.

Se H_0 è vera, allora la v.a. $Y := \sum_{i=1}^{n} X_i$ segue la distribuzione $B(n, p_0)$ di cui conosciamo la distribuzione. La probabilità di commettere errore di prima specie è quindi

$$\begin{aligned}\alpha = \mathbb{P}_{p=p_0}\left(\left|\overline{X} - p_0\right| < \varepsilon\right) &= \mathbb{P}_{p=p_0}\left(|Y - np_0| < n\varepsilon\right) \\ &= \mathbb{P}\left(Y < n(p_0 + \varepsilon)\right) - \mathbb{P}\left(Y \leq n(p_0 - \varepsilon)\right).\end{aligned}$$

Test unilaterale

Testiamo l'ipotesi nulla

$$H_0) \qquad p \leq p_0.$$

Stabiliamo il criterio di accettazione: accetto H_0 se $\overline{x} < p_0 + \varepsilon$ e la rifiuto altrimenti.

Se H_0 è vera, allora la v.a. $Y := \sum_{i=1}^{n} X_i$ segue la distribuzione $B(n, p)$ per qualche $p \leq p_0$. La probabilità di commettere errore di prima specie è quindi

$$\begin{aligned}\mathbb{P}_{p\leq p_0}\left(\overline{X} \geq p_0 + \varepsilon\right) &= \mathbb{P}_{p\leq p_0}\left(Y \geq n(p_0 + \varepsilon)\right) \\ &\leq \mathbb{P}_{p_0}\left(Y \geq n(p_0 + \varepsilon)\right).\end{aligned}$$

Per limitare superiormente il livello di significatività α scelgo dunque ε in modo che $\mathbb{P}_{p_0}\left(Y \geq n(p_0 + \varepsilon)\right) \leq \alpha$.

Test unilaterale

Testiamo l'ipotesi nulla

$$H_0) \qquad p \geq p_0.$$

Stabiliamo il criterio di accettazione: accetto H_0 se $\overline{x} > p_0 - \varepsilon$ e la rifiuto altrimenti.

Se H_0 è vera, allora la v.a. $Y := \sum_{i=1}^{n} X_i$ segue la distribuzione $B(n, p)$ per qualche $p \geq p_0$. La probabilità di commettere errore di prima specie è quindi

$$\begin{aligned}\mathbb{P}_{p\geq p_0}\left(\overline{X} \leq p_0 - \varepsilon\right) &= \mathbb{P}_{p\geq p_0}\left(Y \leq n(p_0 - \varepsilon)\right) \\ &\leq \mathbb{P}_{p_0}\left(Y \leq n(p_0 - \varepsilon)\right).\end{aligned}$$

Per limitare superiormente il livello di significatività α scelgo dunque ε in modo che $\mathbb{P}_{p_0}\left(Y \leq n(p_0 - \varepsilon)\right) \leq \alpha$.

7.5. Test del χ^2

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico. Supponiamo che le v.a. del campione siano discrete a valori $y_1,\ y_2,\ \ldots,\ y_k$. Consideriamo le densità di probabilità

$$p_j := \mathbb{P}\left(X_i = y_j\right), \qquad j = 1, \ldots, k.$$

Vogliamo testare l'ipotesi nulla

$$H_0) \qquad p_j = p_j^0 \qquad \forall j = 1, \ldots, k.$$

Per ogni $j = 1, \ldots, k$ considero le *frequenze campionarie*

$$N_j(\omega) = \#\left\{i \in \{1, \ldots, n\} \colon X_i(\omega) = y_k\right\} \qquad j = 1, \ldots, k$$

e le *frequenze campionarie relative*

$$F_j := \frac{N_j}{n}, \qquad j = 1, \ldots, k.$$

Sicuramente $N_j \sim B(n, p_j)$, quindi $\mathbb{E}\left[N_j\right] = np_j$ e $\mathbb{E}\left[F_j\right] = p_j$.

In particolare H_0 è vera se e solo se $\mathbb{E}\left[N_j\right] = np_j^0 \quad \forall j = 1, \ldots, k$, dunque un criterio di accettazione potrebbe essere quello di accettare H_0 se e solo se $\left|n_j - np_j^0\right| < \varepsilon \quad \forall j = 1, \ldots, k$.

Questo criterio però non ci permette di calcolare la probabilità di errore di prima specie. Vale però il seguente risultato:

Teorema 7.5.1 (di Pearson)**.** *Se $N_j \sim B(n, p_j)$, allora la funzione di ripartizione della v.a.*

$$T_n := \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$$

converge, per $n \to \infty$, alla funzione di ripartizione associata alla distribuzione χ^2_{k-1}.

Osservazione 7.5.1. Nelle applicazioni l'approssimazione è considerata accettabile se $np_j \geq 5 \quad \forall j = 1, \ldots, k$.

Formuliamo allora il seguente criterio di accettazione:

$$\text{accetto l'ipotesi nulla } H_0 \text{ se e solo se } t_n := \sum_{j=1}^{k} \frac{(n_j - np_j^0)^2}{np_j^0} < \varepsilon.$$

La probabilità di commettere errore di prima specie è allora

$$\alpha := \mathbb{P}\left(T_n \geq \varepsilon\right) \simeq 1 - F_{\chi^2_{k-1}}(\varepsilon).$$

Scelgo dunque ε tale che $F_{\chi^2_{k-1}}(\varepsilon) = 1 - \alpha$, cioè $\varepsilon = \chi^2_{k-1,1-\alpha}$.

Osservazione 7.5.2. Il test si può applicare anche nel caso in cui $y_1,\ y_2,\ \ldots,\ y_k$ siano sostituiti da classi di modalità $I_1,\ I_2,\ \ldots,\ I_k$.

7.5.1. Test di normalità

Supponiamo di aver un campione $X_1,\ X_2,\ \ldots,\ X_n$. Vogliamo testare l'ipotesi

$$H_0) \qquad \text{Il campione è normale}$$

Si può procedere nel seguente modo:

1. stimiamo μ e σ^2 rispettivamente con $\overline{x}$ e s^2;
2. standardizziamo i dati ponendo $z_i := \dfrac{x_i - \mu}{\sigma}$. Se il campione segue la distribuzione $\mathcal{N}\left(\mu, \sigma^2\right)$, allora $Z_i \sim \mathcal{N}\left(0, 1\right)$;
3. suddividiamo la retta reale in intervalli $I_1,\ I_2,\ \ldots,\ I_k$ (simmetrici rispetto all'origine), ivi comprese due semirette simmetriche $[a, +\infty)$ e $(-\infty, -a]$;
4. contiamo $n_j := \#\left\{i \in \{1, \ldots, n\} \colon z_i \in I_j\right\}$;
5. calcoliamo $p_j^0 := \mathbb{P}\left(Z_i \in I_j\right)$;
6. consideriamo la v.a. $T_k^{(n)} := \displaystyle\sum_{j=1}^{k} \frac{(N_j - np_j^0)^2}{np_j^0}$. Si può dimostrare che per $n \to \infty$ la funzione di ripartizione di $T_k^{(n)}$ converge alla funzione di ripartizione associata alla distribuzione χ^2_{k-2-1}, dove il -2 è dovuto al fatto che abbiamo sostituito i due parametri μ e σ^2 con le loro stime provenienti dai dati $\overline{x}$ e s^2;
7. accettiamo l'ipotesi nulla se $t_k^{(n)} < \varepsilon$. Se imponiamo un livello di significatività α, sceglierermo allora $\varepsilon = \chi^2_{k-3,1-\alpha}$.