Analisi Reale - Appunti di Statistica

Laura Poggiolini

B204 – a.a. 2017–18

Indice

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

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

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

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

I caratteri numerici a loro volta si possono suddividere in

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

1.1 Campione statistico, modalità e classi modali

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

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

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

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

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

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

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

1.2 Frequenza assoluta e frequenza relativa

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

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

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

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

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

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

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

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

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

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

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

1.3 Moda e valori modali

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

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

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

Possiamo visualizzare con degli istogrammi, vedi Figura 1.3

1.4 Mediana

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

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

e distinguiamo due casi:

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

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

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

- n pari: $n = 2m$

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

  Il dato $x_{(m)}$ è maggiore-uguale di $m - 1$ dati e minore-uguale di m dati. Il dato $x_{(m+1)}$ è maggiore-uguale di m dati e minore-uguale di $m - 1$ dati.

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

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

Consideriamo un campione relativo ad un carattere numerico

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

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

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

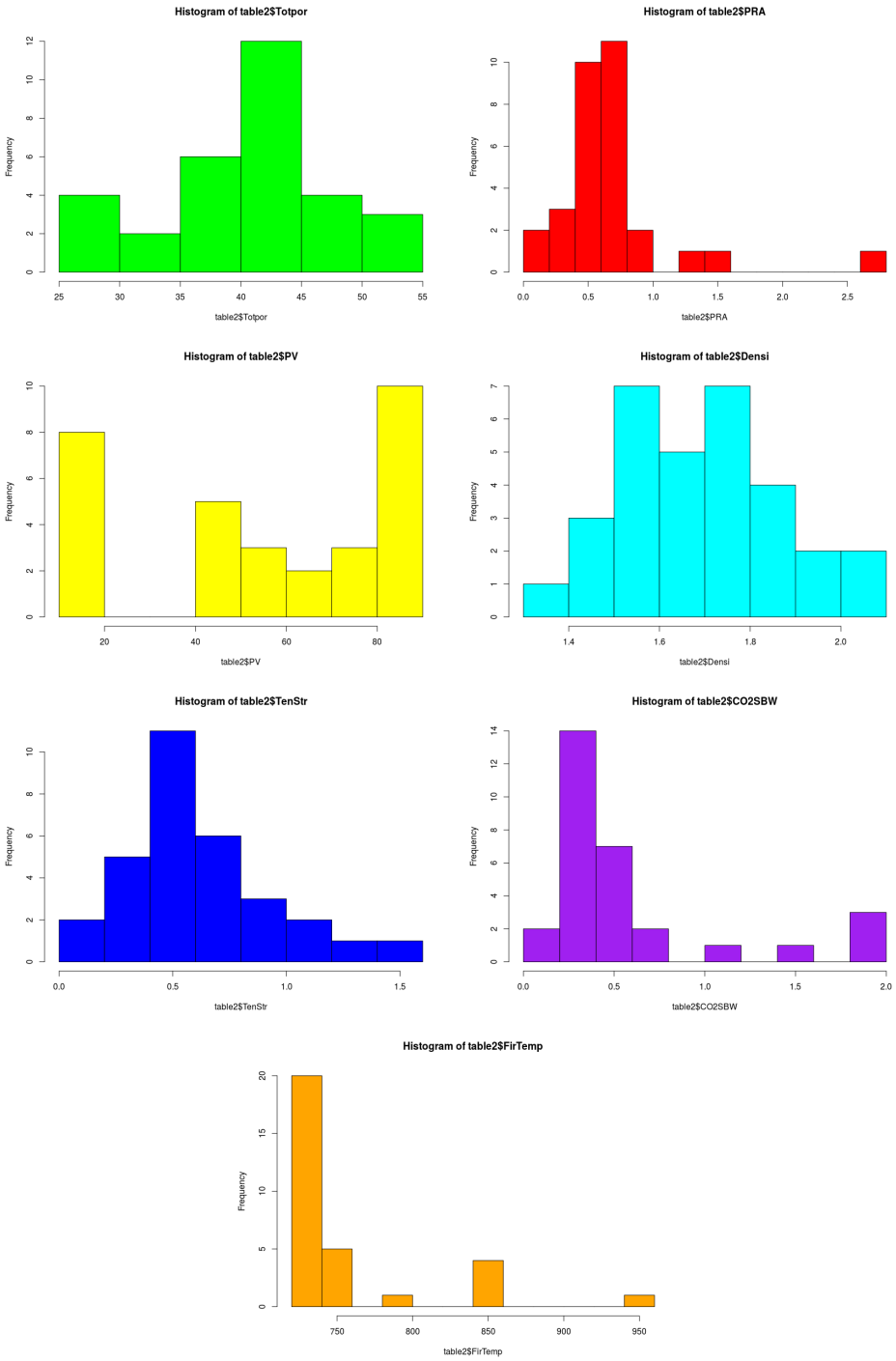

Figura 1.1: Alcuni istogrammi dall'Esempio 1.5.1

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

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

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

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

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

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

La radice quadrata della varianza campionaria

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

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

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

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

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

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

Inseriamo la tabella in R

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

```
)
> View(table2)
```

	Code	Totpor	PRA	PV	Densi	TenStr	CO2SBW	FirTemp
1	AS1	41.46	0.528	80.0	1.55	0.403	0.38	740
2	AS2	47.21	0.467	81.2	1.65	0.645	0.70	740
3	AS3	43.67	0.697	78.5	1.71	0.527	0.46	740
4	AS4	52.39	0.422	77.3	1.52	0.143	0.48	740
5	AS5	44.70	0.411	87.4	1.50	0.593	0.29	740
6	AS6	51.33	0.422	88.6	1.48	0.463	0.33	740
7	AS7	31.46	0.718	80.6	1.90	0.955	0.23	740
8	AS8	40.90	0.458	80.4	1.68	0.195	0.41	740
9	AS9	45.54	0.492	80.8	1.62	1.328	0.50	750
10	AS10	45.62	0.734	86.2	1.62	1.405	0.34	750
11	AS11	44.14	0.730	85.7	1.59	0.256	0.42	750
12	AS12	40.71	0.543	87.8	1.75	0.309	0.20	750
13	AS13	35.70	0.686	84.3	1.52	0.472	0.05	740
14	C1	40.29	0.306	43.5	1.76	0.520	0.43	740
15	C2	36.57	0.625	42.3	1.75	0.738	0.36	740
16	C3	42.13	0.249	63.2	1.63	0.410	0.25	740
17	C4	37.83	0.731	47.9	2.02	0.601	0.28	740
18	C5	42.18	0.407	59.4	1.58	0.376	0.34	740
19	C6	41.60	0.446	42.8	1.85	0.473	0.26	740
20	C7	32.66	0.664	64.3	1.85	0.695	0.25	740
21	C8	36.07	0.673	58.2	1.78	0.624	0.29	740
22	C9	36.04	1.397	55.6	1.73	0.582	0.38	740
23	C10	36.64	0.861	45.2	1.75	0.650	0.47	740
24	R1	42.89	0.785	10.2	1.54	0.453	1.04	850
25	R2	26.85	0.315	14.7	2.01	1.124	1.86	960
26	R3	28.55	0.158	18.6	1.92	0.937	1.96	850
27	R4	29.86	0.158	15.3	1.89	1.020	1.48	850
28	R5	45.70	0.984	12.8	1.50	0.328	--	800
29	R6	54.64	1.525	12.5	1.34	0.267	0.67	750
30	R7	27.55	2.657	14.6	1.92	0.892	0.40	730
31	R8	40.82	0.622	15.3	1.57	0.502	1.94	860

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

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

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

Richiediamo anche varianza campionaria e deviazione standard.

```
> medianaTotPor <- median(table2$Totpor);
> meanTotPor <- mean(table2$Totpor);
> VarTotPor <- var(table2$Totpor);
> StdTotPor <- sd(table2$Totpor)
> medianaTotPor; meanTotPor; VarTotPor; StdTotPor
[1] 40.9
[1] 40.11935
[1] 49.52185
[1] 7.037176
> medianaPRA <- median(table2$PRA);
> meanPRA <- mean(table2$PRA);
VarPRA <- var(table2$PRA);
> StdPRA <- sd(table2$PRA)
> medianaPRA; meanPRA; VarPRA; StdPRA
[1] 0.622
[1] 0.6732581
[1] 0.226613
[1] 0.4760389
> medianaPV <- median(table2$PV);
> meanPV <- mean(table2$PV);
> VarPV <- var(table2$PV);
> StdPV <- sd(table2$PV)
> medianaPV; meanPV; VarPV; StdPV
[1] 59.4
[1] 55.32903
[1] 815.0935
[1] 28.54984
> medianaDensi <- median(table2$Densi);
> meanDensi <- mean(table2$Densi);
> VarDensi <- var(table2$Densi);
> StdDensi <- sd(table2$Densi)
> medianaDensi; meanDensi; VarDensi; StdDensi
[1] 1.68
[1] 1.692903
[1] 0.02894129
[1] 0.1701214
> medianaTenStr <- median(table2$TenStr);
> meanTenStr <- mean(table2$TenStr);
> VarTenStr <- var(table2$TenStr);
> StdTenStr <- sd(table2$TenStr)
> medianaTenStr; meanTenStr; VarTenStr; StdTenStr
[1] 0.527
[1] 0.6092258
[1] 0.09882738
[1] 0.3143682
```

```
> medianaCO2SBW <- median(na.omit(table2$CO2SBW));
> meanCO2SBW <- mean(na.omit(table2$CO2SBW));
> VarCO2SBW <- var(na.omit(table2$CO2SBW));
> StdCO2SBW <- sd(na.omit(table2$CO2SBW))
> medianaCO2SBW; meanCO2SBW; VarCO2SBW;  StdCO2SBW
[1] 0.39
[1] 0.5816667
[1] 0.2765868
[1] 0.5259152
> medianaFirTemp <- median(table2$FirTemp);
> meanFirTemp <- mean(table2$FirTemp);
> VarFirTemp <- var(table2$FirTemp);
> StdFirTemp <- sd(table2$FirTemp)
> medianaFirTemp; meanFirTemp; VarFirTemp; StdFirTemp
[1] 740
[1] 764.8387
[1] 2805.806
[1] 52.96986
```

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

2.1 Covarianza e coefficiente di correlazione

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

Abbiamo dunque due vettori di dati

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

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

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

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

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

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

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

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

Osservando che $\rho\,[x,y]$ non è altro che il rapporto tra $\langle x-(\overline{x},\ldots,\overline{x}), y-(\overline{y},\ldots,\overline{y})\rangle$ (prodotto scalare) e $\|x-(\overline{x},\ldots,\overline{x})\|\|y-(\overline{y},\ldots,\overline{y})\|$ (prodotto delle norme) si dimostrano facilmente le seguenti proprietà:

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

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

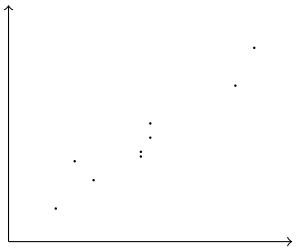

Figura 2.1: Campione bivariato *pressoché lineare*

2.2 Retta di regressione

Supponiamo di avere un campione bivariato

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

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

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

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

Si ha

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

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

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

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.

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

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

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

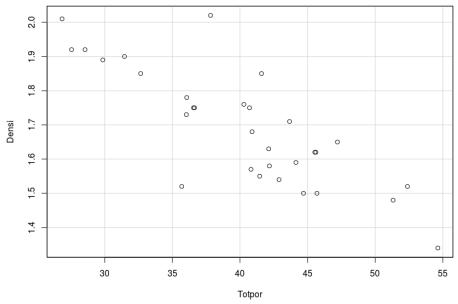

Figura 2.2: Porosità totale versus Densità

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

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

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

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

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

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

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

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

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

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

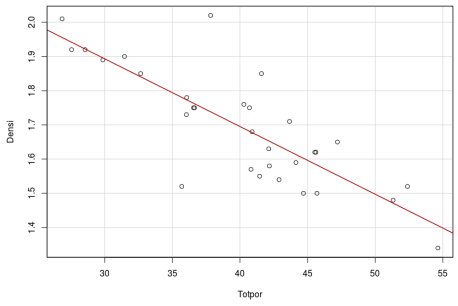

Figura 2.3: Retta di regressione lineare

Parte II

Statistica inferenziale

3. Campioni statistici

3.1 Introduzione

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

Possiamo sintetizzare il modello matematico che applichiamo come segue

- Se rileviamo un carattere su una popolazione di n individui, consideriamo ciascun dato rilevato come il valore assunto da $X_1,\ X_2,\ \ldots,\ X_n$ variabili aleatorie aventi tutte la stessa distribuzione μ e che (molto spesso) si possono supporre indipendenti.
- La distribuzione μ è (parzialmente) incognita; si cercano informazioni su μ a partire dai dati rilevati. Le informazioni ricavate sulla distribuzione μ sono di natura probabilistica. Per esempio, non riusciremo ad ottenere informazioni del tipo *la media della distribuzione μ è* 50 ma informazioni del tipo *la media della distribuzione μ è compresa tra 49.8 e 50.2 con probabilità del 90%.*

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

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

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

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

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

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

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

Definizione 3.1.2 (Statistica)**.** Sia $X_1,\ \ldots,\ X_n$ un campione statistico. Una funzione (non dipendente da parametri) di $X_1,\ \ldots,\ X_n$ si dice una statistica.

Osservazione 3.1.2. Chiariamo cosa si intende per statistica: $3X_1 - 2X_2$ è una statistica; $\max\{X_1,\ \ldots,\ X_n\}$ è una statistica. $X_1 - \mu$ $\mu \in \mathbb{R}$ *non* è una statistica.

3.2 Media campionaria e varianza campionaria

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

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

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

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

Proposizione 3.2.1. *Sia $X_1, \ldots, X_n$ un campione statistico di numerosità n con 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

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

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

3.2.1 La disuguaglianza di Chebychev 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 3.2.1 (Disuguaglianza di Chebychev)**.** *Se X è una variabile aleatoria con media μ e varianza non superiore a σ^2, allora*

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

Osservazione 3.2.1. La disuguaglianza di Chebychev 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 3.2.1. Sia $X_1, \ldots, X_n$ 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 Chebychev segue facilmente il seguente

Teorema 3.2.2 (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

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

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 Chebychev ci dice con precisione quanto è *probabilisticamente accettabile* questa sostituzione.

Esempio 3.2.2. 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 $X_1, \ldots, X_n$ 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 Chebychev

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

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 $\dfrac{100}{n} \leq \dfrac{1}{10}$ 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 $\overline{x} = \dfrac{1}{1000}\displaystyle\sum_{i=1}^{1000} x_i$ 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 3.2.3. Sia $X_1, \ldots, X_n$ un campione statistico di numerosità n, bernoulliano di parametro (incognito) $p \in [0,1]$. Dunque

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

Allora, per la disuguaglianza di Chebychev

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

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

3.2.2 La distribuzione gaussiana $N\left(\mu, \sigma^2\right)$ e il teorema del limite centrale

Ricordiamo che la distribuzione gaussiana di parametri $\mu \in \mathbb{R}$ e $\sigma^2 > 0$, $N\left(\mu, \sigma^2\right)$, 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 $N\left(\mu, \sigma^2\right)$, allora

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

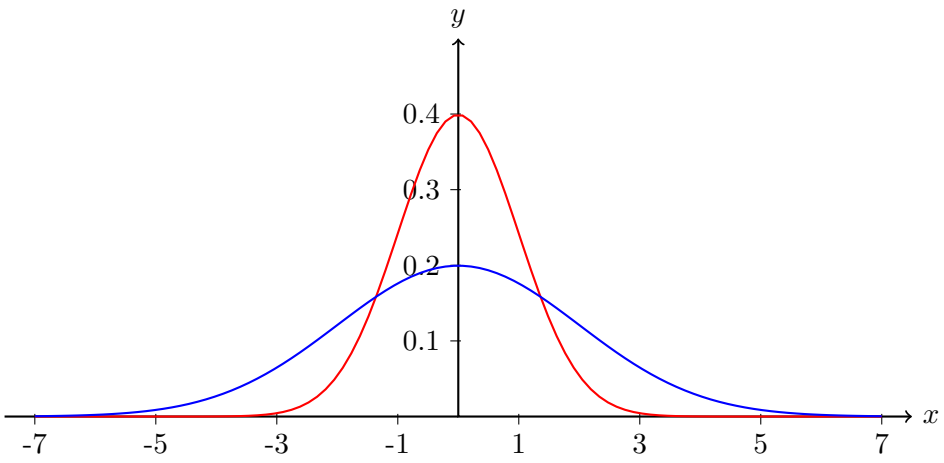

Figura 3.1: Densità associate alle distribuzioni $N(0,1)$ (in rosso) e $N(0,4)$ (in blu)

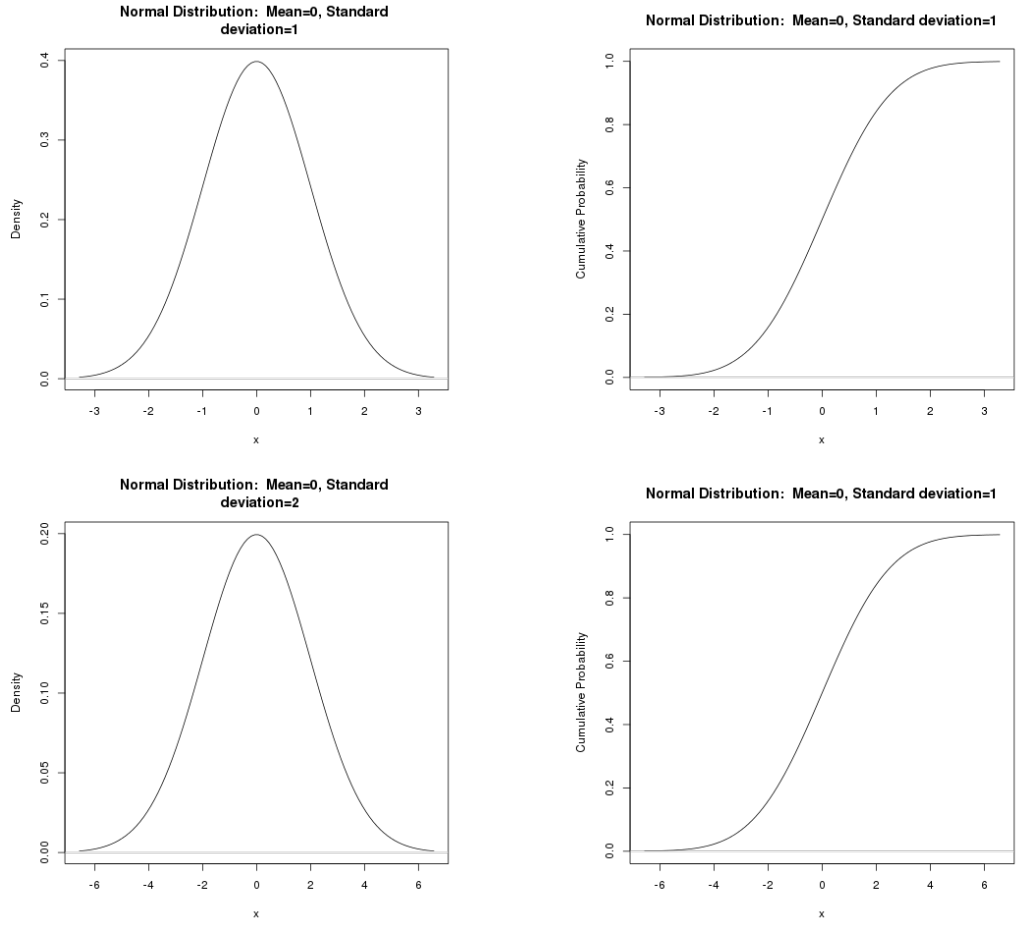

Figura 3.2: $N(0,1)$ e $N(0,4)$, densità e funzione di ripartizione

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 $N(0,1)$ 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

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

Ricordiamo alcune proprietà che abbiamo già visto:

Proprietà 3.2.1.

1. Se X è una v.a. gaussiana di media μ e varianza σ^2: $\mathbb{P}_X = N\left(\mu, \sigma^2\right)$ 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 $\alpha^2\sigma^2$: $\mathbb{P}_{\alpha X+\beta} = N\left(\alpha\mu + \beta, \alpha^2\sigma^2\right)$. In particolare $Y := \dfrac{X - \mu}{\sigma}$ è una v.a. gaussiana standard: $\mathbb{P}_Y = N(0,1)$.

2. Siano $X_1, \ldots, X_n$ v.a. indipendenti con X_i gaussiana di media μ_i e varianza σ_i^2: $\mathbb{P}_{X_i} = N\left(\mu_i, \sigma_i^2\right)$ $\forall i = 1, \ldots, n$. Allora la v.a. $S_n := X_1 + X_2 + \ldots + X_n$ è gaussiana di media pari alla somma delle medie e varianza pari alla somma delle varianze:

$$\mathbb{P}_{S_n} = N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Teorema 3.2.3 (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 $N(0,1)$.*

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

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

Allora

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

ed il limite è uniforme in $t \in \mathbb{R}$.

Osservazione 3.2.2. Una formulazione equivalente della tesi del teorema del limite centrale è

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

Esempio 3.2.4. 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 $\mu = \mathbb{E}[X_i] \quad \forall i = 1, \dots, n$ e dunque è anche $\mu = \mathbb{E}\left[\overline{X}\right]$. Applicando la disuguaglianza di Chebychev 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

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

Perché questa stima *sembra* tanto migliore di quella ottenuta con la disuguaglianza di Chebychev? Perché non abbiamo un'indicazione sul significato del primo dei $\simeq$. 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 3.2.4 (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 $N(0,1)$.*

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 3.2.4, otteniamo dunque

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

3.3 Alcune distribuzioni legate alla distribuzione gaussiana

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

Si tratta della distribuzione $\Gamma(\alpha, \lambda)$ dove $\alpha = \frac{n}{2}$, $\lambda = \frac{1}{2}$. È dunque 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 3.3.1. Abbiamo visto che $\forall a > 0$ si ha $\Gamma(a+1) = a\Gamma(a)$ e che $\Gamma(1) = 1$. Inoltre $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. Infatti (con la sostituzione $x = y^2$)

$$\Gamma\left(\frac{1}{2}\right) = \int_0^{+\infty} x^{-1/2}\mathrm{e}^{-x/2}\mathrm{d}x = \int_0^{+\infty} 2\,\mathrm{e}^{-y^2}\mathrm{d}y = \int_{\mathbb{R}} \mathrm{e}^{-y^2}\mathrm{d}y = \sqrt{\pi}.$$

Quindi

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

Proprietà 3.3.1. Se X è una v.a. con distribuzione χ^2 a n gradi di libertà, $\mathbb{P}_X = \chi^2_n$, allora

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

Dimostrazione. Poiché una v.a. con distribuzione $\Gamma(\alpha,\lambda)$ ha valore atteso α/λ e varianza α/λ^2, in particolare per una v.a. con distribuzione di Pearson abbiamo

$$\mathbb{E}[X] = \frac{\frac{n}{2}}{\frac{1}{2}} = n, \qquad \mathrm{Var}[X] = \frac{\frac{n}{2}}{\left(\frac{1}{2}\right)^2} = 2n.$$

□

Lemma 3.3.1. *Se X e Y sono due variabili aleatorie indipendenti, con distribuzioni $\mathbb{P}_X = \Gamma(\alpha,\lambda)$, $\mathbb{P}_Y = \Gamma(\beta,\lambda)$, allora la v.a. $X+Y$ ha distribuzione $\Gamma(\alpha+\beta,\lambda)$.*

Dimostrazione. Sappiamo che la distribuzione di $X+Y$ è a.c. con densità $h(x)$ data dal prodotto di convoluzione delle densità associate alle distribuzioni $\Gamma(\alpha,\lambda)$ e $\Gamma(\beta,\lambda)$. Dunque $h(x) = 0$ per $x \le 0$. Per $x > 0$ abbiamo invece

$$\begin{aligned} h(x) &= \int_0^x \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1}\mathrm{e}^{-\lambda y}\frac{\lambda^\beta}{\Gamma(\beta)}(x-y)^{\beta-1}\mathrm{e}^{-\lambda(x-y)}\mathrm{d}y \\ &= \mathrm{e}^{-\lambda x}\frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)}\int_0^x y^{\alpha-1}(x-y)^{\beta-1}\mathrm{d}y = \qquad \text{(sostituisco } y = xt\text{)} \\ &= \mathrm{e}^{-\lambda x}\frac{\lambda^{\alpha+\beta}x^{\alpha+\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t = Cx^{\alpha+\beta-1}\mathrm{e}^{-\lambda x} \end{aligned}$$

dove $C = \dfrac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)}\displaystyle\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t$. Poiché h deve essere una densità di probabilità può solo essere $C = \dfrac{\lambda^{\alpha+\beta}}{\Gamma(\alpha+\beta)}$. □

Teorema 3.3.2. *Se X e Y sono due variabili di Pearson indipendenti, $\mathbb{P}_X = \chi^2_n$, $\mathbb{P}_Y = \chi^2_k$, allora la v.a. $X+Y$ segue la distribuzione di Pearson a $n+k$ gradi di libertà:*

$$\mathbb{P}_{X+Y} = \chi^2_{n+k}.$$

Dimostrazione. Scegliendo $\alpha = \frac{n}{2}$, $\beta = \frac{k}{2}$, $\lambda = \frac{1}{2}$ nel Lemma 3.3.1, si ottiene la tesi. □

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

Teorema 3.3.3. *Se X è una v.a. gaussiana standard, $\mathbb{P}_X = N(0,1)$, allora X^2 segue la distribuzione di Pearson ad un grado di libertà, $\mathbb{P}_{X^2} = \chi^2_1$.*

Dimostrazione. Sappiamo che $\mathbb{P}_X = N(0,1) = f(x)\mathrm{d}x$ con $f(x) = \dfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$. Dunque $\mathbb{P}_{X^2} = g(x)\mathrm{d}x$ con

$$g(x) = \begin{cases} 0 & x \leq 0, \\ \frac{1}{\sqrt{2\pi}} x^{-1/2}\mathrm{e}^{-x/2} & x > 0, \end{cases}$$

cioè $\mathbb{P}_{X^2} = \chi^2_1$. □

Teorema 3.3.4. *Se $X_1, \ldots, X_n$ 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.*

Dimostrazione. Poiché la v.a. $\dfrac{X_i - \mu_i}{\sigma_i}$ha distribuzione gaussiana standard, applicando i teoremi 3.3.3 e 3.3.2 ed il principio di induzione si ottiene la tesi. □

Corollario 3.3.5. *Se $X_1, \ldots, X_n$ è un campione statistico gaussiano, con media μ e varianza σ^2, allora la v.a. $\displaystyle\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$ segue una distribuzione χ^2 con n gradi di libertà.*

Esempio 3.3.1. 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/2017-18_analisi_reale/alcuni_appunti")
> .x <- seq(0.015, 18.015, length.out=100)
> plot(.x, dchisq(.x, df=3), xlab="x", ylab="Density",
+ main=paste("ChiSquared Distribution:  Degrees of freedom=3"), type="l")
> plot(.x, pchisq(.x, df=3), xlab="x", ylab="Density",
+ main=paste("ChiSquared Distribution:  Degrees of freedom=3"), type="l")
> abline(h=0.36, col="red")
> pchisq(c(0.36), df=3, lower.tail=TRUE)
[1] 0.05162424
```

Il seguente teorema raccoglie alcune importanti proprietà dei campioni statistici gaussiani e delle loro media e varianza campionarie.

Teorema 3.3.6. *Sia $X_1, \ldots, X_n$ un campione statistico gaussiano di numerosità n, valore atteso μ e varianza σ^2.*
Allora, la media campionaria $\overline{X}$ e la varianza campionaria S^2 sono v.a. indipendenti.

Sia $Z_1, Z_2, \ldots Z_n$ la standardizzazione del campione statistico $X_1, \ldots, X_n$ 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 $Z_1, \ldots, Z_n$.
Allora $\overline{Z} = \dfrac{\overline{X} - \mu}{\sigma}$ e la v.a. $\displaystyle\sum_{i=1}^{n} \left(Z_i - \overline{Z}\right)^2$ sono indipendenti e quest'ultima segue una distribuzione χ^2 con $n-1$ gradi di libertà.

Dimostrazione. **1. $\mathbf{n = 2}$.** Sappiamo che $\mathbb{P}_{X_1+X_2} = N\left(2\mu, 2\sigma^2\right)$ e $\mathbb{P}_{\overline{X}} = N\left(\mu, \sigma^2/2\right)$. Inoltre

$$S^2 = (X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 = \frac{1}{2}(X_1 - X_2)^2.$$

Dunque $\overline{X}$ e S^2 sono indipendenti se e solo se $X_1 + X_2$ e $X_1 - X_2$ sono indipendenti. Poiché $\mathbb{P}_{-X_2} = N\left(-\mu, \sigma^2\right)$ abbiamo che $\mathbb{P}_{X_1-X_2} = N\left(0, 2\sigma^2\right)$.

Per provare che $U := X_1 + X_2$ e $V := X_1 - X_2$ sono indipendenti ne calcoliamo la densità congiunta e mostriamo che è uguale al prodotto delle densità marginali. Abbiamo già visto che $\mathbb{P}_{X_1+X_2} = N\left(2\mu, 2\sigma^2\right)$. Inoltre, poiché $\mathbb{P}_{-X_2} = N\left(-\mu, \sigma^2\right)$ abbiamo che $\mathbb{P}_{X_1-X_2} = N\left(0, 2\sigma^2\right)$. Posto

$$\varphi\colon (x, y) \in \mathbb{R}^2 \mapsto (x + y, x - y) \in \mathbb{R}^2$$

abbiamo

$$(U, V) = \varphi \circ (X_1, X_2)$$

dunque, per ogni funzione boreliana non-negativa $\psi\colon \mathbb{R}^2 \to \mathbb{R}$ abbiamo

$$\int_{\mathbb{R}^2} \psi(u, v) \mathbb{P}_{U,V}(\mathrm{d}u\mathrm{d}v) = \int_{\mathbb{R}^2} \psi(x + y, x - y) \mathbb{P}_{X_1,X_2}(\mathrm{d}x\mathrm{d}y)$$

$$= \int_{\mathbb{R}^2} \psi(x + y, x - y) \frac{1}{2\pi\sigma^2} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) \mathrm{d}x\mathrm{d}y$$

con il cambiamento di variabile $u = x + y$, $v = x - y$

$$= \int_{\mathbb{R}^2} \psi(u, v) \frac{1}{2\pi(\sqrt{2}\sigma)^2} \exp\left(\frac{-(u-2\mu)^2}{2(\sqrt{2}\sigma)^2}\right) \exp\left(\frac{-v^2}{2(\sqrt{2}\sigma)^2}\right) \mathrm{d}u\mathrm{d}v$$

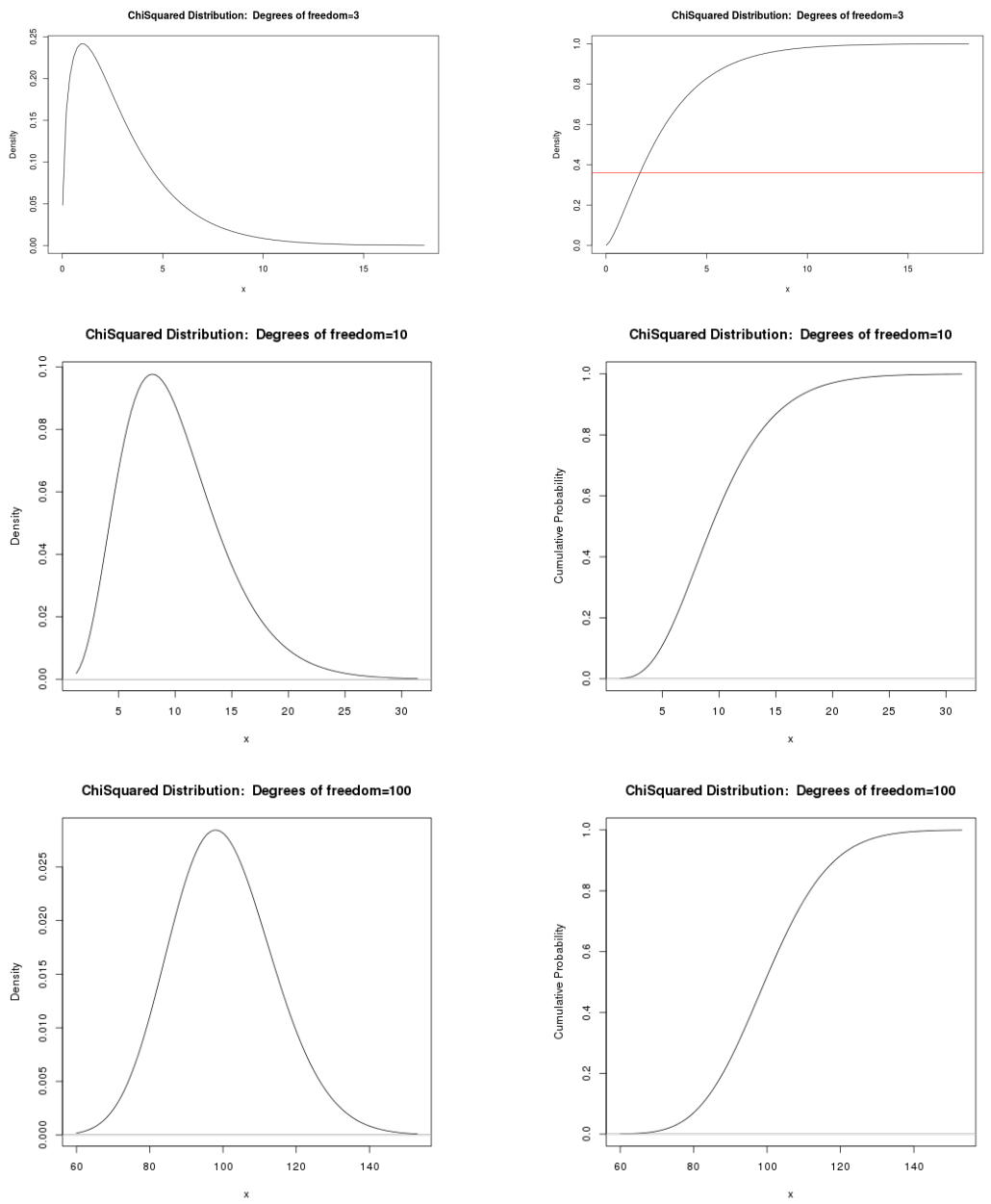

Figura 3.3: χ^2_3, χ^2_{10} e χ^2_{100}, densità e funzione di ripartizione

ovvero la densità congiunta è il prodotto delle densità marginali

$$f_{X_1+X_2}(u) = \frac{1}{\sqrt{2\pi(\sqrt{2}\sigma)^2}} \exp\left(\frac{-(u-2\mu)^2}{2(\sqrt{2}\sigma)^2}\right), \quad f_{X_1-X_2}(v) = \frac{1}{\sqrt{2\pi(\sqrt{2}\sigma)^2}} \exp\left(\frac{-v^2}{2(\sqrt{2}\sigma)^2}\right).$$

Inoltre, se Z_1 e Z_2 sono gaussiane standard indipendenti abbiamo:

$$(Z_1 - \overline{Z})^2 + (Z_2 - \overline{Z})^2 = \frac{1}{2}(Z_1 - Z_2)^2 = \left(\frac{Z_1 - Z_2}{\sqrt{2}}\right)^2.$$

La v.a. $Z_1 - Z_2$ ha distribuzione $N(0,2)$, dunque $\dfrac{Z_1 - Z_2}{\sqrt{2}}$ ha distribuzione $N(0,1)$. Applicando il Teorema 3.3.3 otteniamo la tesi.

2. $\mathbf{n \geq 3}$. Procediamo per induzione, supponendo che $\overline{X}_{n-1}$ e S^2_{n-1} siano indipendenti. Osserviamo che

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\left((n-1)\overline{X}_{n-1} + X_n\right) = \frac{n-1}{n}\overline{X}_{n-1} + \frac{1}{n}X_n \tag{3.1}$$

e dunque

$$\overline{X}_n - \overline{X}_{n-1} = \frac{1}{n}\left(X_n - \overline{X}_{n-1}\right).$$

Abbiamo dunque

$$\begin{aligned}
S_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_{n-1} + \overline{X}_{n-1} - \overline{X}_n\right)^2 \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\left(X_i - \overline{X}_{n-1}\right)^2 + 2\sum_{i=1}^{n}\left(\overline{X}_{n-1} - \overline{X}_n\right)\left(X_i - \overline{X}_{n-1}\right) + \sum_{i=1}^{n}\left(\overline{X}_{n-1} - \overline{X}_n\right)^2\right) \\
&= \frac{1}{n-1}\left((n-2)S^2_{n-1} + \left(X_n - \overline{X}_{n-1}\right)^2 + 2\left(\overline{X}_{n-1} - \overline{X}_n\right)n\left(\overline{X}_n - \overline{X}_{n-1}\right) + n\left(\overline{X}_{n-1} - \overline{X}_n\right)^2\right) \\
&= \frac{1}{n-1}\left((n-2)S^2_{n-1} + \left(X_n - \overline{X}_{n-1}\right)^2 - \frac{2}{n}\left(X_n - \overline{X}_{n-1}\right)\left(X_n - \overline{X}_{n-1}\right) + \frac{1}{n}\left(X_n - \overline{X}_{n-1}\right)^2\right) \\
&= \frac{1}{n-1}\left((n-2)S^2_{n-1} + \frac{n-1}{n}\left(X_n - \overline{X}_{n-1}\right)^2\right) \qquad (3.2)
\end{aligned}$$

Per la (3.1) e l'ipotesi di induzione $\overline{X}_n$ è indipendente da S^2_{n-1}. Avremo dunque che S_n^2 e $\overline{X}_n$ sono indipendenti se e solo se $\overline{X}_n$ e $X_n - \overline{X}_{n-1}$ sono indipendenti.

Sappiamo che $\mathbb{P}_{X_n} = N\left(\mu, \dfrac{\sigma^2}{n}\right)$, dunque

$$\mathbb{P}_{\overline{X}_n} = N\left(\mu, \frac{\sigma^2}{n}\right), \quad \mathbb{P}_{\overline{X}_{n-1}} = N\left(\mu, \frac{\sigma^2}{n-1}\right), \quad \mathbb{P}_{X_n - \overline{X}_{n-1}} = N\left(0, \sigma^2\frac{n}{n-1}\right),$$

Devo provare che $U := \dfrac{n-1}{n}\overline{X}_{n-1} + \dfrac{1}{n}X_n$ e $V = X_n - \overline{X}_{n-1}$ sono indipendenti. Osserviamo che

$$(U,V) = \varphi \circ \left(\overline{X}_{n-1}, X_n\right), \qquad \varphi(x,y) = \left(\frac{n-1}{n}x + \frac{1}{n}y, y - x\right).$$

Sia dunque $\psi\colon \mathbb{R}^2 \to \mathbb{R}$ una funzione di Borel non negativa. Abbiamo

$$\int_{\mathbb{R}^2} \psi(u,v)\mathbb{P}_{U,V}(\mathrm{d}u\mathrm{d}v) = \int_{\mathbb{R}^2} \psi\left(\frac{n-1}{n}x + \frac{1}{n}y, y-x\right) \mathbb{P}_{\overline{X}_{n-1},X_n}\mathrm{d}x\mathrm{d}y$$
$$= \int_{\mathbb{R}^2} \psi\left(\frac{n-1}{n}x + \frac{1}{n}y, y-x\right) \frac{\sqrt{n-1}}{2\pi\sigma^2} \exp\left(\frac{-(n-1)(x-\mu)^2 - (y-\mu)^2}{2\sigma^2}\right)\mathrm{d}x\mathrm{d}y$$

con il cambiamento di variabile $u = \dfrac{n-1}{n}x + \dfrac{1}{n}y$, $v = y - x$

$$= \int_{\mathbb{R}^2} \psi(u,v) \frac{\sqrt{n-1}}{2\pi\sigma^2} \exp\left(\frac{-(u-\mu)^2\left(\sqrt{n}\right)^2}{2\sigma^2}\right) \exp\left(\frac{-v^2\left(\sqrt{\dfrac{n-1}{n}}\right)^2}{2\sigma^2}\right)\mathrm{d}u\mathrm{d}v$$

$$= \int_{\mathbb{R}^2} \psi(u,v) \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n}}} \exp\left(\frac{-(u-\mu)^2}{2\left(\frac{\sigma}{\sqrt{n}}\right)^2}\right) \frac{1}{\sqrt{2\pi\sigma^2\frac{n}{n-1}}} \exp\left(\frac{-v^2}{2\left(\sigma\sqrt{\dfrac{n-1}{n}}\right)^2}\right)\mathrm{d}u\mathrm{d}v$$

ovvero la densità congiunta è il prodotto delle densità marginali. Questo prova l'indipendenza di U e V e dunque la prima parte della tesi.

Per dimostrare la seconda parte della tesi, osserviamo che essa è sicuramente vera per $n-1$, grazie al Teorema 3.3.3. Procediamo per induzione e riconsideriamo ora la formula (3.2) e supponiamo che essa non sia relativa al campione $X_1, \ldots, X_n$ ma alla sua versione standardizzata $Z_1, \ldots, Z_n$:

$$\sum_{i=1}^{n}\left(Z_i - \overline{Z}_n\right)^2 = (n-1)S_n^2 = (n-2)S_{n-1}^2 + \left(\sqrt{\frac{n-1}{n}}\left(Z_n - \overline{Z}_{n-1}\right)\right)^2.$$

Poiché il campione $Z_1, \ldots, Z_n$ è gaussiano standard, $\mathbb{P}_{Z_n - \overline{Z}_{n-1}} = N\left(0, \dfrac{n}{n-1}\right)$ dunque la v.a. $\sqrt{\dfrac{n-1}{n}}\left(Z_n - \overline{Z}_{n-1}\right)$ è gaussiana standard e quindi il suo quadrato segue una distribuzione di Pearson con un grado di libertà. D'altra parte, per induzione, $\sum_{i=1}^{n-1}\left(Z_i - \overline{Z}_{n-1}\right)^2 = (n-2)S_{n-1}^2(Z)$ segue una distribuzione di Pearson a $n-2$ gradi di libertà. Per il Teorema 3.3.2 otteniamo la tesi. □

Corollario 3.3.7. *Sia $X_1, \ldots, X_n$ 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

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

□

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

Si chiama così la distribuzione associata alla densità

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

Proprietà 3.3.2. 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 3.3.2. 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 3.3.8. *Se Z è una v.a. gaussiana standard, $\mathbb{P}_Z = N(0,1)$, se Y segue la distribuzione χ^2 con n gradi di libertà, $\mathbb{P}_Y = \chi^2_n$ 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: $\mathbb{P}_T = t(n)$.*

Dimostrazione. Possiamo scrivere $T = \varphi \circ (Y, Z)$ dove $\varphi\colon (y,z) \in \mathbb{R}^2 \mapsto \begin{cases} \dfrac{z\sqrt{n}}{y} & y > 0 \\ 0 & y \leq 0 \end{cases} \in \mathbb{R}$.

Sia dunque $\psi\colon \mathbb{R} \to \mathbb{R}$ una funzione di Borel non negativa.

$$\begin{aligned}
&\int_{\mathbb{R}} \psi(t)\mathbb{P}_T(\mathrm{d}t) = \int_{y>0, z\in\mathbb{R}} \psi\left(\frac{z\sqrt{n}}{\sqrt{y}}\right)\mathbb{P}_{Y,Z}(\mathrm{d}y\mathrm{d}z) \\
&= \int_{y>0,z\in\mathbb{R}} \psi\left(\frac{z\sqrt{n}}{\sqrt{y}}\right)\frac{1}{\sqrt{2\pi}}\frac{1}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{1}{2}\right)^{\frac{n}{2}} y^{\frac{n}{2}-1}\exp\left(\frac{-y}{2}\right)\exp\left(\frac{-z^2}{2}\right)\mathrm{d}y\mathrm{d}z
\end{aligned}$$

con il cambio di variabile $t = \dfrac{z\sqrt{n}}{\sqrt{y}}$, $z = \dfrac{t\sqrt{y}}{\sqrt{n}}$, $\mathrm{d}z = \dfrac{\sqrt{y}}{\sqrt{n}}\mathrm{d}t$,

$$= \int_{\mathbb{R}} \psi(t)\frac{1}{\sqrt{2n\pi}}\frac{1}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{1}{2}\right)^{\frac{n}{2}}\left(\int_0^{+\infty} y^{\frac{1}{2}}y^{\frac{n}{2}-1}\exp\left(\frac{-y}{2}\right)\exp\left(\frac{-yt^2}{2n}\right)\right)\mathrm{d}t$$

con il cambio di variabile $u = \dfrac{y}{2}\left(1+\dfrac{t^2}{n}\right)$, $y = 2u\left(1+\dfrac{t^2}{n}\right)^{-1}$, $\mathrm{d}y = 2\left(1+\dfrac{t^2}{n}\right)^{-1}\mathrm{d}u$,

$$\begin{aligned}
&= \int_{\mathbb{R}} \psi(t)\frac{1}{\sqrt{2n\pi}}\frac{1}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{1}{2}\right)^{\frac{n}{2}}\left(\int_0^{+\infty}(2u)^{\frac{n+1}{2}-1}\exp(-u)\left(1+\frac{t^2}{n}\right)^{\frac{-(n+1)}{2}}\mathrm{d}u\right)\mathrm{d}t \\
&= \int_{\mathbb{R}} \psi(t)\frac{1}{\sqrt{2n\pi}}\frac{1}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{1}{2}\right)^{\frac{n}{2}}\left(1+\frac{t^2}{n}\right)^{\frac{-(n+1)}{2}}\Gamma\left(\frac{n+1}{2}\right)\mathrm{d}t
\end{aligned}$$

da cui la tesi. □

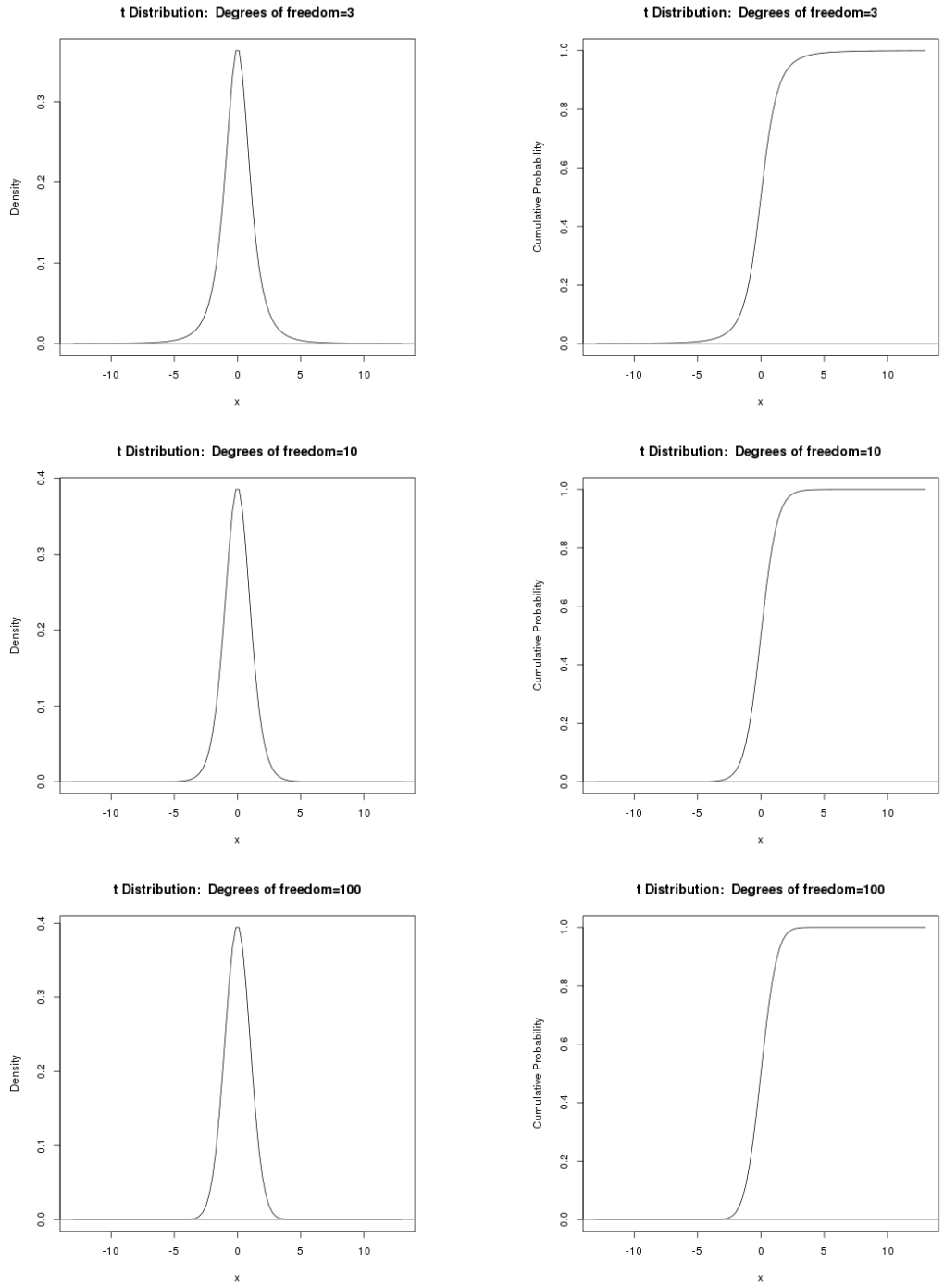

Figura 3.4: $t(3)$, $t(10)$, $t(100)$, densità e funzione di ripartizione

Corollario 3.3.9. *Se $X_1, \ldots, X_n$ è un campione statistico gaussiano di numerosità n, valore atteso 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: $\mathbb{P}_T = t(n-1)$.

Dimostrazione. Basta applicare il teorema 3.3.8 con $Z = \dfrac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ e $Y = V = (n-1)\dfrac{S^2}{\sigma^2}$. □

4. Stimatori di massima versosimiglianza

Sia $X_1,\ \ldots,\ X_n$ un campione statistico e sia $Y = \varphi\,(X_1,\ \ldots,\ X_n)$ una sua statistica. Se Y ha lo scopo di stimare un parametro θ della distribuzione del campione, diciamo che Y è uno *stimatore del parametro* θ.

Supponiamo di conoscere la distribuzione del campione a meno di un parametro θ e supponiamo che tale distribuzione sia discreta o assolutamente continua e dunque dotata di densità (discreta o meno). Tale densità dipenderà dal parametro θ e la indico col simbolo $g(x|\theta)$.La distribuzione congiunta si indica col simbolo $f(x_1,\ \ldots,\ x_n|\theta)$ e sappiamo che

$$f(x_1,\ \ldots,\ x_n|\theta) = g(x_1|\theta)\cdot\ldots\cdot g(x_n|\theta) = \prod_{i=1}^{n} g(x_i|\theta).$$

Interpreto $f(x_1,\ \ldots,\ x_n|\theta)$ come la *plausibilità* che la n-upla $x_1,\ \ldots,\ x_n$ si realizzi nel campione empirico quando il parametro incognito prende il valore θ. Sappiamo infatti che, se f è continua nel punto $(x_1,\ \ldots,\ x_n, \theta)$, allora

$$\begin{aligned}\mathbb{P}\left(\|X_1 - x_1\| < \frac{\delta}{2}, \ldots, \|X_n - x_n\| < \frac{\delta}{2}\right)\\ = \mathbb{P}\left((X_1,\ \ldots,\ X_n) \in \prod_{i=1}^{n}\left(x_i - \frac{\delta}{2}, x_i + \frac{\delta}{2}\right)\right) \simeq f\,(x_1,\ \ldots,\ x_n|\theta)\,\delta^n\end{aligned}$$

Dunque: dato il campione empirico $x_1,\ \ldots,\ x_n$, cerco $\widehat{\theta} = \widehat{\theta}(x_1,\ \ldots,\ x_n)$ che massimizza la funzione $f(x_1,\ \ldots,\ x_n|\theta)$. La statistica $\widehat{\theta} = \widehat{\theta}(X_1,\ \ldots,\ X_n)$ si dirà *stimatore di massima verosimiglianza del parametro* θ.

Osservazione 4.0.1. Poiché la funzione $\ln\colon (0,+\infty) \to \mathbb{R}$ è strettamente monotona crescente, massimizzare $f(x, n_1,\ \ldots,\ x, n_|\theta) = \prod_{i=1}^{n} g(x_i|\theta)$ equivale a massimizzare la funzione $\ln f(x, n_1,\ \ldots,\ x, n_|\theta) = \sum_{i=1}^{n} \ln g(x_i|\theta)$ e si ha

$$\frac{\partial}{\partial\theta}\ln f(x, n_1,\ \ldots,\ x, n_|\theta) = \frac{\partial}{\partial\theta}\sum_{i=1}^{n}\ln g(x_i|\theta) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln g(x_i|\theta) = \sum_{i=1}^{n}\frac{1}{g(x_i|\theta)}\frac{\partial g(x_i|\theta)}{\partial\theta}$$

4.1 Distribuzione di Bernoulli

Sappiamo che la distribuzione di Bernoulli dipende dal solo parametro $p = \mathbb{P}X = 1$. Sia dunque $X_1,\ \ldots,\ X_n$ un campione statistico di Bernoulli di parametro incognito $p \in [0,1]$.

Realizzo n prove di Bernoulli e ottengo il campione empirico $x_1,\ \ldots,\ x_n$, $x_i \in \{0,1\}$.

$$f(x_1,\ \ldots,\ x_n|p) = \mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n\right) = p^k(1-p)^{n-k},$$
$$k = k(x_1,\ \ldots,\ x_n) := \sum_{i=1}^{n} x_i.$$

Abbiamo

$$\frac{\partial f}{\partial p} = kp^{k-1}(1-p)^{n-k} - (n-k)p^k(1-p)^{n-k-1}$$
$$= p^{k-1}(1-p)^{n-k-1}(k-np) \gtreqless 0 \iff k - np \gtreqless 0 \iff p \lesseqgtr \frac{k}{n}.$$

Poiché $k = \sum_{i=1}^{n} x_i$, lo stimatore di massima verosimiglianza per il parametro p è $\frac{\sum_{i=1}^{n} X_i}{n}$ cioè la media campionaria $\overline{X}$.

4.2 Distribuzione di Poisson

La distribuzione di Poisson è concentrata sugli interi nonnegativi e dipende da un solo parametro:

$$g(x|\lambda) = \mathrm{e}^{-\lambda}\,\frac{\lambda^x}{x!}, \qquad x = 0, 1, \ldots$$

e dunque

$$f(x_1,\ \ldots,\ x_n|\lambda) = \prod_{i=1}^{n}\left(\mathrm{e}^{-\lambda}\,\frac{\lambda^{x_i}}{x_i!}\right)$$

$$\ln f(x_1,\ \ldots,\ x_n|\lambda) = \sum_{i=1}^{n} \ln\left(\mathrm{e}^{-\lambda}\,\frac{\lambda^{x_i}}{x_i!}\right)$$
$$= \sum_{i=1}^{n}\left(-\lambda + x_i\ln(\lambda) - \ln(x_i!)\right) = -n\lambda + n\overline{x}\ln(\lambda) - \sum_{i=1}^{n}\ln(x_i!)$$

Da cui

$$\frac{\partial}{\partial\lambda}\ln f(x_1,\ \ldots,\ x_n|\lambda) = n\left(-\lambda + \frac{\overline{x}}{\lambda}\right) \gtreqless 0 \iff \lambda \lesseqgtr \overline{x}.$$

Quindi anche in questo caso o stimatore di massima verosimiglianza per il parametro λ è la media campionaria $\overline{X}$.

4.3 Distribuzione gaussiana

In questo caso la densità dipende da due parametri, $\mu \in \mathbb{R}$ e $\sigma > 0$:

$$f(x_1,\ \ldots,\ x_n|\mu,\sigma) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)$$
$$= (2\pi)^{\frac{-n}{2}}(\sigma)^{-n}\exp\left(\frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right)$$

cosicché

$$\begin{aligned}\ln f(x_1,\ \ldots,\ x_n|\mu,\sigma) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\ &= -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.\end{aligned}$$

Si ha quindi

$$\begin{aligned}\frac{\partial}{\partial\mu}\ln f(x_1,\ \ldots,\ x_n|\mu,\sigma) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = n\left(\overline{x}-\mu\right), \\ \frac{\partial}{\partial\sigma}\ln f(x_1,\ \ldots,\ x_n|\mu,\sigma) &= \frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{1}{\sigma^3}\left(-n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2\right).\end{aligned}$$

Dunque le due derivate parziali si annullano contemporaneamente se e solo se

$$\mu = \overline{x}, \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

Dunque la media campionaria $\overline{X}$ è uno stimatore di massima verosimiglianza per il valore atteso μ mentre $\dfrac{n-1}{n}S^2$ è uno stimatore di massima verosimiglianza per la varianza σ^2.

4.4 Distribuzione uniforme su un intervallo

Se (a,b) è l'intervallo, allora la densità del campione è

$$g(x|a,b) = \begin{cases}\frac{1}{b-a} & x\in[a,b], \\ 0 & \text{altrimenti}\end{cases}$$

da cui

$$f(x_1,\ \ldots,\ x_n|a,b) = \begin{cases}\frac{1}{(b-a)^n} & x_i\in[a,b] \quad \forall i=1,\ldots,n, \\ 0 & \text{altrimenti.}\end{cases}$$

Devo massimizzare $\frac{1}{(b-a)^n}$ con il vincolo $a\le x_i\le b$ per ogni $i=1,\ldots,n$. Devo dunque minimizzare la lunghezza dell'intervallo $b-a$ con il vincolo $a\le x_i\le b$ per ogni $i=1,\ldots,n$. È dunque

$$a = \min\{x_1,\ \ldots,\ x_n\}, \qquad b = \min\{x_1,\ \ldots,\ x_n\}.$$

Dunque

$$\min\{X_1,\ \ldots,\ X_n\}, \qquad \max\{X_1,\ \ldots,\ X_n\}$$

sono stimatori di massima verosimiglianza rispettivamente per l'estremo inferiore e per l'estremo superiore dell'intervallo.

5. Intervalli di confidenza

La media campionaria e la varianza campionaria ci offrono una stima dei parametri valore atteso 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 5.0.1 (Intervallo di confidenza)**.** Sia $X_1, \ldots, X_n$ un campione statistico e sia θ un parametro (ignoto) che caratterizza la distribuzione del campione.
Siano $L_i = l_i(X_1, \ldots, X_n)$ e $L_s = l_s(X_1, \ldots, X_n)$ 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 5.0.1.

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 Chebychev ci ha fornito un intervallo di confidenza per il valore atteso 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 Chebychev 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 $\left(\overline{X} - \dfrac{\sigma}{\sqrt{\alpha}}, \overline{X} + \dfrac{\sigma}{\sqrt{\alpha}}\right)$ è un intervallo di confidenza di livello $1-\alpha$ per il valore atteso μ del campione.

5.1 Stima per intervalli del valore atteso di campioni gaussiani

5.1.1 Campione gaussiano di cui è nota la varianza

Intervallo bilaterale

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

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

$$\begin{aligned}
&\mathbb{P}\left(|Z| \leq z_{1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq Z \leq z_{1-\frac{\alpha}{2}}\right) \\
&= \mathbb{P}\left(Z \leq z_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(Z \leq -z_{1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(Z \leq z_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(Z \leq z_{\frac{\alpha}{2}}\right) \qquad (5.1) \\
&= \Phi\left(z_{1-\frac{\alpha}{2}}\right) - \Phi\left(z_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.
\end{aligned}$$

Sappiamo che $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ e che dunque $\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$. Applichiamo quindi la disuguaglianza (5.1) a $\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$. Si ha:

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

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 il valore atteso μ del campione.

Osservazione 5.1.1 (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

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

ovvero

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

Intervallo unilaterale superiore

Sia $Z \sim N(0,1)$. 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 $Z \sim N(0,1)$. 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 $1-\alpha$.

5.1.2 Campione gaussiano di cui non è nota la varianza

Intervallo bilaterale

Sia $X_1, \ldots, X_n$ un campione gaussiano di valore atteso 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

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

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 il valore atteso μ del campione.

Intervallo unilaterale superiore

Sappiamo che

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

Abbiamo dunque

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

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

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

Abbiamo dunque

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

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 $1-\alpha$.

5.2 Stima per intervalli della varianza di campioni gaussiani

Intervallo bilaterale

Sia $X_1, \ldots, X_n$ un campione gaussiano di valore atteso 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 5.2.1. $\chi^2_{n-1,\alpha} > 0$ 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

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

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

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

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 5.2.1. Calcoliamo gli intervalli di confidenza per il carattere Totpor dei dati tratti da [2], nell'ipotesi che si tratti della realizzazione di v.a. normali.

```
> setwd("~/Documents/didattica/2017-18_analisi_reale/alcuni_appunti/esempio_statistica")
>
> library(readr)
>
> table2 <- read_delim("~/Documents/didattica/2017-18_analisi_reale/alcuni_appunti/
table2.csv", "\t", escape_double = FALSE, trim_ws = TRUE)
Parsed with column specification:
cols(
  Code = col_character(),
  Totpor = col_double(),
  PRA = col_double(),
  PV = col_double(),
  Densi = col_double(),
  TenStr = col_double(),
  CO2SBW = col_double(),
  FirTemp = col_integer()
)
>
> ## 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)}
>
> # 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)
+ }
>
> # 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)
+ }
>
>
> numSummary(table2[,c("Totpor", "PRA", "PV", "Densi", "TenStr", "CO2SBW", "FirTemp")],
+  statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
              mean         sd      0%      25%     50%      75%    100%  n NA
Totpor   40.1193548  7.0371760  26.850  36.0550  40.900  44.4200  54.640 31  0
PRA       0.6732581  0.4760389   0.158   0.4220   0.622   0.7305   2.657 31  0
PV       55.3290323 28.5498417  10.200  30.4500  59.400  80.7000  88.600 31  0
Densi     1.6929032  0.1701214   1.340   1.5600   1.680   1.8150   2.020 31  0
TenStr    0.6092258  0.3143682   0.143   0.4065   0.527   0.7165   1.405 31  0
CO2SBW    0.5816667  0.5259152   0.050   0.2900   0.390   0.4950   1.960 30  1
FirTemp 764.8387097 52.9698636 730.000 740.0000 740.000 750.0000 960.000 31  0
>
> bilat.norm(table2$Totpor, 7.04, .9)
[1] 38.03957 42.19914
> bilat.norm(table2$Totpor, 7.04, .95)
[1] 37.64113 42.59758
>
> bilat.stud(table2$Totpor, .9)
[1] 37.97416 42.26455
> bilat.stud(table2$Totpor, .95)
[1] 37.53810 42.70061
>
> bilat.chi(table2$Totpor, .9)
[1] 33.94002 80.33757
> bilat.chi(table2$Totpor, .95)
[1] 31.62366 88.48047
>
```

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

Per chiarirsi le idee vediamo prima un esempio.

Esempio 6.0.1. Ho una moneta. Voglio verificare se è bilanciata o meno. La lancio n volte. Pongo $X_i = \begin{cases} 1 & \text{se all'}i\text{-esimo lancio esce testa,} \\ 0 & \text{se all'}i\text{-esimo lancio esce croce.} \end{cases}$, $i = 1, \dots, n$.

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:

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

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

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 $\left| \displaystyle\sum_{i=1}^{n} x_i - \frac{n}{2} \right| \geq n\varepsilon$. 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 i.i.d con $\mathbb{P}_{X_i} = B(p)$, 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 $p = 0.5$ cosicché $\mathbb{P}_Y = B(n, 0.5)$ 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

```
> ## definisco la funzione che calcola
> ## la probabilitÃ  di errore di prima specie
> alpha.binom = function(n,p,tolle) {
+ infe = n*(p - tolle);
+ supe = n*(p + tolle);
+ supep = supe;
+ if(supe == floor(supe)) supep = supe-1;
+ infe = round(infe, digits = 0);
+ c(floor(infe), floor(supe),
+   pbinom(infe, size=n, prob=p, lower.tail=TRUE) +
+         pbinom(supep, size=n, prob=p, lower.tail=FALSE))
+ }
> alpha.binom(50, .5, .1)
[1] 20.0000000 30.0000000  0.2026388
> alpha.binom(100, .5, .1)
[1] 40.00000000 60.00000000  0.05688793
> alpha.binom(200, .5, .1)
[1] 8.000000e+01 1.200000e+02 5.685156e-03
> alpha.binom(300, .5, .1)
[1] 1.2000e+02 1.8000e+02 6.3422e-04
> alpha.binom(400, .5, .1)
[1] 1.600000e+02 2.400000e+02 7.426568e-05
> alpha.binom(500, .5, .1)
[1] 2.000000e+02 3.000000e+02 8.940067e-06
> alpha.binom(50, .5, .05)
[1] 22.0000000 27.0000000  0.4798877
> alpha.binom(100, .5, .05)
[1] 45.0000000 55.0000000  0.3197273
> alpha.binom(200, .5, .05)
[1]  90.0000000 110.0000000   0.1581653
> alpha.binom(300, .5, .05)
[1] 135.0000000 165.0000000   0.0939037
> alpha.binom(400, .5, .05)
[1] 180.00000000 220.00000000   0.04563548
```

```
> alpha.binom(500, .5, .05)
[1] 225.00000000 275.00000000   0.02832616
```

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 $\mathbb{P}_Y = B(n,p)$ 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 6.0.2. 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 $\mathbb{P}_Y = B(n,p)$ per qualche $p \leq 0.5$. Indico F_Y^p la sua funzione di ripartizione
Vediamo alcuni casi

```
  > ## definisco la funzione che calcola il primo valore
> ## che rifiuto e
> ## la probabilitÃ  di errore di prima specie
> alpha.binom.uni = function(n,p,tolle) {
+ supe = n*(p + tolle);
+ supep = supe;
+ if(supe == floor(supe)) supep = supe-1;
+ c(floor(supe), pbinom(supep, size=n, prob=p, lower.tail=FALSE))
+ }
> alpha.binom.uni(50, .5, .1)
[1] 30.0000000  0.1013194
```

```
> ppp =numeric(0)
> fff =numeric(0)
> beta.p <- matrix(0, nrow = 1000, ncol = 2, byrow = FALSE)
> for (i in 1:1000)  {
+     ppp[i] <- i*0.5/1000
+     fff[i] <- pbinom(c(274), size=500, prob=ppp[i], lower.tail=TRUE)
+     - pbinom(c(225), size=500, prob=ppp[i], lower.tail=TRUE)
+     beta.p[i,1] <- round(ppp[i],6)
+     beta.p[i,2] <- round(fff[i],6)
+         }
> write.csv(beta.p, "betadip.csv", row.names = FALSE)
```

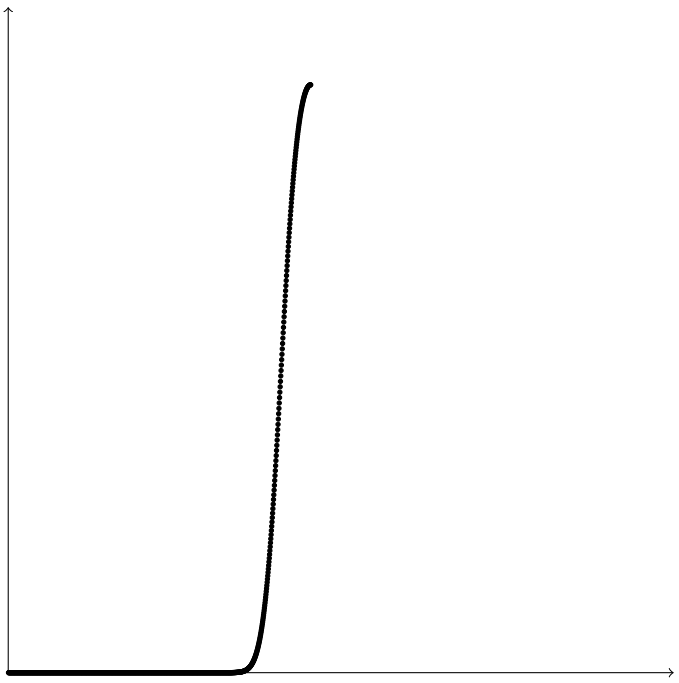

Figura 6.1: $\beta(p)$

```
> alpha.binom.uni(100, .5, .1)
[1] 60.00000000  0.02844397
> alpha.binom.uni(200, .5, .1)
[1] 1.200000e+02 2.842578e-03
> alpha.binom.uni(300, .5, .1)
[1] 1.8000e+02 3.1711e-04
> alpha.binom.uni(400, .5, .1)
[1] 2.400000e+02 3.713284e-05
> alpha.binom.uni(500, .5, .1)
[1] 3.000000e+02 4.470033e-06
> alpha.binom.uni(50, .5, .05)
[1] 27.0000000  0.2399438
> alpha.binom.uni(100, .5, .05)
[1] 55.0000000  0.1356265
> alpha.binom.uni(200, .5, .05)
[1] 110.00000000   0.06868333
> alpha.binom.uni(300, .5, .05)
[1] 165.00000000   0.04695185
> alpha.binom.uni(400, .5, .05)
[1] 220.00000000   0.02011537
> alpha.binom.uni(500, .5, .05)
[1] 275.00000000   0.01416308
```

6.1 Principi generali di un test statistico

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

1. Si definisce l'insieme delle distribuzioni *compatibili* con il campione $X_1, \ldots, X_n$.
2. Si definisce l'ipotesi da testare, detta *ipotesi nulla* (si indica col simbolo H_0). Le ipotesi si possono suddividere in due grandi famiglie:
   - **ipotesi parametriche**: la distribuzione del campione è nota a meno di un parametro θ, scalare o vettoriale. La formula generale di un'ipotesi parametrica è dunque
     $$H_0 : \qquad \theta \in \Theta_0 \subset \Theta$$
     ovvero: il parametro θ appartiene ad uno specificato sottoinsieme Θ_0 del dominio ammissibile per il parametro Θ.
   - **ipotesi non parametriche**: sono ipotesi sul tipo di distribuzione del campione oppure ipotesi che riguardano popolazioni differenti. La formulazione generale di una ipotesi non parametrica è del tipo
     $$H_0 : \qquad F(x) \in \mathcal{F}_0 \subset \mathcal{F}$$
     ovvero: la legge F del campione appartiene ad uno specificato sottoinsieme della famiglia delle leggi ammissibili.

In entrambi i casi l'ipotesi si dice *semplice* se Θ_0 o $\mathcal{F}_0$ è costituito da un solo elemento. Si dice *composta* altrimenti.

3. Si definisce l'ipotesi alternativa H_A che è da considerarsi valida quando si rifiuta H_0.

$$\begin{aligned} H_A: &\quad \theta \in \Theta_1, \quad \Theta_1 := \Theta \setminus \Theta_0 \quad \text{nel caso parametrico,} \\ H_A: &\quad F(x) \in \mathcal{F}_1 \quad \mathcal{F}_1 := \mathcal{F} \setminus \mathcal{F}_0 \quad \text{nel caso non parametrico.} \end{aligned}$$

4. Si definisce una statistica $\varphi(X_1, \ldots, X_n)$ con distribuzione nta quando H_0 è vera.

5. Si suddivide lo spazio $\mathcal{G}$ delle possibili osservazioni in due insiemi disgiunti:

   - $\mathcal{A}$ detta *regione di accettazione di* H_0;
   - $\mathcal{C} := \mathcal{G} \setminus \mathcal{A}$ detta *regione di rifiuto di* H_0 o *regione critica*.

6. Si formula la regola di decisione:

   - accetto H_0 se $\varphi(x_1, \ldots, x_n) \in \mathcal{A}$;
   - rifiuto H_0 se $\varphi(x_1, \ldots, x_n) \notin \mathcal{A}$, ovvero se $\varphi(x_1, \ldots, x_n) \in \mathcal{C}$.

Diciamo che commettiamo *errore di prima specie* se rigettiamo H_0 quando essa in realtà è vera e chiamiamo *livello di significatività del test* la probabilità che ciò accada:

$$\alpha := \mathbb{P}\left(\varphi(X_1, \ldots, X_n) \in \mathcal{C} | H_0\right).$$

Il valore $1 - \alpha$ è detto *livello di fiducia del test*.

Diciamo invece che commettiamo *errore di seconda specie* se accettiamo H_0 quando esssa è falsa. Indichiamo con β la probabilità che cio accada:

$$\beta := \mathbb{P}\left(\varphi(X_1, \ldots, X_n) \in \mathcal{A} | H_A\right).$$

Il valore $1 - \beta$ è detto *potenza del test*. (Vedremo negli esempi successivi relativi a test parametrici che se H_A è un'ipotesi composta, allora β è una funzione $\beta(\theta)$, $\theta \in \Theta_1$.

Come già detto, è prioritario *limitare* la probabilità di commettere errore di prima specie, cioè di limitare la probabilità di rifiutare l'ipotesi nulla quando essa è vera.

6.2 Test parametrici per campioni gaussiani

6.2.1 Test d'ipotesi per il valore atteso di campioni gaussiani di cui è nota la varianza

Test bilaterale

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

$$H_0: \quad \mu = \mu_0, \qquad H_A: \quad \mu \neq \mu_0.$$

Sappiamo che $\mathbb{P}_{X_i} = N\left(\mu_0, \sigma^2\right)$ se e solo se $\mathbb{E}\left[\overline{X}\right] = \mu_0$. 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 significatività (cioè la probabilità di commettere un errore di prima specie) è allora

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

Ma se H_0 è vera, $\mathbb{P}_{\overline{X}} = N\left(\mu_0, \frac{\sigma^2}{n}\right)$ e $Z := \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ ha distribuzione gaussiana standard $N(0,1)$. Dunque

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

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

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

Presi i dati $x_1, \ldots, x_n$, sia $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ la loro media:

$$\text{accetto } H_0 \text{ se } |\overline{x} - \mu_0| < \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \text{ e la rifiuto altrimenti.}$$

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

$$\begin{aligned}
\beta(\mu) &= \mathbb{P}\left(\left|\overline{X} - \mu_0\right| < \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} | \mathbb{E}[X_i] = \mu\right) \\
&= \mathbb{P}\left(\mu_0 - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} < \overline{X} < \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} | \mathbb{E}[X_i] = \mu\right) \\
&= \mathbb{P}\left(\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} - z_{1-\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}} | \mathbb{E}[X_i] = \mu\right) \\
&= \Phi\left(\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}}\right) - \Phi\left(\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{\frac{\alpha}{2}}\right).
\end{aligned} \tag{6.1}$$

Distinguiamo due casi

1. $\mu > \mu_0$
   In questo caso $\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} < 0$ dunque $\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{\frac{\alpha}{2}} < z_{\frac{\alpha}{2}}$ 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. Abbiamo dunque

$$\beta(\mu) \sim \Phi\left(\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}}\right).$$

In particolare

$$\sup_{\mu>\mu_0} \beta(\mu) \sim \Phi\left(z_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}.$$

Supponiamo di voler fissare (oltre ad α) anche $\beta(\mu) = \widehat{\beta}$, per un qualche μ fissato. Con la semplificazione fatta dalla (6.1) otteniamo $\widehat{\beta} \geq \Phi\left(\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}}\right)$. L'unica quantità che possiamo trattare è la numerosità n. Risolvendo l'equazione rispetto a n otteniamo

$$z_{\widehat{\beta}} \geq \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}}$$

e dunque

$$\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\widehat{\beta}} + z_{\frac{\alpha}{2}},$$

cioè

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

2. $\mu < \mu_0$
   In questo caso $\frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} < 0$ e scriviamo la (6.1) 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. Abbiamo dunque Abbiamo dunque

$$\beta(\mu) \sim \Phi\left(\frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}}\right).$$

   In particolare

$$\sup_{\mu<\mu_0} \beta(\mu) \sim \Phi\left(z_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}.$$

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

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

Test unilaterale inferiore con H_0 semplice

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

$$H_0: \quad \mu = \mu_0 \qquad H_A: \quad \mu > \mu_0.$$

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

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

Poiché, se $H - 0$ è vera si ha $\mathbb{P}_{\overline{X}} = N\left(\mu_0, \frac{\sigma^2}{n}\right)$ e $Z := \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ ha distribuzione $N(0,1)$, si ha

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

Dunque scelgo $\varepsilon = \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$. Presi i dati $x_1, \ldots, x_n$, sia dunque $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ la loro media.

$$\text{Accetto } H_0 \text{ se } \overline{x} < \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} \text{ e la rifiuto altrimenti.}$$

Test unilaterale inferiore con H_0 composta

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

$$H_0: \quad \mu \leq \mu_0 \qquad H_A: \quad \mu > \mu_0.$$

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

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

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

$$\begin{aligned}
&\mathbb{P}\left(\overline{X} \geq \mu_0 + \varepsilon | \mathbb{E}\left[\overline{X}\right] = \mu\right) = \mathbb{P}\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \geq \frac{\mu_0 - \mu + \varepsilon}{\frac{\sigma}{\sqrt{n}}} | \mathbb{E}\left[\overline{X}\right] = \mu\right) \\
&= \mathbb{P}\left(Z \geq \frac{(\mu_0 - \mu + \varepsilon)\sqrt{n}}{\sigma}\right) = 1 - \mathbb{P}\left(Z \leq \frac{(\mu_0 - \mu + \varepsilon)\sqrt{n}}{\sigma}\right) \\
&= 1 - \Phi\left(\frac{(\mu_0 - \mu + \varepsilon)\sqrt{n}}{\sigma}\right) \leq 1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right).
\end{aligned}$$

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

$$\mathbb{P}\left(\overline{X} > \mu_0 + \varepsilon | \mathbb{E}\left[\overline{X}\right] = \mu\right) \leq \alpha \qquad \forall \mu \leq \mu_0$$

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 $x_1, \ldots, x_n$, sia dunque $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ la loro media.

$$\text{Accetto } H_0 \text{ se } \overline{x} < \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} \text{ e la rifiuto altrimenti.}$$

Test unilaterale superiore con H_0 semplice

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

$$H_0 : \qquad \mu = \mu_0 \qquad\qquad H_a \colon \mu < \mu_0$$

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

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

Poiché, se H_0 è vera, $\mathbb{P}_{\overline{X}} = N\left(\mu_0, \frac{\sigma^2}{n}\right)$, e $Z := \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ ha distribuzione $N(0,1)$, si ha

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

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

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

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

$$\text{Accetto } H_0 \text{ se } \overline{x} > \mu_0 - \frac{\sigma}{\sqrt{n}} z_{1-\alpha} \text{ e la rifiuto altrimenti.}$$

Test unilaterale superiore con H_0 composta

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

$$H_0 : \qquad \mu \geq \mu_0 \qquad\qquad H_A : \mu < \mu_0.$$

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

$$\mathbb{P}\left(\overline{X} \leq \mu_0 - \varepsilon | \mathbb{E}\left[\overline{X}\right] \leq \mu_0\right).$$

Poiché, se $\mathbb{P}_{X_i} = N\left(\mu, \sigma^2\right)$ si ha $\mathbb{P}_{\overline{X}} = N\left(\mu, \dfrac{\sigma^2}{n}\right)$, e $Z := \dfrac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ ha distribuzione $N(0,1)$, abbiamo anche

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

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

$$\mathbb{P}\left(\overline{X} \leq \mu_0 - \varepsilon | \mathbb{E}\left[\overline{X}\right] = \mu \geq \mu_0\right) \leq \alpha \qquad \forall \mu \geq \mu_0$$

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 $x_1, \ldots, x_n$, sia dunque $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ la loro media.

$$\text{Accetto } H_0 \text{ se } \overline{x} > \mu_0 - \frac{\sigma}{\sqrt{n}} z_{1-\alpha} \text{ e la rifiuto altrimenti.}$$

6.2.2 Campione gaussiano di cui non è nota la varianza

Test bilaterale

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

$$H_0 : \qquad \mu = \mu_0 \qquad\qquad H_A : \qquad \mu \neq \mu_0/$$

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 $\mathbb{E}\left[\dfrac{(\overline{X}-\mu_0)\sqrt{n}}{S}\right] = \mathbb{E}\left[\overline{X}-\mu_0\right]\sqrt{n}\,\mathbb{E}\left[\dfrac{1}{\sqrt{S^2}}\right] = 0$. Dunque considero $t := \dfrac{(\overline{x}-\mu_0)\sqrt{n}}{s}$ e accetto l'ipotesi nulla H_0 se $|t| \leq \varepsilon$.

Sappiamo che, se $\mu = \mu_0$, allora $T := \dfrac{(\overline{X}-\mu_0)\sqrt{n}}{S}$ ha distribuzione $t(n-1)$. Il livello di di significatività è allora $\alpha = \mathbb{P}\left(|T| \geq \varepsilon\right)$ e si ha

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

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 $x_1, \ldots, x_n$, dunque accetto H_0 se $|t| \leq t_{n-1,1-\frac{\alpha}{2}}$ e la rifiuto altrimenti, ovvero

$$\text{accetto } H_0 \text{ se } \mu_0 - \frac{t_{n-1,1-\frac{\alpha}{2}}\, s}{\sqrt{n}} \leq \overline{x} \leq \mu_0 + \frac{t_{n-1,1-\frac{\alpha}{2}}\, s}{\sqrt{n}} \text{ e la rifiuto altrimenti.}$$

Test unilaterale superiore con ipotesi nulla semplice

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

$$H_0: \quad \mu = \mu_0, \qquad\qquad H_0: \quad \mu > \mu_0$$

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

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

dove $\mathbb{P}_T = t(n-1)$. 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 $x_1, \ldots, x_n$, 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

$$\text{accetto } H_0 \text{ se } \overline{x} \leq \mu_0 + \frac{t_{n-1,1-\alpha}\, s}{\sqrt{n}} \text{ e la rifiuto altrimenti.}$$

Test unilaterale superiore con ipotesi nulla composta

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

$$H_0: \quad \mu \leq \mu_0, \qquad\qquad H_0: \quad \mu > \mu_0$$

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

$$\mathbb{P}\left(\frac{(\overline{X}-\mu_0)\sqrt{n}}{S} > \varepsilon | \mathbb{E}\left[X\right] = \mu \leq \mu_0\right).$$

Se H_0 è vera, allora $\mathbb{E}\left[X\right] = \mu \leq \mu_0$ e dunque

$$\frac{(\overline{X}-\mu_0)\sqrt{n}}{S} \leq \frac{(\overline{X}-\mu)\sqrt{n}}{S} =: T, \qquad \mathbb{P}_T = t(n-1).$$

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

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

Se vogliamo controllare 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 $x_1, \ldots, x_n$, 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

$$\text{accetto } H_0 \text{ se } \overline{x} \leq \mu_0 + \frac{t_{n-1,1-\alpha}\, s}{\sqrt{n}} \text{ e la rifiuto altrimenti.}$$

Test unilaterale inferiore con ipotesi nulla semplice

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

$$H_0: \quad \mu = \mu_0, \qquad\qquad H_A: \quad \mu < \mu_0.$$

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

$$\alpha = \mathbb{P}\left(\frac{(\overline{X}-\mu_0)\sqrt{n}}{S} < -\varepsilon | \mu = \mu_0\right) = \mathbb{P}\left(T < -\varepsilon\right) = F_T(-\varepsilon)$$

dove $\mathbb{P}_T = t(n-1)$. Se vogliamo stabilire il livello di significatività α dovremmo scegliere ε in modo che

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

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

Presi i dati $x_1, \ldots, x_n$, 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.

Test unilaterale inferiore con ipotesi nulla composta

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

$$H_0: \quad \mu \geq \mu_0, \qquad H_A: \quad \mu < \mu_0.$$

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

$$\mathbb{P}\left(\frac{(\overline{X}-\mu_0)\sqrt{n}}{S} < -\varepsilon | \mathbb{E}\left[X\right] = \mu \geq \mu_0\right).$$

Se H_0 è vera, allora $\mathbb{E}\left[X\right] = \mu \geq \mu_0$ e dunque

$$\frac{(\overline{X}-\mu_0)\sqrt{n}}{S} \geq \frac{(\overline{X}-\mu)\sqrt{n}}{S} =: T, \qquad \mathbb{P}_T = t(n-1).$$

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 $\mu \geq \mu_0$ si ha

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

Se vogliamo controllare 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 $x_1, \ldots, x_n$, 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

$$\text{accetto } H_0 \text{ se } \overline{x} \geq \mu_0 - \frac{t_{n-1,1-\alpha}\, s}{\sqrt{n}} \text{ e la rifiuto altrimenti.}$$

6.3 Test d'ipotesi per la varianza di campioni gaussiani

Test bilaterale

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

$$H_0: \quad \sigma^2 = \sigma_0^2 \qquad H_A: \quad \sigma^2 \neq \sigma_0^2$$

H_0 è vera se e solo se $\mathbb{E}\left[S^2\right] = \sigma_0^2$ ovvero se e solo se $\mathbb{E}\left[\dfrac{S^2}{\sigma_0^2}\right] = 1$. 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. Sappiamo che se H_0 è vera, allora la v.a. $V := \dfrac{(n-1)S^2}{\sigma^2}$ ha distribuzione χ^2_{n-1}.

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

Una possibile scelta è allora

$$\begin{aligned}\mathbb{P}\left(V > (n-1)(1+\varepsilon_2)\right) &= \frac{\alpha}{2} \qquad \text{cioè} \quad (n-1)(1+\varepsilon_2) = \chi^2_{n-1,1-\frac{\alpha}{2}} \\ \mathbb{P}\left(V < (n-1)(1-\varepsilon_1)\right) &= \frac{\alpha}{2} \qquad \text{cioè} \quad (n-1)(1-\varepsilon_1) = \chi^2_{n-1,\frac{\alpha}{2}}.\end{aligned}$$

Dunque accetto H_0 se $\chi^2_{n-1,\frac{\alpha}{2}} < \dfrac{(n-1)s^2}{\sigma_0^2} < \chi^2_{n-1,1-\frac{\alpha}{2}}$ ovvero

$$\text{accetto } H_0 \text{ se } \frac{\sigma_0^2}{n-1}\chi^2_{n-1,\frac{\alpha}{2}} < s^2 < \frac{\sigma_0^2}{n-1}\chi^2_{n-1,1-\frac{\alpha}{2}} \text{ e la rifiuto altrimenti.}$$

Test unilaterale inferiore con ipotesi semplice

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

$$H_0: \quad \sigma^2 = \sigma_0^2 \qquad\qquad H_A: \quad \sigma^2 > \sigma_0^2.$$

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

Se la varianza è σ_0^2, allora $V := \dfrac{(n-1)S^2}{\sigma_0^2}$ ha distribuzione χ^2_{n-1} e la probabilità di commettere errore di prima specie è

$$\mathbb{P}\left(\frac{S^2}{\sigma_0^2} > 1+\varepsilon | \sigma^2 = \sigma_0^2\right) = \mathbb{P}\left(\frac{(n-1)S^2}{\sigma_0^2} > (n-1)(1+\varepsilon) | \sigma^2 = \sigma_0^2\right) = 1 - F_V\left((n-1)(1+\varepsilon)\right).$$

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

$$\text{acccetto } H_0 \text{ se } s^2 < \frac{\sigma_0^2}{n-1}\chi^2_{n-1,1-\alpha} \text{ e la rifiuto altrimenti.}$$

Test unilaterale inferiore con ipotesi composta

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

$$H_0: \qquad \sigma^2 \leq \sigma_0^2 \qquad\qquad H_A: \qquad \sigma^2 > \sigma_0^2.$$

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 $V := \dfrac{(n-1)S^2}{\sigma^2}$ ha distribuzione χ^2_{n-1} e la probabilità di commettere errore di prima specie è

$$\begin{aligned}
&\mathbb{P}\left(\frac{S^2}{\sigma_0^2} > 1 + \varepsilon | \mathrm{Var}\left[X_i\right] = \sigma^2 \leq \sigma_0\right) = \mathbb{P}\left(\frac{(n-1)S^2}{\sigma^2} > \frac{\sigma_0^2}{\sigma^2}(n-1)(1+\varepsilon) | \mathrm{Var}\left[X_i\right] = \sigma^2 \leq \sigma_0\right) \\
&= \mathbb{P}\left(V > \frac{\sigma_0^2}{\sigma^2}(n-1)(1+\varepsilon)\right) = 1 - F_V\left(\frac{\sigma_0^2}{\sigma^2}(n-1)(1+\varepsilon)\right) \\
&\leq 1 - F_V\left((n-1)(1+\varepsilon)\right)
\end{aligned}$$

dove abbiamo usato la monotonia di F_V e il fatto che $\sigma^2 \leq \sigma_0^2$ implica $\dfrac{\sigma^2}{\sigma_0^2} \leq 1$.

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

$$\text{accetto } H_0 \text{ se } s^2 < \frac{\sigma_0^2}{n-1}\chi^2_{n-1,1-\alpha} \text{ e la rifiuto altrimenti.}$$

Test unilaterale superiore con ipotesi semplice

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

$$H_0: \qquad \sigma^2 = \sigma_0^2 \qquad\qquad H_0: \qquad \sigma^2 < \sigma_0^2.$$

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

Se H_0 è vera, allora $V := \dfrac{(n-1)S^2}{\sigma_0^2}$ ha distribuzione χ^2_{n-1} e la probabilità di commettere errore di prima specie è

$$\alpha = \mathbb{P}\left(\frac{S^2}{\sigma_0^2} < 1 - \varepsilon | \sigma^2 = \sigma_0^2\right) = F_V\left((n-1)(1-\varepsilon)\right).$$

Deve quindi essere

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

Dunque accetto l'ipotesi nulla H_0 se $\dfrac{(n-1)s^2}{\sigma_0^2} > \chi^2_{n-1,\alpha}$ ovvero

$$\text{accetto } H_0 \text{ se } s^2 > \frac{\sigma_0^2}{n-1}\chi^2_{n-1,\alpha} \text{ e la rifiuto altrimenti.}$$

Test unilaterale superiore con ipotesi composta

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

$$H_0: \quad \sigma^2 \geq \sigma_0^2 \qquad\qquad H_A: \quad \sigma^2 < \sigma_0^2.$$

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 $V := \dfrac{(n-1)S^2}{\sigma^2}$ ha distribuzione χ^2_{n-1} e la probabilità di commettere errore di prima specie è

$$\begin{aligned}
&\mathbb{P}\left(\frac{S^2}{\sigma_0^2} < 1-\varepsilon | \mathrm{Var}\left[X_i\right] = \sigma^2 \geq \sigma_0^2\right) \\
&\qquad = \mathbb{P}\left(\frac{(n-1)S^2}{\sigma^2} < \frac{\sigma_0^2}{\sigma^2}(n-1)(1-\varepsilon) | \mathrm{Var}\left[X_i\right] = \sigma^2 \geq \sigma_0^2\right) \\
&\qquad\qquad = F_V\left(\frac{\sigma_0^2}{\sigma^2}(n-1)(1-\varepsilon)\right) \leq F_V\left((n-1)(1-\varepsilon)\right).
\end{aligned}$$

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

$$\text{accetto } H_0 \text{ se } s^2 > \frac{\sigma_0^2}{n-1}\chi^2_{n-1,\alpha} \text{ e la rifiuto altrimenti.}$$

7. Test di ipotesi per il confronto di campioni gaussiani

7.1 Test d'ipotesi per la differenza dei valori attesi di campioni gaussiani

Supponiamo di avere due campioni, entrambi gaussiani e tra di loro indipendenti

$$X\colon X_1,\ \ldots,\ X_n \qquad \mathbb{P}_{X_i} = N\left(\mu_X, \sigma_X^2\right),$$
$$Y\colon Y_1,\ \ldots,\ Y_k \qquad \mathbb{P}_{Y_j} = N\left(\mu_Y, \sigma_Y^2\right).$$

Vogliamo testare

$$H_0: \qquad \mu_X - \mu_Y = d \qquad\qquad H_A: \qquad \mu_X - \mu_Y \neq d.$$

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

Distinguiamo tre diversi casi

7.1.1 Le varianze σ_X^2 e σ_Y^2 sono note

Sappiamo che $\mathbb{P}_{\overline{X}} = N\left(\mu_x, \frac{\sigma_X^2}{n}\right)$, $\mathbb{P}_{\overline{Y}} = N\left(\mu_Y, \frac{\sigma_Y^2}{k}\right)$. Considero la v.a. $W := \overline{X} - \overline{Y}$. Poiché i due campioni sono indipendenti, anche $\overline{X}$ e $\overline{Y}$ sono indipendenti, abbiamo che

$$\mathbb{P}_W = N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{k}\right).$$

Dunque H_0 è vera se e solo se $\mathbb{P}_W = N\left(d, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{k}\right)$. Stabilisco quindi il seguente criterio di accettazione:

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

La probabilità di commettere errore di prima specie vale allora

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

D'altra parte, se H_0 è vera, allora $Z := \dfrac{W-d}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{k}}}$ ha distribuzione gaussiana standard $N(0,1)$, 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

$$\text{accetto l'ipotesi } H_0 \text{ se } |\overline{x} - \overline{y}| < z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{k}} \text{ e la rifiuto altrimenti.}$$

Osservazione 7.1.1. Se $\sigma_X^2 = \sigma_Y^2 = \sigma_0^2$ e $k = n$, allora $\varepsilon = z_{1-\frac{\alpha}{2}} \sigma_0 \sqrt{\frac{2}{n}}$.

7.1.2 Le varianze σ_X^2 e σ_Y^2 sono ignote ma si possono ritenere uguali

Consideriamo le due varianze campionarie

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

Indico con σ^2 il comune valore di σ_X^2 e σ_Y^2. Sappiamo che $V_X := \dfrac{(n-1)S_X^2}{\sigma^2}$ segue la distribuzione χ^2_{n-1}, e che $V_Y := \dfrac{(k-1)S_Y^2}{\sigma^2}$ segue la distribuzione χ^2_{k-1}. Inoltre, poiché i due campioni sono indipendenti, anche V_X e V_Y sono idipendenti. Dunque, per il Teorema 3.3.2, $V_X + V_Y$ segue la distribuzione $\chi^2_{n-1+k-1} = \chi^2_{n+k-2}$

D'altra parte

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

Se definiamo la statistica:

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

abbiamo

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

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

$$Z := \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{k}}}$$

ha distribuzione gaussiana standard $N(0,1)$. Considero

$$T := \frac{\overline{X} - \overline{Y} - d}{\sigma \sqrt{\frac{1}{n} + \frac{1}{k}}} \frac{\sqrt{n+k-2}}{\sqrt{V_X + V_Y}} = \frac{\overline{X} - \overline{Y} - d}{\sqrt{\frac{1}{n} + \frac{1}{k}}} \frac{\sqrt{n+k-2}}{\sqrt{(n-1)S_X^2 + (k-1)S_Y^2}}.$$

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 $\overline{X} - \overline{Y}$ e $V_X + V_Y$ sono indipendenti, e dunque $\mu_X - \mu_Y = d$ se e solo se e $\mathbb{E}[T] = 0$. Infatti, per l'indipendenza, si ha

$$\mathbb{E}[T] = \frac{\mathbb{E}\left[\overline{X} - \overline{Y} - d\right]}{\sigma \sqrt{\frac{1}{n} + \frac{1}{k}}} \sqrt{n+k-2} \mathbb{E}\left[\frac{1}{\sqrt{V_X + V_Y}}\right].$$

Come criterio di accettazione per l'ipotesi nulla H_0 scelgo pertanto $|t| < \varepsilon$.

Inoltre, se H_0 è vera, allora per il Teorema 3.3.8 la v.a. T segue la distribuzione $t(n+k-2)$. La probabilità di commettere errore di prima specie è quindi $\alpha = \mathbb{P}(|T| \geq \varepsilon)$. Fissato il livello di significatività α, devo dunque scegliere $\varepsilon = t_{n+k-2,1-\frac{\alpha}{2}}$.

Siano $x\colon x_1, \ldots, x_n$ e $y\colon y_1, \ldots, y_k$ i dati, $\overline{x}$ e $\overline{y}$ le rispettive medie, s_x^2 e s_y^2 le rispettive varianze:

$$\text{accetto } H_0 \text{ se } \frac{|\overline{x} - \overline{y} - d|}{\sqrt{\frac{1}{n} + \frac{1}{k}}} \frac{\sqrt{n+k-2}}{\sqrt{(n-1)s_X^2 + (k-1)s_Y^2}} < t_{n+k-2,1-\frac{\alpha}{2}}, \text{ e la rifiuto altrimenti.}$$

7.2 Test d'ipotesi per l'uguaglianza delle varianze di campioni gaussiani

Introduciamo prima una nuova distribuzione.

7.2.1 Distribuzione di Fisher-Snedecor a k e n gradi di libertà

Si può dimostrare che la funzione

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

è una densità di probabilità. La distribuzione assolutamente continua ad essa associata si dice *distribuzione di Fisher-Snedecor a k ed n gradi di libertà*, o semplicemente *distribuzione di Fisher a k ed n gradi di libertà*.

Si può dimostrare che se F è una variabile aleatoria con questa distribuzione, allora

$$\mathbb{E}[F] = \begin{cases} \frac{n}{n-2} & n > 2, \\ +\infty & n = 1, 2, \end{cases} \qquad \operatorname{Var}[F] = \begin{cases} \frac{2n^2(k+n-2)}{k(n-2)^2(n-4)} & n > 4, \\ +\infty & n = 3, 4, \\ \text{non esiste} & n = 1, 2. \end{cases}$$

Teorema 7.2.1. *Siano U e V variabili aleatorie indipendenti con distribuzioni $\mathbb{P}_U = \chi_k^2$, $\mathbb{P}_V = \chi_n^2$. Allora la v.a. $F := \dfrac{U/k}{V/n}$ segue la distribuzione di Fisher-Snedecor con k ed n gradi di libertà.*

Dimostrazione. Sappiamo che $\mathbb{P}_U = f(u)\mathrm{d}u$, $\mathbb{P}_V = g(v)\mathrm{d}v$ dove

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

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

Possiamo scrivere $F = \varphi \circ (U, V)$ dove

$$\varphi\colon (u,v) \in \mathbb{R}^2 \mapsto \begin{cases} \dfrac{un}{kv} & v \neq 0, \\ 0 & v = 0. \end{cases}$$

Sia $\psi\colon \mathbb{R} \to \mathbb{R}$ una funzione di Borel non negativa. Abbiamo

$$\int_{\mathbb{R}} \psi(t)\mathrm{d}t = \int_{\mathbb{R}^2} \psi(\varphi(u,v))\mathbb{P}_{U,V}(\mathrm{d}u\mathrm{d}v)$$
$$= \int_{(0,+\infty)^2} \psi\left(\frac{nu}{kv}\right) \frac{1}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{1}{2}\right)^{\frac{k+n}{2}} u^{\frac{k}{2}-1} v^{\frac{n}{2}-1} \exp\left(\frac{-(u+v)}{2}\right) \mathrm{d}u\mathrm{d}v$$

sostituiamo $t = \dfrac{nu}{kv}$, $u = \dfrac{kv}{n}t$, $\mathrm{d}u = \dfrac{kv}{n}\mathrm{d}t$

$$= \int_0^{+\infty} \psi(t) \frac{1}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{1}{2}\right)^{\frac{k+n}{2}} \left(\frac{k}{n}\right)^{\frac{k}{2}} t^{\frac{k}{2}-1} \left(\int_0^{+\infty} y^{\frac{k+n}{2}-1} \exp\left(\frac{-v}{2}\left(1+\frac{kt}{n}\right)\right) \mathrm{d}v \right) \mathrm{d}t$$

sostituiamo $x = \dfrac{v}{2}\left(1+\dfrac{kt}{n}\right) = \dfrac{v}{2}\dfrac{n+kt}{n}$, $v = \dfrac{2nx}{n+kt}$, $\mathrm{d}v = \dfrac{2n}{n+kt}\mathrm{d}x$

$$= \int_0^{+\infty} \psi(t) \frac{1}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{1}{2}\right)^{\frac{k+n}{2}} \left(\frac{k}{n}\right)^{\frac{k}{2}} t^{\frac{k}{2}-1} \left(\int_0^{+\infty} \left(\frac{2n}{n+kt}\right)^{\frac{k+n}{2}} x^{\frac{k+n}{2}-1} \mathrm{e}^{-x} \mathrm{d}x \right) \mathrm{d}t$$
$$= \int_0^{+\infty} \psi(t) \frac{\Gamma\left(\frac{k+n}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{n}{n+kt}\right)^{\frac{k+n}{2}} \left(\frac{k}{n}\right)^{\frac{k}{2}} t^{\frac{k}{2}-1} \mathrm{d}t$$

da cui la tesi. □

Osservazione 7.2.1. Indichiamo con $f_{k,n,\alpha}$ il quantile di livello α associato alla distribuzione di Fisher di parametri k ed n. Siano U e V sono come nel Teorema 7.2.1: U e V variabili aleatorie indipendenti con distribuzioni $\mathbb{P}_U = \chi^2_k$, $\mathbb{P}_V = \chi^2_n$ e sia $\alpha \in (0,1)$. Si ha

$$\alpha = \mathbb{P}\left(\frac{U/k}{V/n} \leq f_{k,n,\alpha}\right) = \mathbb{P}\left(\left(\frac{U/k}{V/n}\right)^{-1} \geq \frac{1}{f_{k,n,\alpha}}\right)$$
$$= \mathbb{P}\left(\frac{V/n}{U/k} \geq \frac{1}{f_{k,n,\alpha}}\right) = 1 - \mathbb{P}\left(\frac{V/n}{U/k} \leq \frac{1}{f_{k,n,\alpha}}\right)$$

ovvero $\mathbb{P}\left(\dfrac{V/n}{U/k} \leq \dfrac{1}{f_{k,n,\alpha}}\right) = 1-\alpha$ cioè $\dfrac{1}{f_{k,n,\alpha}} = f_{n,k,1-\alpha}$.

7.3 Test d'ipotesi per l'uguaglianza delle varianze di campioni gaussiani

Supponiamo di avere due campioni, entrambi gaussiani e tra di loro indipendenti

$$\begin{aligned} X &: X_1, \ \ldots, \ X_k \qquad \mathbb{P}_{X_i} = N\left(\mu_X, \sigma_X^2\right), \\ Y &: Y_1, \ \ldots, \ Y_n \qquad \mathbb{P}_{Y_j} = N\left(\mu_Y, \sigma_Y^2\right). \end{aligned}$$

Vogliamo testare

$$H_0 : \qquad \sigma_X^2 = \sigma_Y^2 \qquad\qquad H_A : \qquad \sigma_2^X \neq \sigma_Y^2.$$

Sappiamo che S_X^2 e S_Y^2 sono stimatori non distorti di σ_X^2 e σ_Y^2, rispettivamente. Dunque:

$$\text{accettiamo } H_0 \text{ se } 1 - \varepsilon_1 < \frac{s_X^2}{s_Y^2} < 1 + \varepsilon_2, \text{ rifiutiamo altrimenti.}$$

Per scegliere ε_1 ed ε_2 in base al livello di significatività desiderato, consideriamo le v.a.

$$V_X = \frac{(k-1)S_X^2}{\sigma_X^2}, \qquad V_Y = \frac{(n-1)S_Y^2}{\sigma_Y^2}.$$

Sappiamo che $\mathbb{P}_{V_X} = \chi^2_{k-1}$, $\mathbb{P}_{V_Y} = \chi^2_{n-1}$. Dunque, la v.a. $\dfrac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$ segue la distribuzione di Fisher con $k-1$ ed $n-1$ gradi di libertà. In particolare H_0 è vera se e solo se $F := \dfrac{S_X^2}{S_Y^2}$ segue la distribuzione di Fisher con $k-1$ ed $n-1$ gradi di libertà.

Abbiamo dunque

$$\alpha = \mathbb{P}\left(\frac{S_X^2}{S_Y^2} \leq 1 - \varepsilon_1 | \sigma_X^2 = \sigma_Y^2\right) + \mathbb{P}\left(\frac{S_X^2}{S_Y^2} \geq 1 + \varepsilon_2 | \sigma_X^2 = \sigma_Y^2\right).$$

Scegliamo di *distribuire equamente l'errore* imponendo

$$\begin{aligned} \frac{\alpha}{2} &= \mathbb{P}\left(\frac{S_X^2}{S_Y^2} \leq 1 - \varepsilon_1 | \sigma_X^2 = \sigma_Y^2\right) = \mathbb{P}\left(F \leq 1 - \varepsilon_1\right) \\ \frac{\alpha}{2} &= \mathbb{P}\left(\frac{S_X^2}{S_Y^2} \geq 1 + \varepsilon_2 | \sigma_X^2 = \sigma_Y^2\right) = \mathbb{P}\left(F \geq 1 + \varepsilon_2\right) = 1 - \mathbb{P}\left(F \leq 1 + \varepsilon_2\right). \end{aligned}$$

Dovrà dunque essere $1 - \varepsilon_1 = f_{k-1,n-1,\frac{\alpha}{2}}$, $1 + \varepsilon_2 = f_{k-1,n-1,1-\frac{\alpha}{2}}$. In definitiva:

$$\text{accetto } H_0 \text{ se } f_{k-1,n-1,\frac{\alpha}{2}} < \frac{s_X^2}{s_Y^2} < f_{k-1,n-1,1-\frac{\alpha}{2}}. \text{ Rifiuto altrimenti.}$$

8. Test del χ^2 e test di Smirnov-Kolmogorov

8.1 Stimatori di massima verosimiglianza per distribuzioni con densità finita

Supponiamo di avere un campione statistico $X_1,\ \ldots,\ X_n$ e di sapere che esso è relativo ad una distribuzione su un insieme finito $t_1,\ \ldots,\ t_k$. Dunque conosco la distribuzione se conosco $p_j := \mathbb{P}\left(X_i = t_j\right)$ per ogni $j = 1, \ldots, k$.

Dato il campione sperimentale $x_1,\ \ldots,\ x_n$, cerchiamo gli stimatori di massima verosimiglianza per i parametri $p_1,\ \ldots,\ p_k$. Tra i dati rilevati $x_1,\ \ldots,\ x_n$ ce ne sono:
n_1 che valgono t_1,
n_2 che valgono t_2,
$\ldots$,
n_k che valgono t_k,
con la condizione $n_1 + n_2 + \ldots + n_k = n$.

La densità congiunta di $(X_1,\ \ldots,\ X_n)$ in $x_1,\ \ldots,\ x_n$ è dunque

$$f(x_1,\ \ldots,\ x_n | p_1,\ \ldots,\ p_k) = {p_1}^{n_1} {p_2}^{n_2} \ldots {p_k}^{n_k} = \prod_{j=1}^{k} {p_j}^{n_j}$$

e perciò

$$g(x_1,\ \ldots,\ x_n | p_1,\ \ldots,\ p_k) := \log f(x_1,\ \ldots,\ x_n | p_1,\ \ldots,\ p_k) = \sum_{j=1}^{k} n_j \log p_j.$$

Usiamo i moltiplicatori di Lagrange per massimizzare g rispetto ai $p_1,\ \ldots,\ p_k$ ammissibili:

$$G(p_1,\ \ldots,\ p_k, \lambda) = \sum_{j=1}^{k} n_j \log p_j - \lambda \left(\sum_{j=1}^{k} p_j - 1 \right).$$

$$\frac{\partial G}{\partial \lambda} = -\left(\sum_{j=1}^{k} p_j - 1 \right), \qquad \frac{\partial G}{\partial p_j} = \frac{n_i}{p_j} - \lambda \qquad \forall j = 1, \ldots, k.$$

Da cui otteniamo

$$p_j = \frac{n_j}{n} \quad \forall j = 1, \ldots, k,$$

ovvero lo stimatore di massima verosimiglianza per la densità in t_j è la frequenza relativa del carattere t_j nel campione $x_1,\ \ldots,\ x_n$.

8.2 Test del χ^2

Sia $Y_1, \ldots, Y_n$ un campione statistico. Supponiamo che le v.a. del campione siano discrete a valori $t_1, \ldots, t_k$. Consideriamo le densità di probabilità

$$p_j := \mathbb{P}(Y_i = t_j), \qquad j = 1, \ldots, k.$$

Siano $p_1^0, \ldots, p_k^0$ dei numeri assegnati, tali che $p_j^0 \geq 0 \quad \forall j = 1, \ldots k$ e $\sum_{j=1}^k p_j^0 = 1$.
Vogliamo testare

$$H_0: \quad p_j = p_j^0 \qquad \forall j = 1, \ldots, k \qquad H_A: \exists \bar{j} \in \{1, \ldots, k\}: p_{\bar{j}} \neq p_{\bar{j}}^0.$$

Per ogni $j = 1, \ldots, k$ considero

$$X_j = \#\{i \in \{1, \ldots, n\}: Y_i = t_j\} \qquad j = 1, \ldots, k.$$

Sicuramente $\mathbb{P}_{X_j} = B(n, p_j)$, quindi $\mathbb{E}[X_j] = np_j$, $\mathrm{Var}[X_j] = np_j(1-p_j)$. Inoltre $(X_j - np_j)^2$ mi dice quanto sia verosimile che $\mathbb{P}(Y_i = t_j) = p_j$. Posso stabilire un criterio di accettazione considerando una opportuna combinazione lineare $\sum_{j=1}^k a_j (X_j - np_j)^2$ con coefficienti $a_1, \ldots, a_k$ positivi. Si può dimostrare che vale il seguente

Teorema 8.2.1 (di Pearson)**.** *Se $\mathbb{P}_{X_j} = Bin(n, p_j)$, allora la legge della v.a. $\sum_{j=1}^k \frac{(X_j - np_j)^2}{np_j}$ converge, per $n \to \infty$, alla legge associata alla distribuzione χ^2_{k-1}.*

Osservazione 8.2.1. L'approssimazione è considerata accettabile se $np_j \geq 5 \quad \forall j = 1, \ldots, k$.

Formuliamo allora il seguente criterio di accettazione. Siano $n_1, \ldots, n_k$ le frequenze assolute dei caratteri $t_1, \ldots, t_k$ nel campione empirico $x_1, \ldots, x_n$

$$\text{accetto } H_0 \text{ se } t_n := \sum_{j=1}^k \frac{(n_j - np_j^0)^2}{np_j^0} < \varepsilon. \text{ Rifiuto altrimenti}$$

La probabilità di commettere errore di prima specie è allora

$$\alpha := \mathbb{P}\left(\sum_{j=1}^k \frac{(X_j - np_j^0)^2}{np_j^0} \geq \varepsilon | p_j = p_j^0 \quad \forall j = 1, \ldots, k\right) \simeq 1 - F_{\chi^2_{k-1}}(\varepsilon).$$

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

Osservazione 8.2.2. Non dimostriamo il Teorema 8.2.1 ma ne vediamo la sua *plausibilità* nel caso $k = 2$. Si ha

$$T = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2}, \qquad p_1 + p_2 = 1, \qquad X_1 + X_2 = n,$$

da cui

$$T = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{n(1-p_1)} = \frac{(X_1 - np_1)^2}{np_1(1-p_1)} = \left(\frac{X_1 - \mathbb{E}[X_1]}{\sqrt{\mathrm{Var}[X_1]}}\right)^2.$$

Possiamo scrivere $X_1 = \sum_{r=1}^{n_1} Z_r$ dove $Z_1, \ldots, Z_{n_1}$ sono i.i.d. con $\mathbb{P}_{Z_r} = B(p_1)$. Dunque

$$T = \left(\frac{\sum_{r=1}^{n_1} Z_r - n_1 \mathbb{E}[Z_1]}{\sqrt{n_1 \mathrm{Var}[Z_1]}} \right)^2.$$

Per il teorema del limite centrale la legge di $\dfrac{\sum_{r=1}^{n_1} Z_r - n\mathbb{E}[Z_1]}{\sqrt{n_1 \mathrm{Var}[Z_1]}}$ converge alla legge gaussiana standard e sappiamo che il quadrato di una v.a. con distribuzione $N(0,1)$ segue la distribuzione χ^2 ad un grado di libertà.

8.3 Test di Kolmogorov-Smirnov

Sia $\{X_i\}_{i=1}^{\infty}$ una successione di v.a. i.i.d. con legge F_0. Pongo

$$Y_i(\omega, t) = \mathbb{1}_{(-\infty,t]}(X_i(\omega)) = \begin{cases} 1 & X_i(\omega) \le t, \\ 1 & X_i(\omega) > t. \end{cases}$$

Si ha $\mathbb{E}[Y_i(\cdot,t)] = \mathbb{P}(X_i \le t) = F_0(t)$, $\mathrm{Var}[Y_i(\cdot,t)] = F_0(t)(1-F_0(t)) \le 1$.

Per ogni $n \in \mathbb{N}$ sia $g_n \colon (x_1, \ldots, x_n, t) \in \mathbb{R}^n \times \mathbb{R} \mapsto \dfrac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(x_i) \in \mathbb{R}$.

Considero la v.a.

$$G_n(\omega, t) = g_n \circ (X_1(\omega), \ldots, X_n(\omega), t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(X_i(\omega)) = \frac{1}{n} \sum_{i=1}^{n} Y_i(\omega, t).$$

Per la disuguaglianza di Chebychev, Teorema 3.2.1,

$$\mathbb{P}(|G_n(\cdot,t) - F_0(t)| > \varepsilon) \le \frac{1}{n\varepsilon^2} \qquad \forall \varepsilon > 0, \quad \forall t \in \mathbb{R}.$$

Dunque

$$\lim_{n\to\infty} \mathbb{P}(|G_n(\cdot,t) - F_0(t)| > \varepsilon) = 0, \text{ uniformemente per } t \in \mathbb{R}.$$

Osserviamo che $G_n(\omega, t) = \dfrac{1}{n} \#\{i \in \{1, \ldots, n\} \colon X_i(\omega) \le t\}$ dunque $G_n(\omega, \cdot)$ è una funzione costante a tratti, monotona crescente che prende valori in $0, \dfrac{1}{n}, \ldots, \dfrac{n-1}{n}, 1$ (li prende tutti se e solo se i valori $X_1(\omega), \ldots, X_n(\omega)$ sono tutti distinti).

Consideriamo allora il seguente test d'ipotesi per un campione statistico $X_1, \ldots, X_n$ di cui rilevo i dati $x_1, \ldots, x_n$. Sia $F_0 \colon \mathbb{R} \to [0,1]$ una funzione monotona crescente, tale che $\lim_{t\to-\infty} F_0(t) = 0$, $\lim_{t\to+\infty} F_0(t) = 1$. Supponiamo inoltre che F_0 sia continua. Voglio testare

$$H_0 : \quad F_0 \text{ è la legge del campione}, \qquad H_A : \quad \exists t \in \mathbb{R} \colon F_0(t) \ne \mathbb{P}(X_i \le t).$$

Sia $d_n := \sup_{t\in\mathbb{R}} |g_n(x_1, \ldots, x_n, t)|$. Accetto H_0 se $d_n < \varepsilon$, rifiuto altrimenti. Vediamo se possiamo scegliere ε in base al livello di significatività desiderato.

Lemma 8.3.1. *Se X è una v.a. con legge F, allora $F(X)$ è uniformemente distribuita sull'intervallo $[0,1]$.*

Dimostrazione. Dimostriamo il lemma limitatamente al caso assolutamente continuo. Sia f la densità della distribuzione di X: $\mathbb{P}_X = f(x)dx$ e sia $\psi\colon \mathbb{R} \to \mathbb{R}$ una funzione di Borel non negativa. Si ha

$$\int_{\mathbb{R}} \psi(t)\mathbb{P}_{F(X)}dt = \int_{\mathbb{R}} \psi(F(x))\mathbb{P}_X(dx) = \int_{\mathbb{R}} \psi(F(x))f(x)dx = \int_0^1 \psi(t)dt$$

dove abbiamo effettuato il cambio di variabile $t = F(x)$. □

Teorema 8.3.2. *Sia $X_1, \ldots, X_n$ campione statistico con legge continua F. Sia G_n come prima: $G_n(\omega, t) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(X_i(\omega))$ e sia*

$$D_n(\omega) := \sup_{t\in\mathbb{R}} |G_n(\omega, t) - F(t)| .$$

Allora la legge di D_n non dipende da F.

Dimostrazione. Sia $d \geq 0$

$$\begin{aligned}\mathbb{P}\left(D_n \geq d\right) &= \mathbb{P}\left(\sup_{t\in\mathbb{R}} \left|\frac{1}{n}\#\{i\colon X_i \leq t\} - F(t)\right| \geq d\right) = \\ &= \mathbb{P}\left(\sup_{t\in\mathbb{R}} \left|\frac{1}{n}\#\{i\colon F(X_i) \leq F(t)\} - F(t)\right| \geq d\right).\end{aligned}$$

Infatti, se F è strettamente crescente, allora $X_i \leq t$ se e solo se $F(x_i) \leq F(t)$. Se invece F è crescente, ma non strettamente, l'uguaglianza rimane vera a livello di probabilità perché la probabilità che X_i cada in un intervallo in cui F è costante è comunque nulla.

D'altra parte le v.a. $U_i := F(X_i)$ sono i.i.d con distribuzione uniforme sull'intervallo $[0, 1]$, dunque

$$\begin{aligned}\mathbb{P}\left(D_n \geq d\right) &= \mathbb{P}\left(\sup_{t\in\mathbb{R}} \left|\frac{1}{n}\#\{i\colon U_i \leq F(t)\} - F(t)\right| \geq d\right) = \\ &= \mathbb{P}\left(\sup_{y\in(0,1)} \left|\frac{1}{n}\#\{i\colon U_i \leq y\} - y\right| \geq d\right)\end{aligned}$$

dato che, essendo continua, F assume tutti i valori compresi tra il suo estremo inferiore ed il suo estremo superiore. □

Si può dimostrare che vale il seguente limite

$$\lim_{n\to\infty} \mathbb{P}\left(D_n\sqrt{n} \leq t\right) = \begin{cases} 1 - 2\displaystyle\sum_{j=1}^{\infty}(-1)^{j-1}\exp\left(-2j^2t^2\right) & t > 0, \\ 0 & t \leq 0. \end{cases}$$

Riconsideriamo dunque la probabilità di commettere errore di prima specie.

$$\alpha = \mathbb{P}\left(D_n \geq \varepsilon\right) = \mathbb{P}\left(D_n\sqrt{n} \geq \varepsilon\sqrt{n}\right) \sim 2\sum_{j=1}^{\infty}(-1)^{j-1}\exp\left(-2j^2\varepsilon^2 n\right) \geq 2\exp\left(-2\varepsilon^2 n\right).$$

Scegliamo dunque $\varepsilon > 0$ tale che $\alpha = 2\exp\left(-2\varepsilon^2 n\right)$ cioè $\varepsilon = \sqrt{\dfrac{1}{2n}\log\dfrac{2}{\alpha}}$. Quindi

accetto H_0 se $\sup_{t\in\mathbb{R}} \left| \frac{1}{n} \#\{i\colon x_i \le t\} - F(t) \right| < \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}$. Rifiuto altrimenti.

Osservazione 8.3.1. Supponiamo di aver ordinato i dati $x_1, \ldots, x_n$ in ordine crescente (per semplicità supponiamo che siano tutti distinti). Abbiamo

$$\begin{aligned}
\sup_{t\in\mathbb{R}} \left| \frac{1}{n} \#\{i\colon x_i \le t\} - F(t) \right| &= \max \Big\{ \sup_{t<x_1} \left| \frac{1}{n} \#\{i\colon x_i \le t\} - F(t) \right|, \\
&\qquad \sup_{t\in[x_1,x_2)} \left| \frac{1}{n} \#\{i\colon x_i \le t\} - F(t) \right|, \ldots, \sup_{t\in[x_{n-1},x_n)} \left| \frac{1}{n} \#\{i\colon x_i \le t\} - F(t) \right|, \\
&\qquad\qquad \sup_{t\ge x_n} \left| \frac{1}{n} \#\{i\colon x_i \le t\} - F(t) \right| \Big\} \\
&= \max \Big\{ \sup_{t<x_1} |F(t)|, \sup_{t\in[x_1,x_2)} \left| \frac{1}{n} - F(t) \right|, \ldots, \sup_{t\in[x_{n-1},x_n)} \left| \frac{n-1}{n} - F(t) \right|, \sup_{t\ge x_n} |1 - F(t)| \Big\} \\
&= \max \Big\{ F(x_1), \left| \frac{1}{n} - F(x_1) \right|, \left| \frac{1}{n} - F(x_2) \right| \ldots, \\
&\qquad\qquad \left| \frac{n-1}{n} - F(x_{n-1}) \right|, \left| \frac{n-1}{n} - F(x_n) \right|, |1 - F(x_n)| \Big\}.
\end{aligned}$$

9. Regressione lineare

Supponiamo di fare un esperimento in cui si può controllare direttamente una variabile di input x. La risposta dell'esperimento dipende da x ma in generale risulta affetta da errore e comunque non deterministica. Se ci sembra che ci sia una relazione di un qualche tipo, per esempio lineare, tra il dato di unput e la risposta dell'esperimento, anche questa relazione sarà affetta da errore: in generale non riusciamo ad osservare $y = ax + b$ ma $y = ax + b + \varepsilon$, dove ε è l'errore.

Per ogni dato di input x_i in $x_1, \ldots, x_n$ vediamo dunque la risposta dell'esperimento come una v.a. Y_i con $\mathbb{E}[Y_i] = ax_i + b$ e i parametri della retta che rappresenta la risposta dell'esperimento in funzione di x come una retta i cui parametri sono v.a.: $y = Ax + B$. La quantità $(Y_i - (Ax_i + B))^2$ è il quadrato della differenza tra l'osservazione ed il valore predetto. La retta, ovvero i parametri A e B che la definiscono, si scelgono minimizzando la somma dei quadrati degli errori, cioè

$$S(A, B) = \sum_{i=1}^{n} (Y_i - (Ax_i + B))^2 \to \min$$

Abbiamo già affrontato questo problema nel caso descrittivo, Sezione 2.2. Si ha dunque

$$A = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{S_{xx}}, \qquad B = \overline{Y} - A\overline{x}, \quad \text{dove} \quad S_{xx} := \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

La retta $Y = Ax + B$ è detta *stima della regressione*. Possiamo scrivere A e B in un'altra forma, più utile a comprenderne la natura.

$$\begin{aligned} A &= \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{S_{xx}} = \frac{\sum_{i=1}^n x_i Y_i - n\overline{x}\,\overline{Y}}{S_{xx}} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) Y_i, \\ B &= \overline{Y} - A\overline{x} = \frac{1}{n} \sum_{i=1}^{n} Y_i - \frac{\overline{x}}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) Y_i = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\overline{x}(x_i - \overline{x})}{S_{xx}} \right) Y_i, \end{aligned} \tag{9.1}$$

Poiché $\mathbb{E}[Y_i] = ax_i + b$, supporrò che le v.a. Y_i siano v.a. indipendenti, gaussiane, ed aventi tutte la stessa varianza σ^2:

$$\mathbb{P}_{Y_i} = N\left(ax_i + b, \sigma^2\right), \quad Y_1, \ldots, Y_n \text{ indipendenti.}$$

Grazie alle equazioni (9.1) abbiamo allora che anche A e B sono gaussiane, in quanto combi-

nazioni lineari di v.a. gaussiane indipendenti. Andiamo a calcolarne valore atteso e varianza.

$$\begin{aligned}
\mathbb{E}[A] &= \mathbb{E}\left[\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})Y_i\right] = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})\mathbb{E}[Y_i] = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})(ax_i+b)\\
&= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})(a(x_i-\overline{x})+a\overline{x}+b) = \frac{1}{S_{xx}}\sum_{i=1}^{n}a(x_i-\overline{x})^2 = a,\\
\mathrm{Var}[A] &= \mathrm{Var}\left[\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})Y_i\right] = \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\mathrm{Var}[Y_i]\\
&= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\sigma^2 = \frac{1}{S_{xx}^2}S_{xx}\sigma^2 = \frac{\sigma^2}{S_{xx}},\\
\mathbb{E}[B] &= \mathbb{E}\left[\overline{Y}-A\overline{x}\right] = \mathbb{E}\left[\overline{Y}\right]-\overline{x}\mathbb{E}[A] = \frac{1}{n}\sum_{i=1}^{n}(ax_i+b)-a\overline{x} = b,\\
\mathrm{Var}[B] &= \mathrm{Var}\left[\sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\overline{x}(x_i-\overline{x})}{S_{xx}}\right)Y_i\right] = \sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\overline{x}(x_i-\overline{x})}{S_{xx}}\right)^2\mathrm{Var}[Y_i]\\
&= \sigma^2\sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\overline{x}(x_i-\overline{x})}{S_{xx}}\right)^2 = \sigma^2\sum_{i=1}^{n}\left(\frac{1}{n^2}+\frac{\overline{x}^2(x_i-\overline{x})^2}{S_{xx}^2}-\frac{2}{n}\frac{\overline{x}(x_i-\overline{x})}{S_{xx}}\right)\\
&= \sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right) = \frac{\sigma^2\sum_{i=1}^{n}x_i^2}{nS_{xx}}.
\end{aligned}$$

Considero la differenza tra la risposta Y_i e la predizione Ax_i+B: $R_i := |Y_i-(Ax_i+B)|$ è detta *residuo*, dunque la quantità che abbiamo ottenuto minimizzando S è la somma dei quadrati dei residui:

$$S_R- = \sum_{i=1}^{n}R_i^2 = \sum_{i=1}^{n}\left(Y_i-(Ax_i+B)\right)^2.$$

Si può dimostrare che la v.a. $\dfrac{S_R}{\sigma^2}$ ha distribuzione χ^2_{n-2} e che A, B e S_R sono indipendenti. Inoltre $\mathbb{E}\left[\dfrac{S_R}{n-2}\right] = \mathbb{E}\left[\dfrac{\sigma^2}{n-2}\dfrac{S_R}{\sigma^2}\right] = \dfrac{\sigma^2}{n-2}\mathbb{E}\left[\dfrac{S_R}{\sigma^2}\right] = \dfrac{\sigma^2}{n-2}(n-2) = \sigma^2$. Riassumendo abbiamo:

Teorema 9.0.1. *Se le v.a. $Y_1, \ldots, Y_n$ sono gaussiane indipendendenti con*

$$\mathbb{P}_{Y_i} = N\left(ax_i+b,\sigma^2\right) \qquad \forall i=1,\ldots,n.$$

Allora le v.a. A, B, S_R sono indipendenti. Hanno distribuzione

$$\mathbb{P}_A = N\left(a,\frac{\sigma^2}{S_{xx}}\right), \qquad \mathbb{P}_B = N\left(b,\frac{\sigma^2\sum_{i=1}^{n}x_i^2}{nS_{xx}}\right), \qquad \mathbb{P}_{\frac{S_R}{\sigma^2}} = \chi^2_{n-2}.$$

Inoltre A, B e $\dfrac{S_R}{n-2}$ sono rispettivamente stimatori non distorti di a, b e σ^2.

Introduciamo una notazione più sintetica:

$$S_{YY} = \sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2 = \sum_{i=1}^{n}Y_i^2-n\overline{Y}^2, \qquad S_{xY} = \sum_{i=1}^{n}(x_i-\overline{x})\left(Y_i-\overline{Y}\right) = \sum_{i=1}^{n}x_iY_i-n\overline{x}\overline{Y}.$$

Abbiamo allora $A = \frac{S_{xY}}{S_{xx}}$, $B = \overline{Y} - A\overline{x}$,

$$\begin{aligned} S_R &= \sum_{i=1}^{n} \left(Y_i - \frac{S_{xY}}{S_{xx}} x_i - \overline{Y} + \frac{S_{xY}}{S_{xx}} \overline{x} \right)^2 = \sum_{i=1}^{n} \left((Y_i - \overline{Y}) - \frac{S_{xY}}{S_{xx}} (x_i - \overline{x}) \right)^2 \\ &= \sum_{i=1}^{n} (Y_i - \overline{Y})^2 + \sum_{i=1}^{n} \frac{S_{xY}^2}{S_{xx}^2} (x_i - \overline{x})^2 - 2 \frac{S_{xY}}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y}) \\ &= S_{YY} - \frac{S_{xY}^2}{S_{xx}} = \frac{S_{xx} S_{YY} - S_{xY}^2}{S_{xx}}. \end{aligned}$$

Possiamo fare inferenza statistica sui parametri a e b della retta di regressione? Cerchiamo un intervallo di confidenza di livello $1 - \alpha$ per il parametro a. Per il Teorema 9.0.1 la v.a. $Z := \frac{A - a}{\frac{\sigma}{S_{xx}}}$ ha distribuzione gaussiana standard, mentre $V_R := \frac{S_R}{\sigma^2}$ ha distribuzione χ^2_{n-2} ed è indipendente da Z Dunque $T := \frac{Z\sqrt{n-2}}{V_R} = \frac{(A-a)S_{xx}\sqrt{n-2}}{\sqrt{S_R}}$ ha distribuzione t di Student con $n-2$ gradi di libertà: $\mathbb{P}_T = t(n-2)$. Abbiamo dunque

$$\begin{aligned} 1 - \alpha &= \mathbb{P}\left(\frac{|A - a| S_{xx}\sqrt{n-2}}{\sqrt{S_R}} < t_{n-2,1-\frac{\alpha}{2}} \right) \\ &= \mathbb{P}\left(A - \frac{\sqrt{S_R}}{S_{xx}\sqrt{n-2}} t_{n-2,1-\frac{\alpha}{2}} < a < A + \frac{\sqrt{S_R}}{S_{xx}\sqrt{n-2}} t_{n-2,1-\frac{\alpha}{2}} \right) \end{aligned}$$

Possiamo anche impostare un test d'ipotesi per il parametro a. Vogliamo testare

$$H_0 : \quad a = \overline{a}, \qquad\qquad H_A : \quad a \neq \overline{a}.$$

Poiché $\frac{(A - \overline{a})S_{xx}\sqrt{n-2}}{\sqrt{S_R}}$ ha valore atteso nullo se e solo se $a = \overline{a}$, accetto H_0 se $\frac{|a(x_1, \ldots, x_n, y_1, \ldots, y_n) - \overline{a}| S_{xx}\sqrt{n-2}}{\sqrt{s_R(x_1, \ldots, x_n, y_1, \ldots, y_n)}} < \varepsilon$, la rifiuto altrimenti. La probabilità di commettere errore di prima specie è

$$PP\frac{|A - \overline{a}| S_{xx}\sqrt{n-2}}{\sqrt{S_R}} \geq \varepsilon | a = \overline{a} = \mathbb{P}\left(|T_{n-2}| \geq \varepsilon \right).$$

Per ottenere livello di significatività pari ad α, dobbiamo dunque prendere $\varepsilon = t_{n-2,1-\frac{\alpha}{2}}$. Infine:

accetto H_0 se $\frac{|a(x_1, \ldots, x_n, y_1, \ldots, y_n) - \overline{a}| S_{xx}\sqrt{n-2}}{\sqrt{s_R(x_1, \ldots, x_n, y_1, \ldots, y_n)}} < t_{n-2,1-\frac{\alpha}{2}}$, la rifiuto altrimenti.

Risultati analoghi si ottengono per il parametro b. La variabile aleatoria $Z_B := \frac{B - b}{\sqrt{\frac{\sigma^2 \sum_{i=1}^n x_i^2}{n S_{xx}}}}$ ha distribuzione gaussiana standard, la v.a. $\frac{S_R}{\sigma^2}$ ha distribuzione χ^2_{n-2} ed è indipendente da Z_B, dunque $T_B := \frac{Z_B\sqrt{n-2}}{\frac{S_R}{\sigma^2}} = \frac{(B-b)\sqrt{n(n-2)S_{xx}}}{\sqrt{S_R \sum_{i=1}^n x_i^2}}$ ha distribuzione $t(n-2)$. Dunque

abbiamo l'intervallo di confidenza di livello $1-\alpha$

$$\left(B - t_{n-2,1-\frac{\alpha}{2}}\sqrt{\frac{S_R\sum_{i=1}^n x_i^2}{n(n-2)S_{xx}}}, B + t_{n-2,1-\frac{\alpha}{2}}\sqrt{\frac{S_R\sum_{i=1}^n x_i^2}{n(n-2)S_{xx}}}\right).$$

Abbiamo anche un test d'ipotesi. Vogliamo testare

$$H_0: \quad b = \overline{b}, \qquad\qquad H_A: \quad b \neq \overline{b}.$$

Poiché $\dfrac{(B-\overline{b})\sqrt{n(n-2)S_{xx}}}{\sqrt{S_R\sum_{i=1}^n x_i^2}}$ ha valore atteso nullo se e solo se $b = \overline{b}$, accettiamo H_0 se $\dfrac{|b(x_1,\ \ldots,\ x_n, y_1,\ \ldots,\ y_n) - \overline{b}|\sqrt{n(n-2)S_{xx}}}{\sqrt{s_R(x_1,\ \ldots,\ x_n, y_1,\ \ldots,\ y_n)\sum_{i=1}^n x_i^2}} < \varepsilon$, rifiutiamo altrimenti. Come per il parametro a, anche qui otteniamo un test di ipotesi con livello di significatività α, scegliendo $\varepsilon = t_{n-2,1-\frac{\alpha}{2}}$.

Accetto H_0 se $\dfrac{|b(x_1,\ \ldots,\ x_n, y_1,\ \ldots,\ y_n) - \overline{b}|\sqrt{n(n-2)S_{xx}}}{\sqrt{s_R(x_1,\ \ldots,\ x_n, y_1,\ \ldots,\ y_n)\sum_{i=1}^n x_i^2}} < t_{n-2,1-\frac{\alpha}{2}}$, rifiuto altrimenti.

9.1 Inferenza sul risultato di un successivo esperimento

Sulla base dei dati $x_1,\ \ldots,\ x_n, y_1,\ \ldots,\ y_n$ supponiamo di aver ottenuto la retta di regressione $y = ax + b$. Se impostiamo il dato di input $x = x_0$, cosa dobbiamo aspettarci come risposta dell'esperimento? Il valore atteso si calcola facilmente:

$$\mathbb{E}[Ax_0 + B] = x_0\mathbb{E}[A] + \mathbb{E}[B] = ax_0 + b.$$

Posso calcolare un intervallo di confidenza o impostare un test d'ipotesi su questa aspettativa? Possiamo scrivere

$$\begin{aligned} Ax_0 + B &= Ax_0 + \overline{Y} - A\overline{x} = A(x_0 - \overline{x}) + \overline{Y} = (x_0 - \overline{x})\sum_{i=1}^n \frac{x_i - \overline{x}}{S_{xx}}Y_i + \frac{1}{n}\sum_{i=1}^n Y_i \\ &= \sum_{i=1}^n \left(\frac{(x_0 - \overline{x})(x_i - \overline{x})}{S_{xx}} + \frac{1}{n}\right)Y_i. \end{aligned}$$

Dunque anche $Ax_0 + B$ è combinazione lineare delle v.a. gaussiane e indipendenti e perció è anch'essa una v.a. gaussiana. Ne abbiamo già calcolato il valore atteso. Per caratterizzarne completamente la distribuzione è dunque sufficiente calcolarne la varianza.

$$\begin{aligned} \mathrm{Var}[Ax_0 + B] &= \mathrm{Var}\left[\sum_{i=1}^n \left(\frac{(x_0 - \overline{x})(x_i - \overline{x})}{S_{xx}} + \frac{1}{n}\right)Y_i\right] \\ &= \sum_{i=1}^n \left(\frac{(x_0 - \overline{x})(x_i - \overline{x})}{S_{xx}} + \frac{1}{n}\right)^2 \mathrm{Var}[Y_i] = \sigma^2\sum_{i=1}^n \left(\frac{(x_0 - \overline{x})(x_i - \overline{x})}{S_{xx}} + \frac{1}{n}\right)^2 \\ &= \sigma^2\sum_{i=1}^n \left(\frac{(x_0 - \overline{x})^2(x_i - \overline{x})^2}{S_{xx}^2} + \frac{1}{n^2} + 2\frac{(x_0 - \overline{x})(x_i - \overline{x})}{nS_{xx}}\right) \\ &= \sigma^2\left(\frac{(x_0 - \overline{x})^2}{S_{xx}} + \frac{1}{n}\right). \end{aligned}$$

Abbiamo dunque che $Ax_0 + B$ è indipendente da S_R e

$$\mathbb{P}_{Ax_0+B} = N\left(ax_0 + b, \sigma^2\left(\frac{(x_0 - \overline{x})^2}{S_{xx}} + \frac{1}{n}\right)\right).$$

Dunque la v.a. $Z_0 := \dfrac{Ax_0 + B - ax_0 - b}{\sigma\sqrt{\dfrac{(x_0 - \overline{x})^2}{S_{xx}} + \dfrac{1}{n}}}$ ha distribuzione gaussiana standard ed è indipendente da $\dfrac{S_R}{\sigma^2}$ che ha distribuzione χ^2_{n-2}. Di conseguenza la v.a.

$$T_0 := \frac{Z_0\sqrt{n-2}}{\sqrt{\frac{S_R}{\sigma^2}}} = \frac{Ax_0 + B - ax_0 - b}{\sqrt{\dfrac{(x_0 - \overline{x})^2}{S_{xx}} + \dfrac{1}{n}}}\frac{\sqrt{n-2}}{\sqrt{S_R}}$$

ha distribuzione $t(n-2)$. Abbiamo dunque l'intervallo di confidenza di livello $1-\alpha$ per il parametro $ax_0 + b$

$$\left(Ax_0 + B - t_{n-2,1-\frac{\alpha}{2}}\sqrt{\frac{S_R\left(\dfrac{(x_0 - \overline{x})^2}{S_{xx}} + \dfrac{1}{n}\right)}{n-2}}, Ax_0 + B + t_{n-2,1-\frac{\alpha}{2}}\sqrt{\frac{S_R\left(\dfrac{(x_0 - \overline{x})^2}{S_{xx}} + \dfrac{1}{n}\right)}{n-2}}\right).$$

Esercizio 9.1.1. Ricavare il test d'ipotesi.

Bibliografia

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

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

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