Università degli Studi
di Salerno

Analisi della popolazione in base al titolo di studio in Italia

Progetto per il corso di
Metodi e Tecniche per
l'Analisi dei Dati

Docente:
Prof. Amelia G. Nobile

Studente:
Jacek Filipczuk
mat. 0522500211

Anno Accademico 2013 - 2014

Indice

Capitolo 1

Introduzione

L'obiettivo del seguente caso di studio e' l'analisi statistica del numero di persone che hanno conseguito un certo titolo di studio in Italia suddivise per regione. I dati usati per svolgere il lavoro sono reperibili nell'archivio del sito dell'ISTAT all'indirizzo `http://dati.istat.it/Index.aspx`, nella sezione Istruzione e Formazione.La tabella dei dati usata e' stata ridimensionata per facilitare l'attivita' di analisi.

1.1 Software Usati

Di seguito verranno elencati i software usati per svolgere il caso di studio proposto, con lo scopo di facilitarne una futura riproduzione.

- RStudio 0.98.501
- pdfTeX 3.1415926-2.4-1.40.13
- R version 3.0.3 (2014-03-06), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=Italian_Italy.1252`, `LC_CTYPE=Italian_Italy.1252`, `LC_MONETARY=Italian_Italy.1252`, `LC_NUMERIC=C`, `LC_TIME=Italian_Italy.1252`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: lattice 0.20-27, xtable 1.7-3
- Loaded via a namespace (and not attached): grid 3.0.3, tools 3.0.3

1.2 Inizializzazione

I dati scaricati dal sito dell'ISTAT sono presenti in formato CSV, di conseguenza la prima operazione da eseguire e' quella di lettura dell'input e memorizzazione in un'apposita struttura dati. Il software usato mette a disposizione il comando `read` per effettuare la lettura del file di input:

```
> dati <- read.csv("Dati.csv", row.names=1)
```

Il parametro row.names=1 sta a indicare che il primo elemento di ogni riga, durante la lettura della tabella, rappresenta il nome della riga stessa. Una volta eseguita la lettura, la tabella risulta essere la seguente:

	licenza elementare	licenza media	diploma 3 anni	diploma 5 anni	laurea
Piemonte	782.56	1252.76	309.49	1064.12	453.35
Valle d'Aosta	21.80	38.76	6.61	29.83	13.18
Liguria	273.60	435.05	84.51	414.92	205.67
Lombardia	1609.08	2687.06	738.86	2415.40	1122.76
Trentino Alto Adige	149.52	281.13	129.45	215.33	102.78
Prov Autonoma Bolzano	73.95	152.42	62.21	91.75	46.59
Prov Autonoma Trento	75.57	128.71	67.24	123.58	56.19
Veneto	850.60	1324.43	442.31	1125.62	486.94
Friuli-Venezia Giulia	192.15	336.79	97.22	308.75	135.44
Emilia-Romagna	781.03	1128.14	269.21	1151.88	518.29
Toscana	785.82	980.43	132.83	958.02	415.56
Umbria	163.95	211.44	46.17	250.65	115.09
Marche	294.80	405.90	74.35	400.66	178.39
Lazio	804.65	1371.11	236.91	1770.26	790.13
Abruzzo	253.05	316.70	55.95	393.23	148.32
Molise	67.12	83.90	8.72	84.88	33.51
Campania	1098.84	1669.99	155.94	1435.93	526.25
Puglia	892.92	1199.97	100.20	948.50	340.83
Basilicata	126.96	148.35	20.78	153.78	55.90
Calabria	433.41	523.42	52.77	536.52	174.98
Sicilia	1041.86	1512.73	91.04	1203.72	422.84
Sardegna	322.43	571.92	36.03	378.29	155.17

Tabella 1.1: Popolazione per titolo di studio

Come si puo' notare, la tabella contiene una suddivisione della popolazione in base alle regioni italiane. Ogni riga contiene il numero di persone che hanno conseguito un titolo di studio in una specifica regione di appartenenza, ogni colonna, invece, indica lo specifico titolo di studio conseguito.

Capitolo 2

Analisi Descrittiva

L'analisi descrittiva si pone l'obiettivo di raccogliere informazioni sulla popolazione di individui che si vuole osservare, riassumendo, organizzando e presentando i dati in maniera ordinata. Questa raccolta di informazioni si divide in due parti:

1. Il calcolo delle **Distribuzioni di Frequenza**, che ci dara' informazioni sulla suddivisione della popolazione tra i vari titoli di studio e tra le regioni.

2. I l calcolo degli **Indici Descrittivi**, che ci daranno informazioni sulla posizione e distribuzione dei dati.

I risulati facilitare la lettura dei risultati dell'analisi, essi verranno rappresentati attraverso opportuni grafici.

2.1 Distribuzioni di Frequenza

Le Distribuzioni di Frequenza hanno come obiettivo quello di evidenziare le diverse modalita' con cui si distribuiscono i dati rispetto alle unita' statistiche osservate, ovvero studiano il loro numero di occorrenze. Il numero esatto di occorrenze e' chiamato **frequenza assoluta**, mentre il numero di occorrenze rispetto al totale viene indicato come **frequenza relativa**.
Piu' formalmente, sia X una variabile e siano $x_1, x_2, ..., x_m$ le modalita' distinte che essa assume. Considerando un campione costituito da n osservazioni di X, si definisce **frequenza assoluta** il valore n_i rappresentante il numero di volte in cui la modalita' x_i e' stata osservata nel campione. Dividendo le frequenze assolute per il numero totale di osservazioni n, si ottengono le **frequenze relative**.
Di seguito andiamo ad analizzare le seguenti frequenze relative:

- le frequenze realitve **dei titoli di studio rispetto al totale**, ovvero come la popolazione si divide in base ai titoli di studio, senza tener conto delle regioni;
- le frequenze relative **delle regioni rispetto al totale**, ovvero come la popolazione si divide alle regioni, senza tener conto del titolo di studio;
- le frequenze realitve **dei singoli titoli di studio**, ovvero come la popolazione avente un determinato titolo di studio si distribuisce sul territorio italiano ;
- le frequenze relative **delle singole regioni** ,ovvero quali sono i titoli di studio piu' diffusi in una determinata regione.

2.1.1 Frequenze relative rispetto al totale

In questa sezione lo scopo e' quello di capire la distribuzione della popolazione, ovvero si vuole calcolare quali sono le regioni e i titoli di studio che hanno il maggior peso in termini della popolazione.
Per effettuare questi calcoli e' necessario, prima, recuperare i seguenti parametri:

- Il totale della popolazione che ha conseguito un titolo;
- Il totale della popolazione per titolo di studio;
- Il totale della popolazione per regione.

Questi valori possono essere calcolati usando i seguenti comandi di `R`:

```
> totalePopolazione <- sum(dati)
> popolazione_per_titolo <- colSums(dati)
> popolazione_per_regione <- rowSums(dati)
```

Otteniamo i seguenti dati mostrati nelle tabelle 2.1 e 2.2.

[redacted]	popolazione_per_titolo
licenza elementare	11095.64
licenza media	16761.11
diploma 3 anni	3218.82
diploma 5 anni	15455.62
laurea	6498.18

Tabella 2.1: Totali per Titolo di Studio

[redacted]	popolazione_per_regione
Piemonte	3862.28
Valle d'Aosta	110.17
Liguria	1413.76
Lombardia	8573.15
Trentino Alto Adige	878.21
Prov Autonoma Bolzano	426.92
Prov Autonoma Trento	451.29
Veneto	4229.90
Friuli-Venezia Giulia	1070.35
Emilia-Romagna	3848.55
Toscana	3272.67
Umbria	787.30
Marche	1354.09
Lazio	4973.06
Abruzzo	1167.26
Molise	278.13
Campania	4886.95
Puglia	3482.42
Basilicata	505.77
Calabria	1721.10
Sicilia	4272.19
Sardegna	1463.85

Tabella 2.2: Totali per Regione

```
> freqRelOrdinate <- sort(popolazione_per_titolo/totalePopolazione)
> barplot(freqRelOrdinate, horiz=TRUE, las=1, border=NA,
...         col=colori[1:5], xlim=c(0, .35),
...         main = "Frequenze Relative per Titoli di Studio")
```

Osserviamo tramite **barplot** le frequenze relative. Possiamo vedere in figura 2.1 che i titoli di studio conseguiti dalla maggior parte della popolazione sono licenza media, diploma 5 anni e licenza elementare conseguiti rispettivamente da circa il 32%, il 29% e il 21% del campione della popolazione, mentre il titolo di studio diploma di 3 anni, definito anche come qualifica professionale, risulta essere quello meno conseguito dalla popolazione, infatti si aggira intorno al 6%.

Vediamo ora invece come sono distribuiti gli studenti fra le regioni.

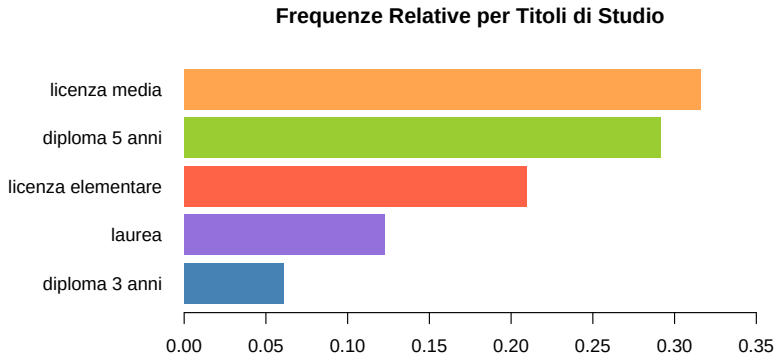

Figura 2.1: Boxplot delle frequenze relative per Titoli di Studio

```
> freqRelOrdinate <- sort(popolazione_per_regione/totalePopolazione)
> barplot(freqRelOrdinate, horiz=TRUE, las=1, border=NA,
...         col=colori, xlim=c(0, .2),
...         main = "Frequenze Relative per Regione")
```

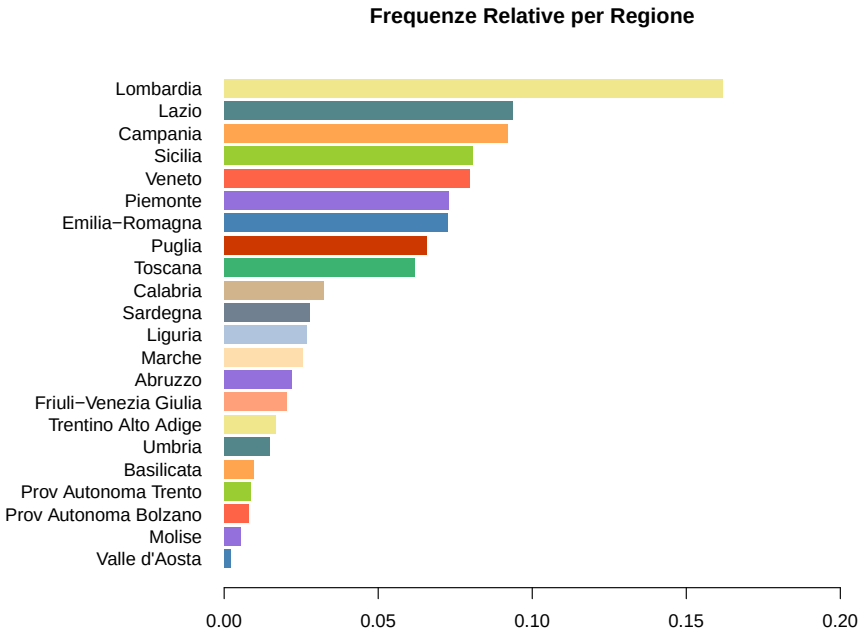

Figura 2.2: Boxplot delle frequenze relative per regione

Si puo' notare dal grafico in figura 2.2 che fra le prime tre regioni in termini di numero della popolazione che ha conseguito un titolo di studio

vi sono presenti la Lombardia, il Lazio e la Campania. Dal grafico risulta che circa il 17% della popolazione avente un titolo di studio proviene dalla Lombardia, seguita da Lazio e Campania con circa il 9% del campione a testa. Tra le regione piazzate nelle ultime posizioni vediamo le due province autonome di Bolzano e Trento con circa l'1.5% seguite dal Molise con circa l'1% e, nell'ultima posizione, abbiamo la Valle d'Aosta con circa lo 0.5%.

2.1.2 Frequenze relative per titolo di studio

In questa parte si effettuera' l'analisi della distribuzione relativa di frequenza dei singoli titoli di studio. Lo scopo e' quello di capire come la popolazione e' distribuita sul territorio in base al proprio titolo di studio. Per effettuare quest'analisi bisogna prima normalizzare le frequenze assolute della popolazione suddivisa in base ai singoli titoli di studio dividendole per il totale della popolazione che ha conseguito quel particolare titolo di studio. Le frequenze realtive cosi' calcolate saranno poi visualizzate tramite grafici di linee interconnesse, per poter osservare immediatamente quali sono le regioni con il maggior numero della popolazione avente un determinato titolo di studio. Siccome i titoli di studio presi in esame sono solo 5, verranno di seguito analizzate tutte le loro frequenze relative.

Distribuzione di frequenza della Licenza Elementare

Cominciamo dal primo titolo di studio. Il grafico a linee interconnesse e' ottenuto tramite le seguenti istruzioni:

```
>   frequenza <- dati$"licenza elementare" / sum(dati$"licenza elementare")
>   plot(frequenza, main="licenza elementare",
...       col="mediumpurple", xaxt="n", xlab="", ylim = c(0, .2))
>   lines(frequenza, col="gray55",type="c")
>   axis(1, at=1:22, labels=row.names(dati), las=2)
```

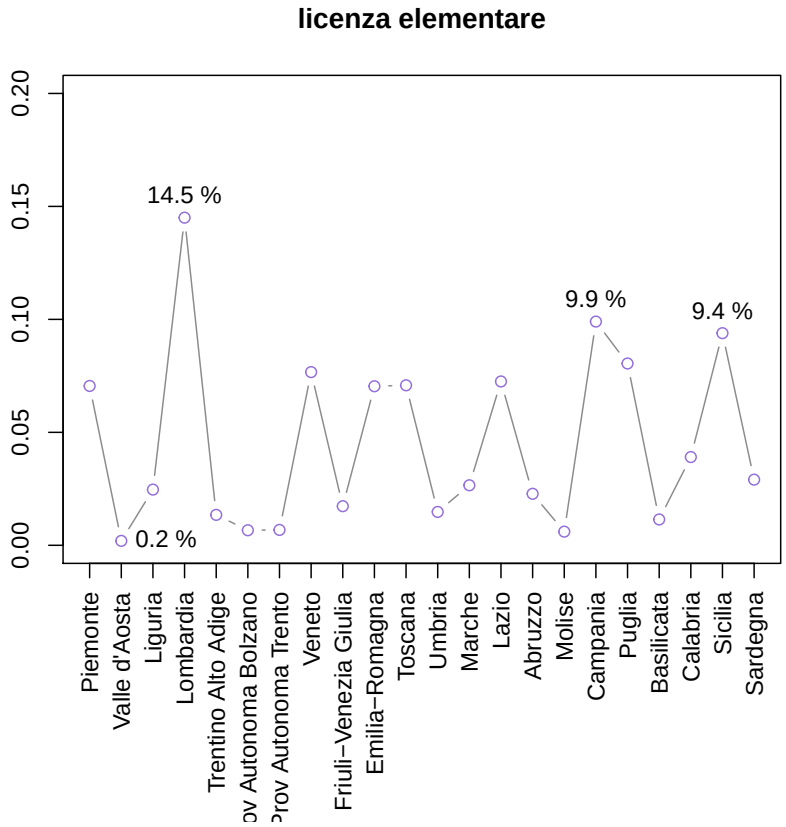

Figura 2.3: Distribuzione di frequenza del titolo di licenza elementare

Il grafico mette in risalto la grande maggioranza della popolazione della Lombardia, circa 14.5%, che ha conseguito il titolo di studio di Licenza Elementare, seguita dalla Campania e dalla Sicilia che hanno una percentuale minore quasi del 5%, ovvero rispettivamente del 9.9% e 9.4%. Nell'ultima posizione, come ci si poteva aspettare, si trova la regione di Valle d'Aosta con il suo modesto 0.2%.

Distribuzione di frequenza della Licenza Media

Passiamo ora a visualizzare il grafico inerente il titolo di studio della Licenza Media. Le istruzioni per ottenerlo sono:

```
>   frequenza <- dati$"licenza media" / sum(dati$"licenza media")
>   plot(frequenza, main="licenza media",
...       col="mediumpurple", xaxt="n", xlab="", ylim = c(0, .2))
>   lines(frequenza, col="gray55",type="c")
>   axis(1, at=1:22, labels=row.names(dati), las=2)
```

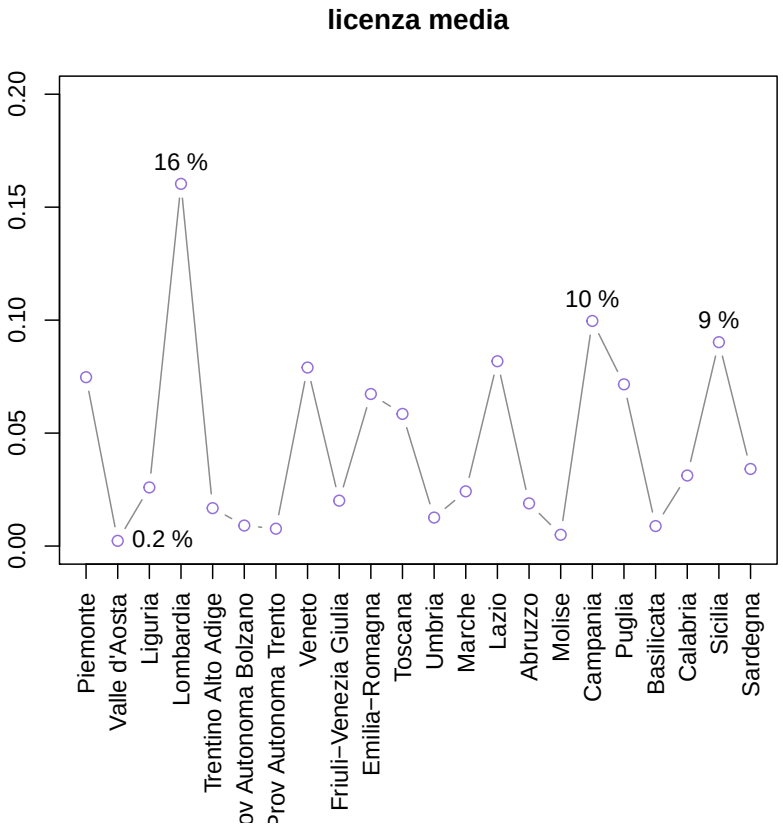

Figura 2.4: Distribuzione di frequenza del titolo di licenza media

Come risulta dal grafico anche in questo caso la regione della Lombardia predomina sulle altre con il suo 16%, seguita dalla Campania con il 10% e dalla Sicilia con il 9%. In ultima posizione la Valle d'Aosta con lo 0.2%.

Distribuzione della frequenza del Diploma 3 anni

Il prossimo titolo di studio preso in analisi e' il Diploma di 3 anni.

```
>   frequenza <- dati$"diploma 3 anni" / sum(dati$"diploma 3 anni")
>   plot(frequenza, main="diploma 3 anni",
...       col="mediumpurple", xaxt="n", xlab="", ylim = c(0, .3))
>   lines(frequenza, col="gray55",type="c")
>   axis(1, at=1:22, labels=row.names(dati), las=2)
```

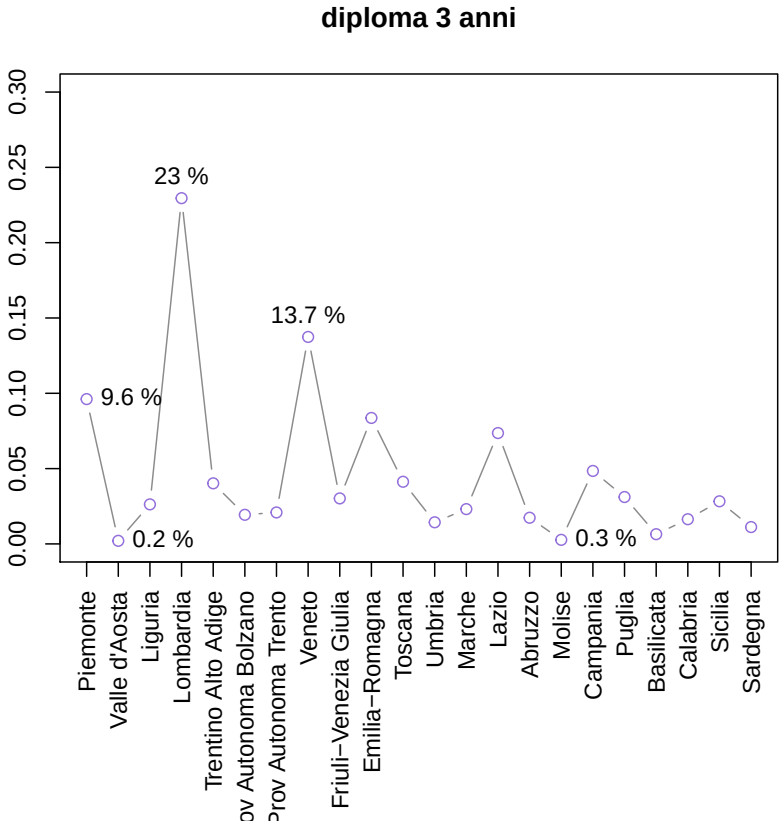

Figura 2.5: Distribuzione di frequenza del titolo di diploma di 3 anni

Il grafico finalmente mostra una novita'. In testa alla classifica rimane ancora la regione Lombardia con il suo 23%, ma questa volta la regione che segue e' quella del Veneto con il 13.7% e il Piemonte con il 9.6%. Anche tra le ultime posizioni possiamo notare che la Valle d'Aosta, con il suo costante 0.2%, non e' piu' sola, infatti e' subito preceduta dal Molise con lo 0.3%.

Distribuzione di frequenza del Diploma di 5 anni

Seguiamo ora con l'analisi del titolo di studio del Diploma di 5 anni.

```
>   frequenza <- dati$"diploma 5 anni" / sum(dati$"diploma 5 anni")
>   plot(frequenza, main="diploma 5 anni",
...       col="mediumpurple", xaxt="n", xlab="", ylim = c(0, .2))
>   lines(frequenza, col="gray55",type="c")
>   axis(1, at=1:22, labels=row.names(dati), las=2)
```

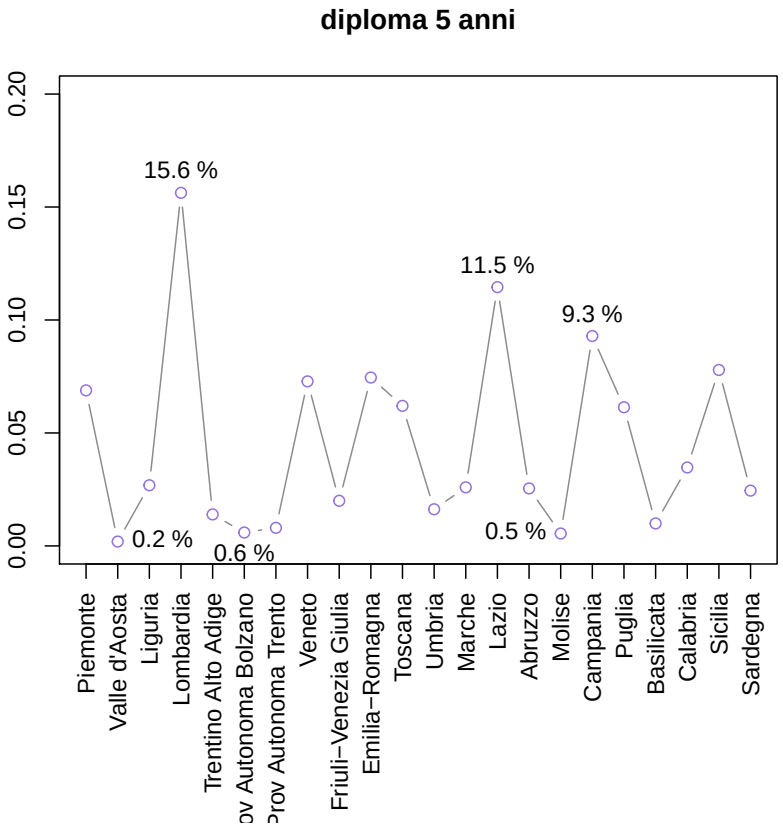

Figura 2.6: Distribuzione di frequenza del titolo di diploma di 5 anni

Anche in questo caso in prima posizione troviamo la Lombardia con il 15.6%, seguita dal Lazio con il 11.5% e dalla Campania con il 9.3%. Tra le ultime posizioni invece troviamo sempre la Valle d'Aosta con lo 0.2%, preceduta dal Molise e dalla Provincia Autonoma di Bolzano con lo 0.5% e lo 0.6% rispettivamente.

Distribuzione di frequenza della Laurea

Passiamo ora all'analisi dell'ultimo titolo di studio, ovvero della Laurea.

```
>   frequenza <- dati$"laurea" / sum(dati$"laurea")
>   plot(frequenza, main="laurea",
...       col="mediumpurple", xaxt="n", xlab="", ylim = c(0, .2))
>   lines(frequenza, col="gray55",type="c")
>   axis(1, at=1:22, labels=row.names(dati), las=2)
```

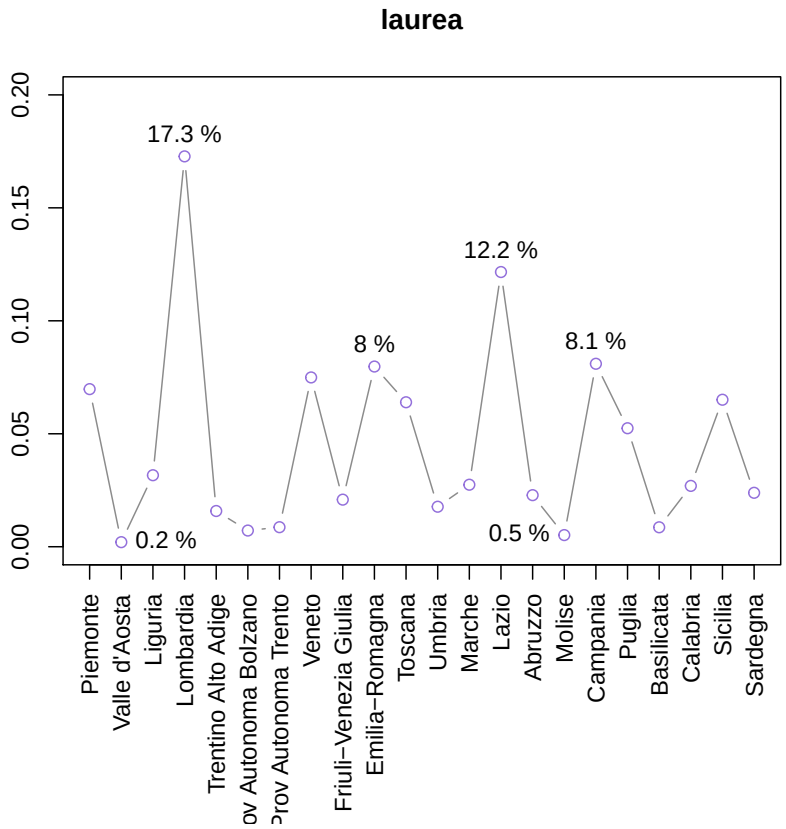

Figura 2.7: Distribuzione di frequenza del titolo di laurea

Come ormai ci potevamo aspettare, in prima posizione abbiamo la Lombardia con il 17.3%, seguita poi dal Lazio con il 12.2% mentre la terza posizione e' combattuta tra la Campania con l'8.1% e l'Emilia-Romagna con l'8%. L'ultima posizione come al solito e' occupata dalla Valle d'Aosta con lo 0.2% preceduta dal Molise con lo 0.5%.

2.1.3 Frequenze relative per Regione

Passiamo ora a esaminare le frequenze relative rispetto alle regioni. Fissata una regione, l'analisi di questo tipo permette di capire come e' distribuita la popolazione rispetto al proprio titolo di studio in quella particolare regione. Andremo, qundi, a scoprire la percentuale della popolazione avente un determinato titolo di studio in una regione fissata. Siccome non e' possibile effettuare l'analisi su tutte le regioni, sceglieremo quelle con il maggior numero di popolazione. L'analisi infatti sara' effettuata sulle seguenti regioni: Lombardia,Campania, Lazio e Sicilia.

Distribuzione di frequenza regione Lombardia

Cominciamo l'analisi dalla regione Lombardia che risultava sempre al primo posto per il numero di popolazione avente qualsiasi titolo di studio. Il calcolo della distribuzione delle frequenza relative rispetto ai titoli di studio sara' effettuato dividendo il numero della popolazione di un singolo titolo di studio per il numero totale della popolazione di quella data regione. Il codice `R` che segue genera il grafico a linee interconnesse mostrato in figura 2.8.

```
> frequenza <- dati["Lombardia",] / sum(dati["Lombardia",])
> plot(as.numeric(frequenza), main="Lombardia",
...       col="tomato", xaxt="n", xlab="", ylim = c(0, .4))
> lines(as.numeric(frequenza), col="gray55", type="c")
> axis(1, at=1:5, labels=colnames(dati), las=2)
```

Dal grafico (figura 2.8) si evince subito che il titolo di studio maggiormente diffuso risulta essere la licenza media, seguita a ruota dal diploma di 5 anni, che insieme costituiscono circa il 60% del totale della popolazione della regione. L'ultima posizione e' occupata dal diploma di 3 anni preceduto dalla laurea.

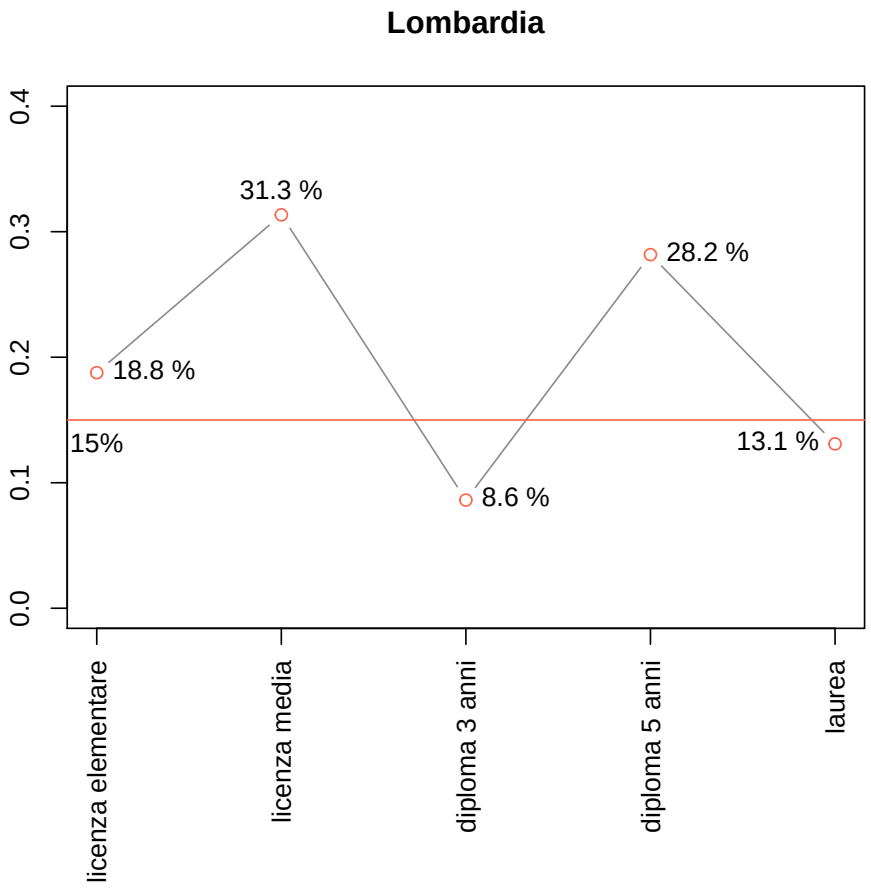

Figura 2.8: Distribuzione di frequenza della regione Lombardia

Distribuzione di frequenza della regione Campania

La seconda regione presa in analisi e' la Campania. Mostriamo il grafico risultante dall'analisi eseguita.

```
> frequenza <- dati["Campania",] / sum(dati["Campania",])
> plot(as.numeric(frequenza), main="Campania",
...     col="tomato", xaxt="n", xlab="", ylim = c(0, .4))
> lines(as.numeric(frequenza), col="gray55", type="c")
> axis(1, at=1:5, labels=colnames(dati), las=2)
```

Il grafico mostrato in figura 2.9 ci fa comprendere che, come al solito la maggior parte della popolazione della Campania ha conseguito un titolo di studio di licenza media o diploma di 5 anni, il dato interessante risulta,

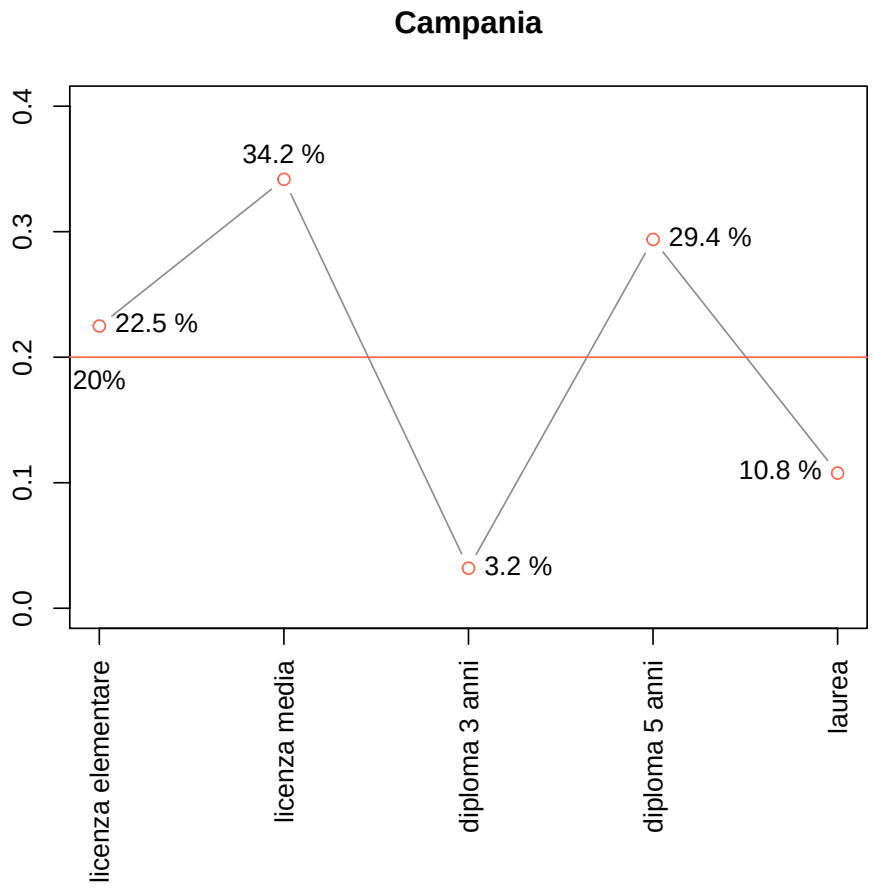

Figura 2.9: Distribuzione di frequenza della regione Campania

invece, il grande numero di persone aventi la licenza elementare, oltre il 22%. Infine possiamo notare che un numero esiguo della popolazione ha conseguito il diploma di 3 anni, poco piu' del 3%.

Distribuzione di frequenza regione Lazio

Passiamo ora a mostrare il grafico della distribuzione delle frequenze inerenti la regione Lazio.

```
> frequenza <- dati["Lazio",] / sum(dati["Lazio",])
> plot(as.numeric(frequenza), main="Lazio",
...     col="tomato", xaxt="n", xlab="", ylim = c(0, .4))
> lines(as.numeric(frequenza), col="gray55", type="c")
> axis(1, at=1:5, labels=colnames(dati), las=2)
```

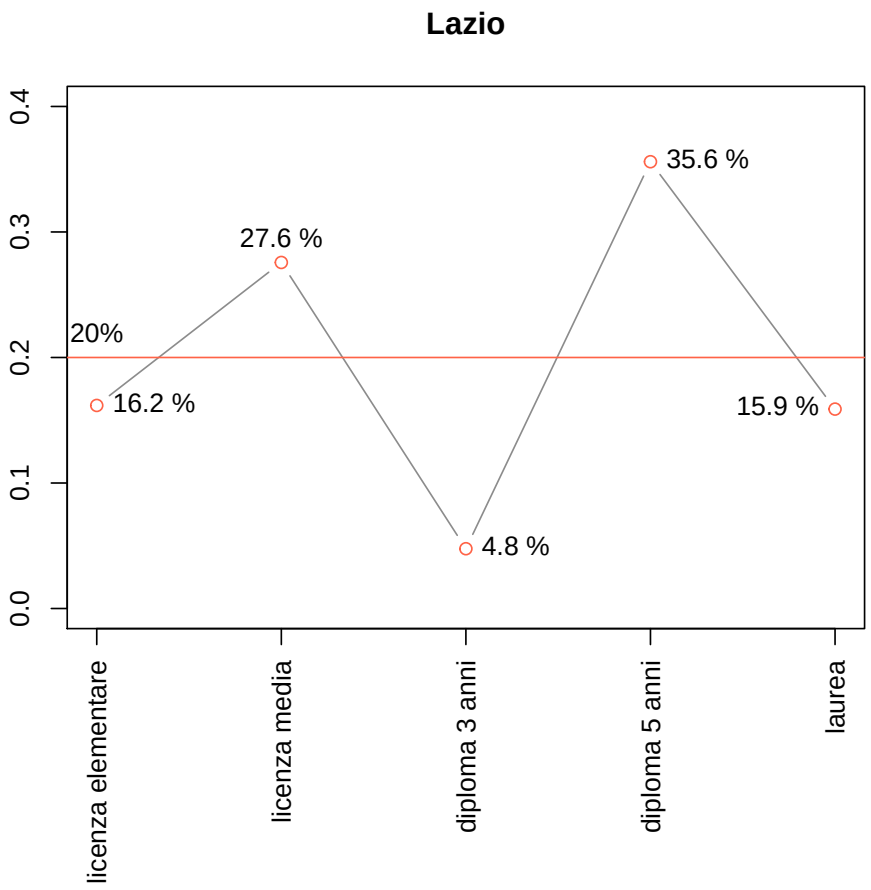

Figura 2.10: Distribuzione di frequenza della regione Lazio

Il grafico della figura 2.10 mostra che oltre il 35% della popolazione del Lazio ha conseguito un diploma di 5 anni e, insieme alla popolazione avente la licenza media, rappresenta oltre il 63% dell'intera popolazione. Un dato

interessante risulta essere la percentuale di popolazione avente la licenza elementare e quella avente la laurea, tutti e due si aggirano intorno al 16%.

Distribuzione di frequenza della regione Sicilia

Infine mostriamo il grafico dell'analisi riguardante l'ultima regione presa in considerazione, ovvero la Sicilia.

```
> frequenza <- dati["Sicilia",] / sum(dati["Sicilia",])
> plot(as.numeric(frequenza), main="Sicilia",
...     col="tomato", xaxt="n", xlab="", ylim = c(0, .4))
> lines(as.numeric(frequenza), col="gray55", type="c")
> axis(1, at=1:5, labels=colnames(dati), las=2)
```

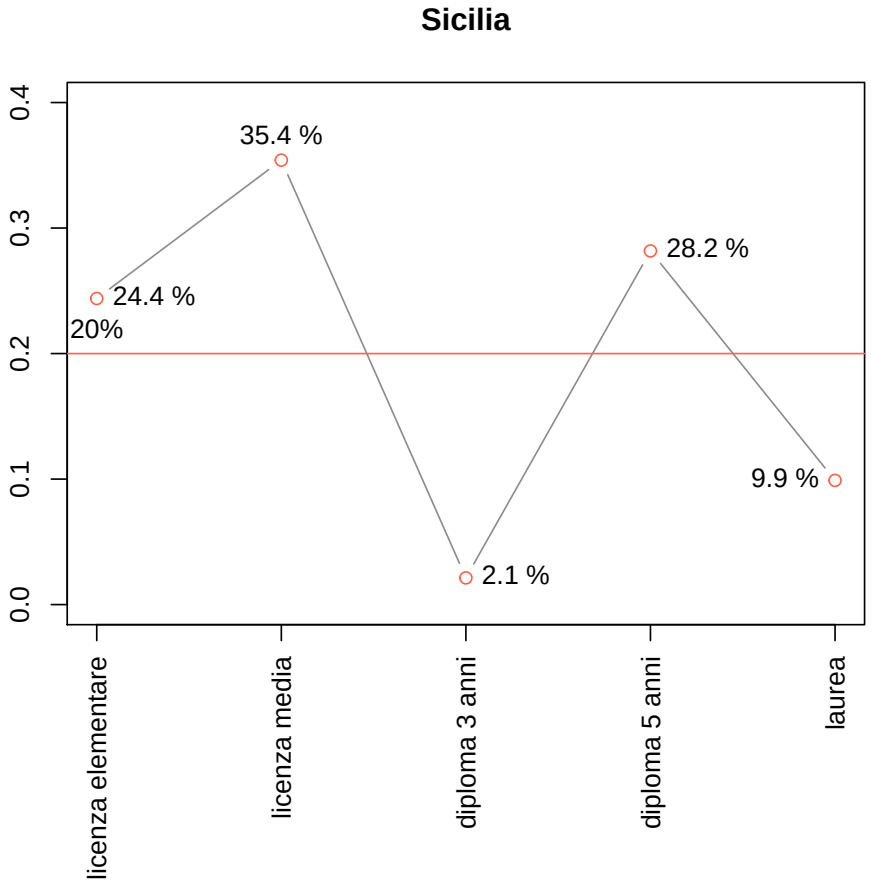

Figura 2.11: Distribuzione di frequenza della regione Sicilia

Come si evince dal grafico in figura 2.11 la regione Sicilia presenta caratteristiche comuni alla regione Campania. Anche qui, infatti, la maggior parte della popolazione risulta avere la licenza elementare, oltre il 34%, seguita da quella avente il diploma di 5 anni, oltre il 28%. Risulta anche un

elevato numero della popolazione avente la licenza elementare, oltre il 24%,e una piccolissima parte avente il diploma di 3 anni, solo il 2.1% del totale.

2.2 Indici Descrittivi

In questa parte daremo una spiegazione degli Indici Descrittivi. Gli Indici Descrittivi sono dei valori statistici che forniscono informazioni caratteristiche dei dati in esame. Essi permettono di sintetizzare in pochi valori le informazioni che consentono di capire come, ad esempio, i dati sono posizionati o quanto sono variabili. Il lato negativo di tale sintesi e' la perdita di informazioni. Di seguito verranno calcolati e analizzati diversi indici descrittivi del campione di dati in esame.

2.2.1 Indici di Posizione e Dispersione

Gli **indici di posizione** sono indici descrittivi che forniscono le informazioni relative alla posizione dei dati nello spazio dei valori. Tra gli indici di posizione abbiamo il minimo, il massimo, la media, la mediana e i quartili. I quartili sono valori ottenuti dividendo l'insieme dei dati ordinati in quattro parti, si ottengono cosi' tre valori:

Primo Quartile (Q_1): valore tale che il 25% dei dati e' minore o uguale a Q_1;

Secondo Quartile (Q_2): valore tale che il 50% dei dati e' minore o uguale a Q_1. Coincide con la **mediana**;

Terzo Quartile (Q_3): valore tale che il 75% dei dati e' minore o uguale di Q_3.

Gli indici di posizione si possono ottenere tramite il comando `R summary`, che prende per parametro l'oggetto contenente i dati su cui calcolare gli indici. Ovvero

```
> indici <- summary(dati)
```

tuttavia, per ottenere una formattazione piu' leggibile, e' stato usato il seguente codice R, che produce direttamente la tabella 2.3.

```
> indici <- t(sapply(dati, summary))
> rws <- seq(1, nrow(indici), by=2)
> colore <- rep("\\rowcolor[gray]{0.95}", length(rws))
> tabella <- xtable(indici, caption="Indici di posizione",
...                   label="tab:statistiche1")
> # align(tabella) <- c( 'l', rep('p{.6in}',6) )
> print(tabella, booktabs=TRUE, size="footnotesize",
...       add.to.row = list(pos = as.list(rws), command = colore))
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
licenza elementare	21.80	153.10	308.60	504.30	799.90	1609.00
licenza media	38.76	228.90	479.20	761.90	1240.00	2687.00
diploma 3 anni	6.61	53.56	87.78	146.30	150.20	738.90
diploma 5 anni	29.83	224.20	407.80	702.50	1110.00	2415.00
laurea	13.18	105.90	176.70	295.40	445.70	1123.00

Tabella 2.3: Indici di posizione

Il punto negativo di tali indici risulta essere il fatto di non tenere conto della dispersione dei dati, pertanto e' necessario recuperare altri tipi di indici, gli **indici di dispersione**. Gli indici di dispersione forniscono informazioni su quanto i dati si discostano dalla loro media, in particolare:

- La **varianza** indica quando i dati si discostino quadraticamente dalla media;
- La **deviazione standard** indica quanto i dati si discostano dalla media, con la stessa unita' di misura dei dati stessi;
- Il **coefficiente di variazione** e' una misura relativa della dispersione dei dati, un numero puro, che si ottiene dividendo la deviazione standard per il valore assoluto della media, ovvero:

$$\sigma^* = \frac{\sigma}{|\mu|}$$

Creiamo una tabella contente gli indici di dispersione. Con il seguente codice inizializziamo un data frame che conterra' la media (per riferimento) e gli indici di dispersione:

```
> numRighe <- 5
> numColonne <- 4
> nomiCol <- c("Media", "Varianza", "Deviazione Standard",
...               "Coefficiente di variazione")
> statistiche <- as.data.frame(matrix(rep(0, numRighe * numColonne),
...                              nrow = numRighe, ncol=numColonne),
...                              row.names= names(dati))
> names(statistiche) <- nomiCol
```

quindi popoliamo il data frame tramite il codice che segue all'interno del quale, per ogni titolo di studio, calcoliamo media, varianza, deviazione standard e coefficiente di variazione:

```
> for(titolo in names(dati)){
...  media <- mean(dati[[titolo]])
...  varianza <- var(dati[[titolo]])
```

```
...  devStd <- sd(dati[[titolo]])
...  cv <- devStd / abs(media)
...  statistiche[titolo,1] <- media
...  statistiche[titolo,2] <- varianza
...  statistiche[titolo,3] <- devStd
...  statistiche[titolo,4] <- cv
...}
```

[redacted]	Media	Varianza	Deviazione Standard	Coefficiente di variazione
licenza elementare	504.35	187153.04	432.61	0.86
licenza media	761.87	465141.28	682.01	0.90
diploma 3 anni	146.31	29164.19	170.78	1.17
diploma 5 anni	702.53	397271.03	630.29	0.90
laurea	295.37	77337.81	278.10	0.94

Tabella 2.4: Indici di dispersione

La tabella 2.4 mostra gli indici di dispersione calcolati. Osservando i coefficienti di variazione notiamo che tutti risultano minori di 1, tranne quello riguardante il diploma di 3 anni. Un coefficiente di variazione piccolo indica dati che si discostano di meno dalla media. Inoltre, per capire di quanto i dati si discostano dalla media, bisogna osservare la deviazione standard, la quale ha la stessa unita' di misura dei dati.

2.2.2 Boxplot

I **boxplot** consentono di visualizzare le informazioni ottenute tramite gli indici descrittivi in modo veloce ed intuitivo. Essi consistono in un rettangolo i cui lati inferiore e superiore rappresentano rispettivamente il primo ed il terzo quartile, attraversato da una linea orizzontale rappresentante la mediana, o secondo quartile. Dal rettangolo escono due segmenti che terminano in corrispondenza dei valori minimo e massimo. L'ampiezza del boxplot e' un indice di dispersione, denominato **scarto interquartile**, in quanto mostra la differenza fra il terzo ed il primo quartile. Se presenti, il boxplot mostra anche la presenza di **outliers**, ovvero valori che si discostano nettamente dalla media. Il comando `R boxplot` consente di visualizzare il boxplot di una serie di dati. Di seguito verranno mostrati i boxplot dei vari titoli di studio.

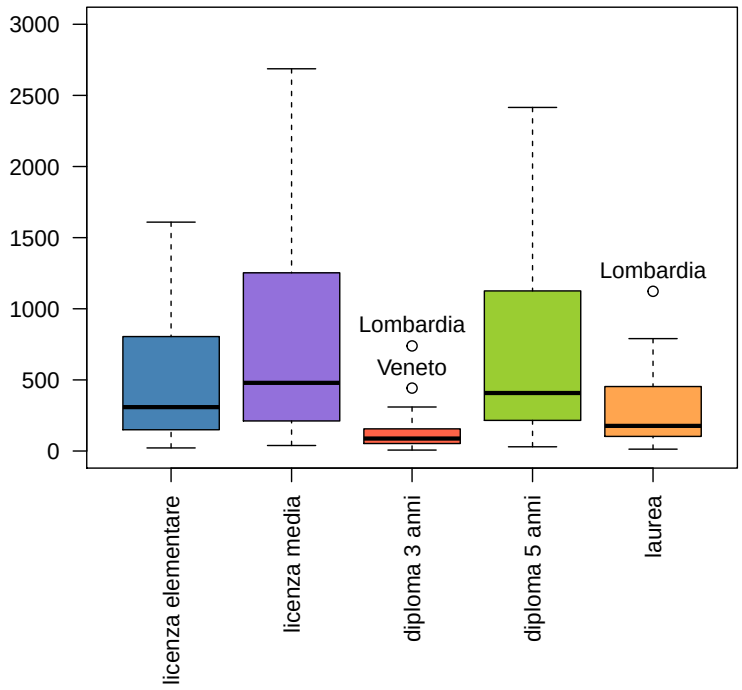

Figura 2.12: Boxplot

Il grafico in figura 2.12 mostra come il diploma di 3 anni risulta avere ben due outliers, ovvero il Veneto e la Lombardia, quest'ultima e' un outlier anche della Laurea. Come si puo' notare il diploma di 3 anni e la Laurea

hanno mediane piuttosto vicine come anche la licenza media e il diploma di 5 anni. Infine possiamo notare che il titolo di studio avente la minore dispersione di dati e' quello del diploma di 3 anni, questo perche' il minimo, il massimo, il primo e il terzo quartile sono tutti molto vicini alla mediana.

Capitolo 3

Analisi dei cluster

L'**analisi dei cluster** e' una metodologia che consiste nel raggruppare entita' in sottoinsiemi (cluster) in base a similarita' , con l'obiettivo di scoprire quali sono, nel caso ve ne siano, le caratteristiche che meglio discriminano i dati. Gli elementi che si trovano all'interno di uno stesso cluster devono risultare piu' "vicini" possibile, mentre elementi presenti all'interno di cluster diversi devono essere quanto piu' "lontani". La metrica utilizzata per misurare la distanza fra gli elementi caraterizza il metodo utilizzato. Nel caso in esame, eseguire l'analisi dei cluster significa vedere quali regioni presentano similitudini nella distribuzione dei vari titoli di studio.

Esistono diversi metodi di clustering, che poissiamo dividere in tre famiglie:

Enumerativi: tali metodi determinano tutte le possibili partizioni di n elementi per poi scegliere come cluster quelli che massimizzano una funzione obiettivo;

Gerarchici: questi metodi si dividono a loro volta in **agglomerativi** e **divisivi**; i primi partono da n cluster contenenti un singolo elemento, per poi combinarli progressivamente a coppie, scegliendo ogni coppia in base ad una funzione di distanza. I secondi, invece, partono da un singolo cluster contenente tutti gli n elementi per poi effettuare separazioni successive, fino ad ottenere cluster di cardinalita' unitaria.

Non gerarchici: sono metodi che, al contrario dei metodi gerarchici, consentono il riassegnamento di un elemento ad un cluster diverso, procedendo iterativamente fino ad un numero massimo di iterazioni o finche' non viene rispetatto un criterio che determina il raggiungimento della divisione migliore.

3.1 Metodi gerarchici

Nell'analisi che si vuole eseguire in questo lavoro ci concentreremo sui metodi gerarchici additivi. Questi metodi si distinguono a seconda di quale criterio viene usato per scegliere i due cluster da unire. I vari tipi di metodi sono di seguito elencati:

- Metodo del legame singolo;
- Metodo del legame completo;
- Metodo del legame medio;
- Metodo del centroide;
- Metodo della mediana.

Ognuno di questi metodi verra' analizzato singolarmente nel prosieguo di questo documento. Indipendentemente dal metodo scelto, pero', bisogna eseguire due operazioni fondamentali, ovvero **normalizzare i dati** e calcolare la **matrice delle distanze**.La prima operazione viene eseguita perche' le varie colonne di una tabella di dati potrebbero non rappresentare tutte la stessa grandezza (metri, kg, litri,...), per questo motivo i dati vanno normalizzati in una stessa scala che risulta indipendente dalla dimensione. La matrice delle distanze, invece, viene semplicemente calcolata per effettuare piu' agevolmente l'analisi. Le due operazioni preliminari sono effettuate tramite il seguente codice `R`:

```
> datiScalati <- scale(dati)
> distanze <- dist(datiScalati, method="euclidean", diag=TRUE, upper=TRUE)
```

Passiamo ora ad analizzare i risultati forniti dai vari metodi di clustering usati.

3.1.1 Metodo del legame singolo

Nel metodo del **legame singolo** la distanza fra due cluster C_1 e C_2 e' definita come la distanza minima fra una qualsiasi coppia di elementi $n_1 \in C_1$ e $n_2 \in C_2$. La coppia di cluster che presenta la distanza minima viene scelta per l'unione. In pratica si sfrutta il concetto che sta dietro il Nearest Neighbour Algorithm per unire due cluster vicini. In `R`, effettuiamo il clustering gerarchico con il metodo del legame singolo tramite la seguente istruzioni:

```
> singleLinkage <- hclust(distanze, method="single")
```

Uno dei pregi dei metodi gerarchici e' di poter decidere in quanti cluster dividere i dati dopo aver eseguito l'algoritmo. Per decidere qual e' il numero di cluster piu' adatto, bisogna determinare quale agglomerazione causa il

maggior aumento della distanza di aggregazione. A questo scopo esaminiamo lo **screeplot**, che permette di visualizzare le distanze di agglomerazione ad ogni passo. Otteniamo lo screeplot tramite le seguenti istruzioni, possiamo osservare il risultato in figura 3.1.

```
> plot(rev(c(0, singleLinkage$heigh)), seq(1,22), type="b",
...     main="Screeplot legame singolo", col = "steelblue3",
...     xlab="Distanza di aggregazione", ylab="Numero di cluster")
```

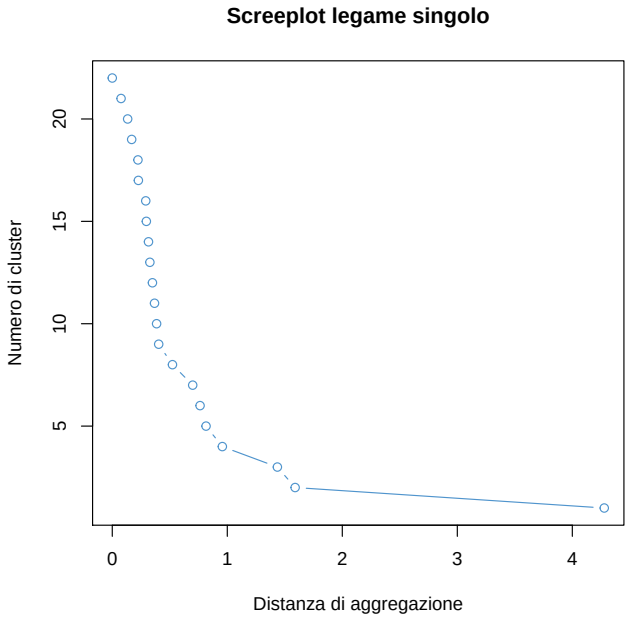

Figura 3.1: Screeplot - metodo del legame singolo

Dal grafico salta subito all'occhio che l'aumento maggiore di distanza si ha nel passaggio da un solo cluster a due cluster, quindi saremmo propensi a usare solo due cluster nell'esecuzione del metodo scelto. Un analisi piu' accurata ci mostra invece che un numero migliore di cluster risulta essere 4. In figura 3.2 possiamo vedere il **dendogramma**, che permette di visualizzare l'ordine di agglomerazione dei cluster. La funzione `plot` permette la visualizzazione del dendogramma, mentre la funzione `rect.hclust` ci permette di evidenziare i cluster ottenuti. Mostreremo di seguito alcuni **dendogrammi** aventi diverso numero di cluster per mettere in risalto il fatto che spetta sempre all'utente la scelta ottimale del numero di cluster, nel caso questi siano visibili facilmente sui grafici usati.

```
> plot(singleLinkage, xlab="Regioni", ylab="",
...      sub="Metodo del legame singolo")
> rect.hclust(singleLinkage, k=2, border="steelblue3")
```

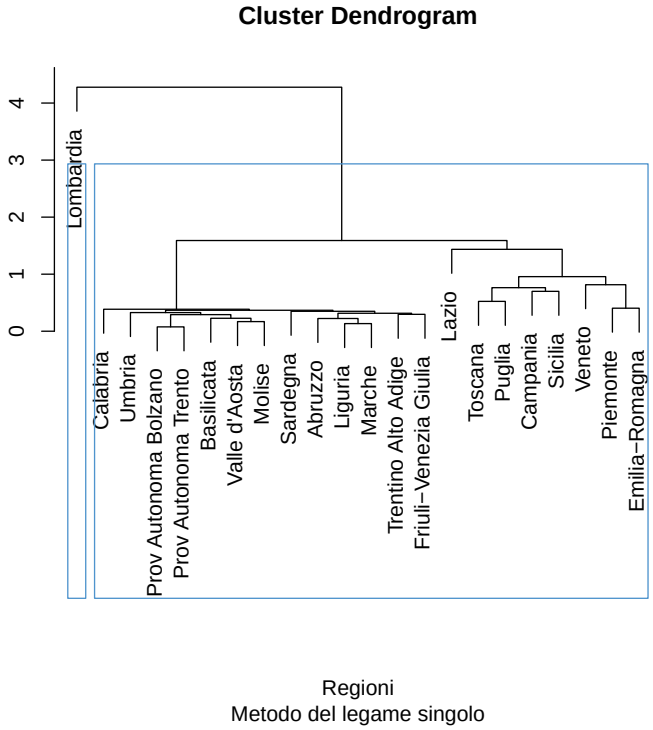

Figura 3.2: Dendogramma - metodo del legame singolo

Dal dendogramma in figura 3.2 si capisce subito che la scelta di utilizzare due cluster sia sbagliata. Infatti nell'analisi delle frequenze ci siamo gia' accorti che la regione **Lombardia** aveva risultati nettamente maggiori rispetto a tutte le altre regione, questa caratteristica si evince ancor piu' chiaramente durante l'analisi dei cluster. La regione Lombardia risulta essere un cluster a se stante, mentre le altre regioni possono essere agglomerate insieme. Infatti, dopo una breve analisi, la scelta del numero di cluster ottimale e' ricaduta sull'uso di 4 cluster, come mostrato in figura 3.3. Da quest'ultimo **dendogramma** si capisce subito di aver effettuato la scelta giusta.

```
> plot(singleLinkage, xlab="Regioni", ylab="",
...     sub="Metodo del legame singolo")
> rect.hclust(singleLinkage, k=4, border="steelblue3")
```

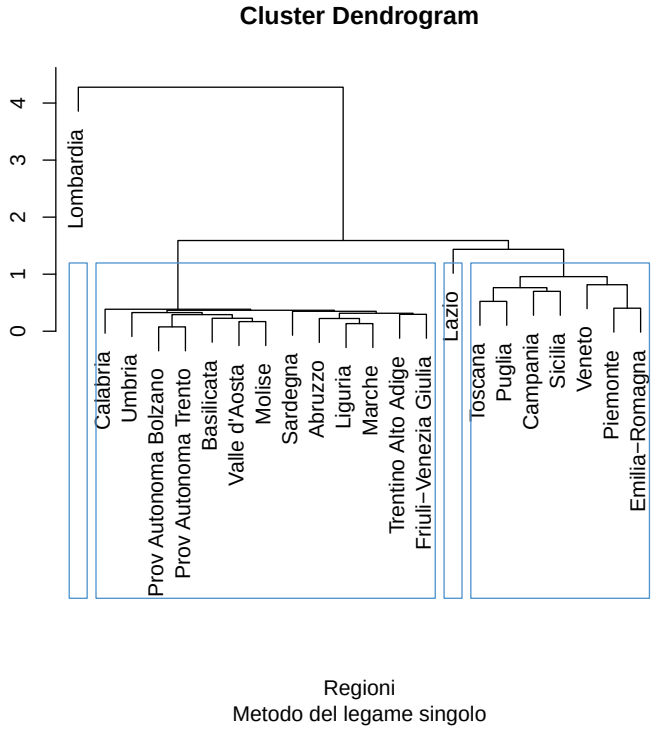

Figura 3.3: Dendogramma - metodo del legame singolo

3.1.2 Metodo del legame completo

Il metodo del **Legame Completo** utilizza una tecnica di agglomerazione completamente opposta a quella sfruttata nel metodo del legame singolo. In questo metodo la distanza fra due cluster C_1 e C_2 e' definita come la distanza massima fra una qualsiasi coppia di elementi $n_1 \in C_1$ e $n_2 \in C_2$. Ad ogni passo viene calcolata questa distanza tra una qualsiasi coppia di cluster e, la coppia con il valore della distanza maggiore, viene agglomerata. Passiamo ora ed eseguire il clustering usando la seguente funzione:

```
> completeLinkage <- hclust(distanze, method="complete")
```

ed esaminiamo lo screeplot di figura 3.4 per decidere il numero ottimale di cluster.

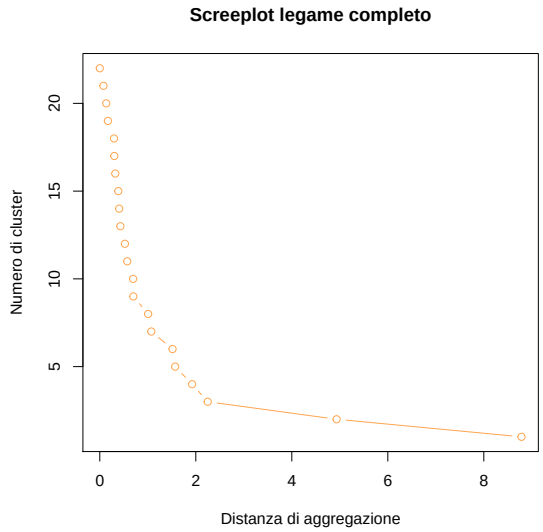

Figura 3.4: Screeplot - metodo del legame completo

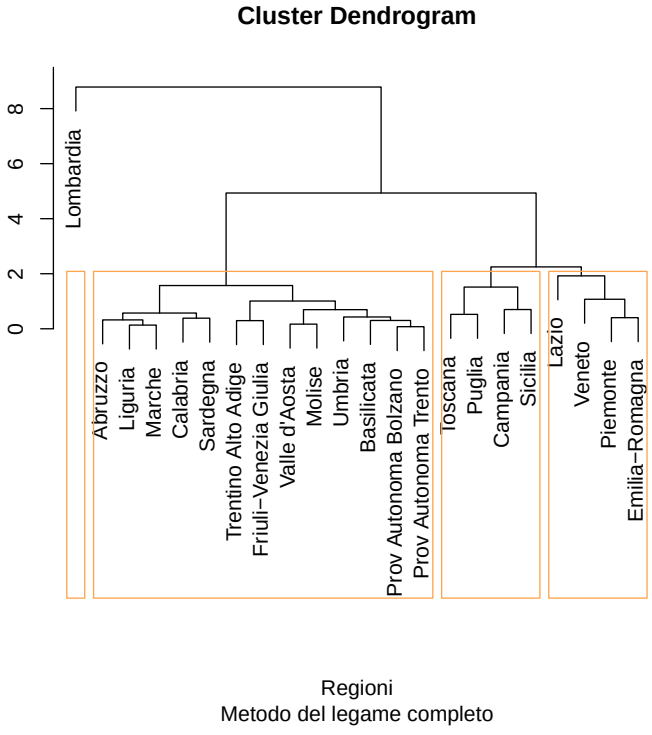

Figura 3.5: Dendogramma - metodo del legame completo

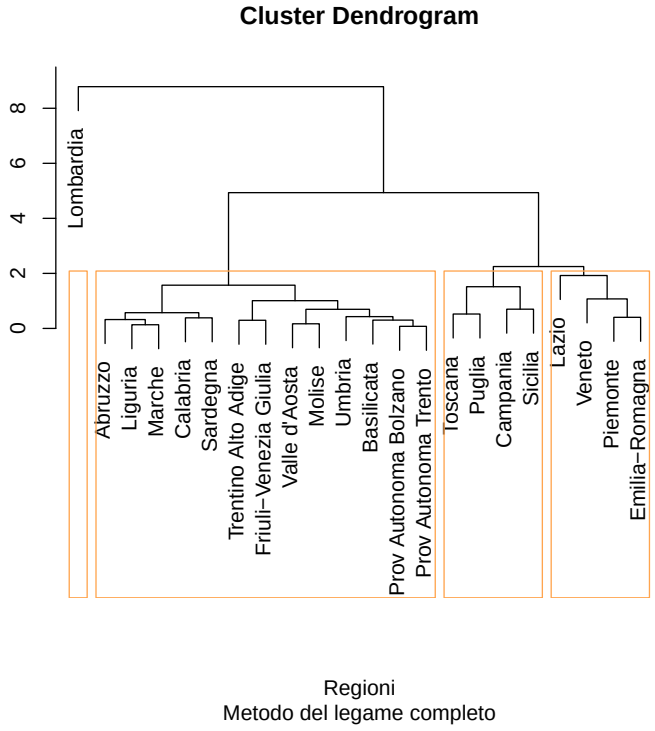

Figura 3.6: Dendogramma - metodo del legame completo

Dal dendogramma in figura 3.5 possiamo osservare che usando il metodo del **legame completo** risulta ancor piu' difficile scegliere il numero ottimale di cluster. Da una prima analisi si e' tentati di utilizzare solo 3 cluster. Avendo, pero', effettuato in precedenza l'analisi con il metodo del legame singolo, anche questa volta scegliamo 4 come numero ottimale di cluster. Il risultato della scelta effettuato lo visualizziamo nel dendogramma in figura 3.6.

3.1.3 Metodo del legame medio

Il metodo del **legame medio** usa come distanza tra due cluster la media delle distanze fra tutti gli elementi che vi appartengono. Anche questa volta eseguiamo il clustering tramite la seguente istruzione:

```
> averageLinkage <- hclust(distanze, method="average")
```

Osservando lo screeplot in figura 3.7 notiamo subito che il risultato mostrato e' uguale a quelli ottenuti tramite i due metodi precedenti.

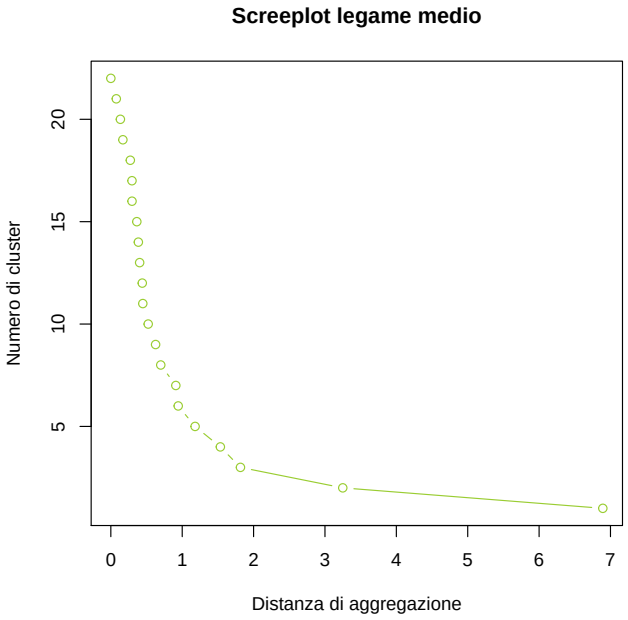

Figura 3.7: Screeplot - metodo del legame medio

Il dendogramma in figura 3.8 ci conferma i sospetti nati osservando lo screeplot 3.7, ovvero che il risultato e' uguale a quelli osservati in precedenza. Di conseguenza scegliamo di nuovo 4 come numero di cluster ottimali e visualizziamo la nostra scelta in figura 3.8.

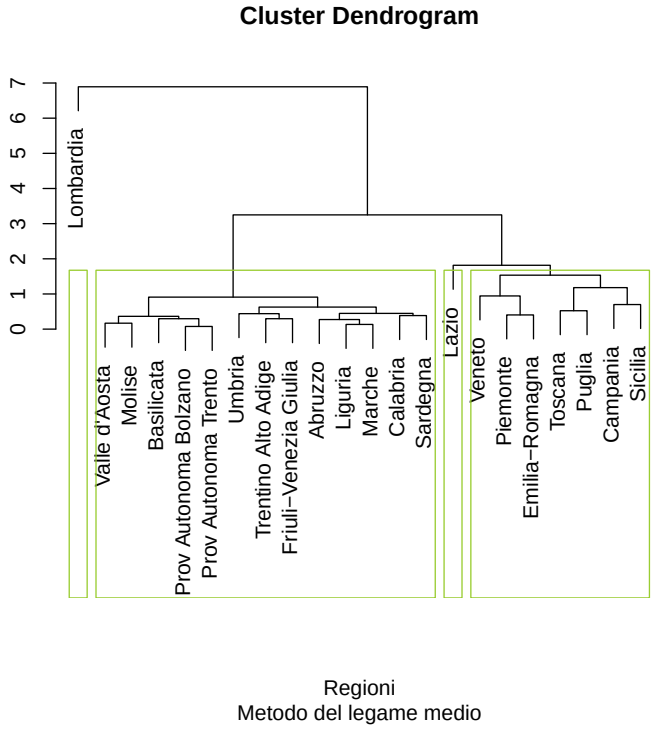

Figura 3.8: Dendogramma - metodo del legame medio

3.1.4 Metodo del centroide

In questo metodo la distanza tra due cluster C_1 e C_2 e' definita come la distanza tra i centroidi dei due cluster. Il centroide e' il punto ottenuto come media campionaria degli elementi di un cluster. La coppia di cluster che presenta la minima distanza quadrata tra centroidi viene scelta per l'unione. Il clustering viene effettuato tramite la seguente istruzione:

```
> centroidHclust <- hclust(distanze**2, method="centroid")
```

Creiamo e analizziamo lo screeplot, mostrato in figura 3.9. Ci accorgiamo subito di essere davanti a un risultato non troppo distante dai precedenti.

Osservando il dendogramma in figura 3.10 possiamo essere tentati dal dire che, questa volta, sia 3 il numero di cluster migliore da usare. Dopo un'analisi leggermente migliore, affermiamo invece che l'uso di 4 cluster ci fornisce un risultato migliore.

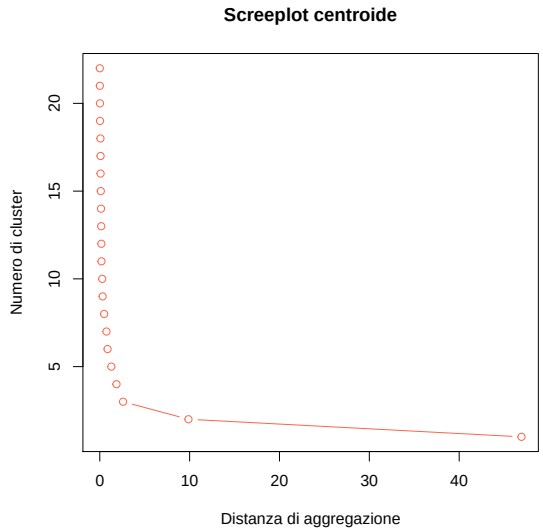

Figura 3.9: Screeplot - metodo del centroide

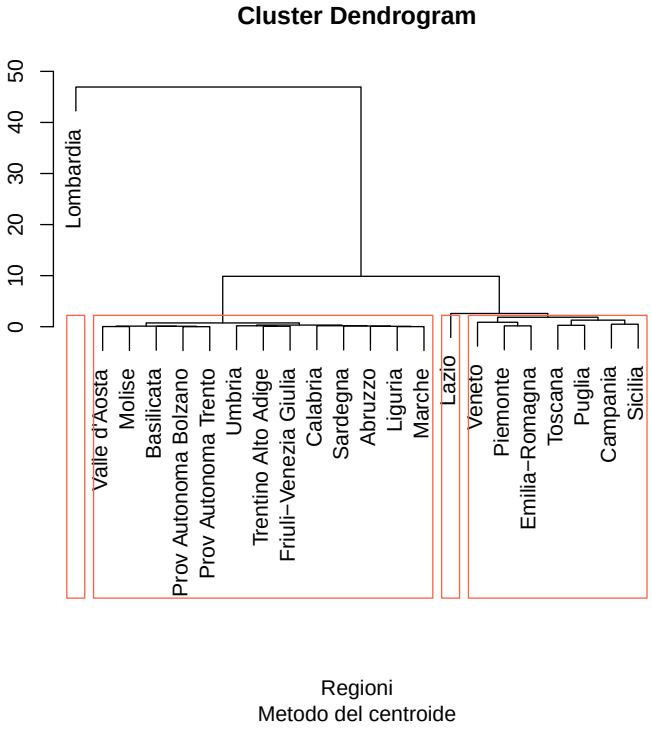

Figura 3.10: Dendogramma - metodo del centroide

3.1.5 Metodo della mediana

Il metodo della **mediana** e' simile a quello del centroide, con la differenza che la procedura e' indipendente dalla numerosita' dei cluster. Infatti, quando due gruppi si aggregano, il nuovo centroide e' calcolato come la semisomma dei due centroidi precedenti. Il seguente comando esegue il clustering con questo metodo:

```
> medianHClust <- hclust(distanze**2, method="median")
```

In figura 3.11 viene rappresentato lo screeplot per questo metodo. Il risultato mostrato sembra ancora una volta simile a quelli ottenuti in precedenza, ma il dendogramma che vedremo di seguito ci dimostra che non e' cosi'.

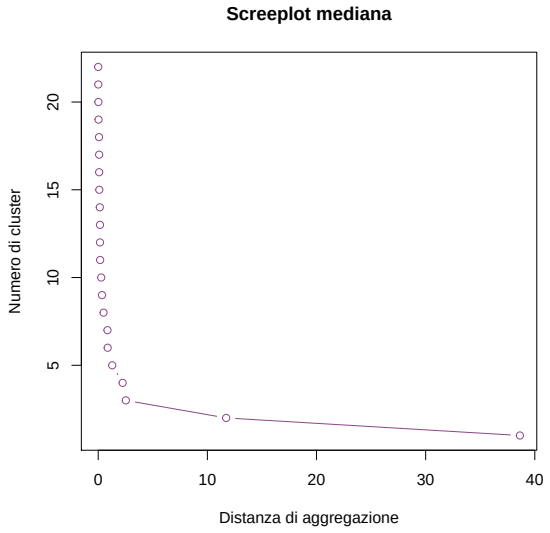

Figura 3.11: Screeplot - metodo della mediana

La figura 3.12 mostra il dendogramma ottenuto in output. Si puo' notare immediatamente che non risulta piu' tanto facile individuare il numero migliore di cluster da usare. Infatti due potrebbero essere le proposte migliori: utilizzare 3 cluster o 5. In quest'analisi si e' scelti di sfruttare il secondo valore come mostrato nel dendogramma.

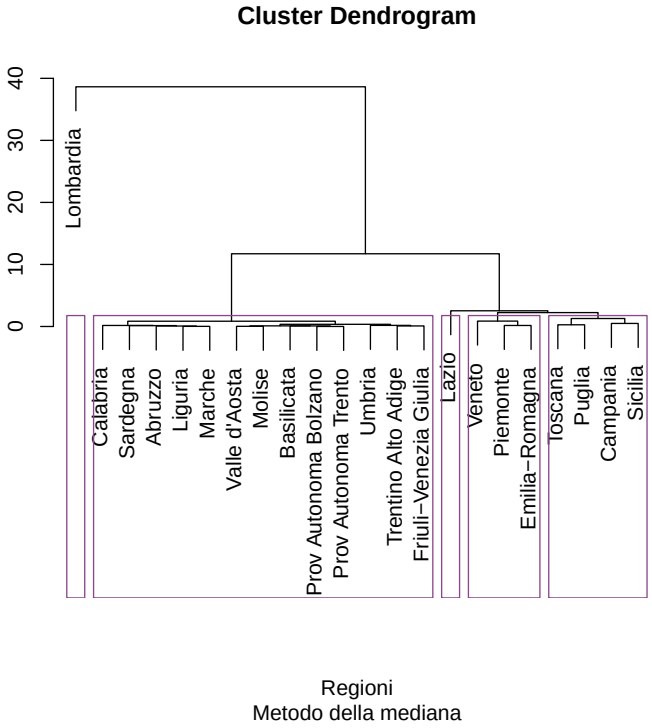

Figura 3.12: Dendogram - metodo della mediana

3.2 Misure di non omogeneita'

Una volta che abbiamo calcolato i cluster, il passo successivo e' quello di verificare quanto siano buoni i risultati ottenuti. Le **misure di non omogeneita' statistica** ci forsniscono proprio quest'informazione, in quanto ci dicono quanto i singoli cluster sono compatti e quanto sono distanti fra loro. Un cluster si dice di buona qualita' se la distanza tra gli elementi di un singolo cluster e' minimizzata mentre la distanza tra cluster differenti e' massimizzata. Di segutio calcoliamo la misura di non omogeneita' sul totale dei dati, e successivamente effettuiamo il calcolo per i cluster ottenuti tramite il metodo del legame singolo.

```
> n <- nrow(dati)
> trHI <- (n-1) * sum(apply(dati, 2, var))
> trHI
[1] 24277414
```

Ora calcoliamo le misure di non omogeneia' dei singoli cluster. Creiamo una matrice per ogni cluster: utilizzando la funzione `cutree` e prendendo

l'ultima colonna otteniamo l'assegnamento delle singole regioni. La funzione `which` permette di ottenere gli indici delle righe che rispettano un certo criterio, utilizziamo questi indici per recuperare i dati dalla matrice dei dati principale.

```
> cluster <- cutree(singleLinkage, k=3)
> slG1 <- dati[which(cluster == 1),]
> slG2 <- dati[which(cluster == 2),]
```

Calcoliamo la misura di non omogeneita' di G1:

```
> n1 <- nrow(slG1)
> trHslG1 <- (n1 - 1) * sum(apply( slG1, 2, var))
> trHslG1
[1] 1197487
```

Calcoliamo la misura di non omogeneita' di G2:

```
> n1 <- nrow(slG2)
> trHslG2 <- (n1 - 1) * sum(apply( slG2, 2, var))
> trHslG2
[1] 882197.8
```

Siccome il cluster che corrisponde alla regione Lombardia e' composto da un unico elemento, il valore della sua misura di non omogeneita' e' pari a zero.

Quindi calcoliamo la non omogenita' **intracluster** (within) e **intercluster** (between):

```
> withinSL <- trHslG1 + trHslG2
> betweenSL <- trHI - trHslG1 - trHslG2
> withinSL
[1] 2079684
> betweenSL
[1] 22197730
```

Useremo lo stesso codice per calcolare la misura di non omogeneita' degli altri metodi di clustering gerarchico. I risultati ottenuti sono riassunti nella tabella 3.1

Le misure appena calcolate confermano la scelta del numero di cluster effettuata precedentemente. Infatti la distanza intracluster risulta molto minore rispetto a quella intercluster. I risultati sono simili per qualsiasi metodo grazie al fatto che i metodi usati hanno restituito quasi sempre gli stessi cluster.

[redacted]	Within	Between
Singolo	2079684.34	22197729.79
Completo	2079684.34	22197729.79
Medio	2079684.34	22197729.79
Centroide	2079684.34	22197729.79
Mediana	2079684.34	22197729.79

Tabella 3.1: Misure di non omogeneita'

3.3 Metodi non gerarchici

I metodi non gerarchici affrontano il problema della clusterizzazione in maniera differente da quelli gerarchici. Questi metodi effettuano un unico partizionamento iniziale che viene raffinato ad ogni iterazione,ovvero riallocano gli elementi ad ogni passo fino al raggiungimento della convergenza (soddisfacimento di determinati criteri) o di un numero massimo di iterazioni. Uno dei metodi piu' diffusi di questa categoria e' il metodo del **k-means**. In questa analisi verra' utilizzato proprio il metodo del **k-means**.

3.3.1 Metodo del K-Means

Il metodo del **k-means** richiede che il numero k di cluster da ottenere venga fissato a priori. L'algoritmo ottiene i cluster cercando di minimizzare la funzione di distorsione

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \, ||x_n - \mu_k||^2$$

dove

- N e' il numero di elementi;
- K e' il numero di cluster;
- $r_{nk} \in \{0, 1\}$ e' una variabile decisionale che vale 1 se l'elemento n e' stato assegnato al cluster k;
- μ_k e' il centro del cluster k-esimo.

Dopo aver inizializzato i centri μ_k (in modo casuale o con centri scelti a priori), l'algoritmo cerca di minimizzare la funzione alternando due passi:

Passo 1: si associa ogni elemento al cluster il cui centro e' piu' vicino, ovvero

$$r_{nk} = \begin{cases} 1 & \text{se } k = \arg\ \min_j ||x_n - \mu_j||^2 \\ 0 & \text{altrimenti} \end{cases}$$

Passo 2: Con i nuovi cluster creati, si ottimizzano i centri dei cluster ricalcolando la media sui nuovi elementi

$$\mu_k = \frac{1}{N_k} \sum_n r_{nk} x_n$$

dove N_k e' il numero di elementi appartenenti al cluster k.

Questi due passi si alternano fino a convergenza o fino ad un numero massimo di iterazioni. In R possiamo eseguire il k-means tramite il comando che segue:

```
> km <- kmeans(datiScalati, centers=3, iter.max=10, nstart=8)
> km
K-means clustering with 3 clusters of sizes 13, 8, 1

Cluster means:
  licenza elementare licenza media diploma 3 anni
1         -0.7304860    -0.7071598     -0.5225055
2          0.8678366     0.7962842      0.4153508
3          2.5536251     2.8228042      3.4697658
  diploma 5 anni     laurea
1     -0.7018333 -0.6690000
2      0.8007824  0.7152273
3      2.7175740  2.9751812

Clustering vector:
             Piemonte          Valle d'Aosta
                    2                      1
              Liguria              Lombardia
                    1                      3
  Trentino Alto Adige Prov Autonoma Bolzano
                    1                      1
 Prov Autonoma Trento                 Veneto
                    1                      2
Friuli-Venezia Giulia         Emilia-Romagna
                    1                      2
              Toscana                 Umbria
                    2                      1
               Marche                  Lazio
                    1                      2
              Abruzzo                 Molise
                    1                      1
             Campania                 Puglia
                    2                      2
```

```
          Basilicata              Calabria
                   1                     1
             Sicilia              Sardegna
                   2                     1

Within cluster sum of squares by cluster:
[1] 3.551436 7.774168 0.000000
 (between_SS / total_SS =  89.2 %)

Available components:

[1] "cluster"      "centers"      "totss"
[4] "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

L'oggetto restituito dal metodo del k-means contiene diversi dati:

- I centri dei cluster associati a ogni oggetto;
- L'associazione degli oggetti ai cluster;
- Le somme di quadrati interne ai cluster;
- Il rapporto tra la somma di quadrati fra i cluster e la somma di quadrati totale.

Le somme di quadrati indicano la bonta' dei cluster. Come detto anche durante l'analisi delle misure di non omogeneita', l'obiettivo e' fare in modo che le somme dei quadrati interne ai cluster abbiano valori bassi, mentre la somma di quadrati fra i cluster sia il piu' alta possibile. Queste caratteristiche significano che i cluster sono molto compatti e distanti fra loro, di conseguenza risultano ben separati. Il rapporto fra la somma di quadrati fra cluster e la somma di quadrati totale ci fornisce una misura della divisione effettuata: nel nostro caso in esame abbiamo come valore del rapporto 89.2%. Questo risultato ci dice che la distanza tra cluster e' molto maggiore rispetto alla distanza all'interno dei cluster. La divisione in cluster risulta essere molto buona. Il codice `R` che segue permette di stampare un grafico che mostra la divisione delle regioni su piani cartesiani che hanno come assi i gruppi di corsi di laurea.

```
> nomi[1]<- "licenza\nelementare"
> splom(datiScalati,groups=km$cluster,pch=21, pscales=0,
...       panel=function(x, y,i,j,groups, ...) {
...           panel.points(x, y, pch=21,col=groups)
...           panel.points(km$center[,j],km$center[,i],
...                        pch=10)
...       },varnames=nomi,auto.key=FALSE)
```

Il grafico in figura 3.13 ci mostra che, scegliendo una qualsiasi coppia di caratteristiche, si ottiene quasi sempre una netta separazione dei punti, tanto che e' possibile effettuarla tramite dei confini di decisione. Capitano casi particolari in cui vi e' un leggero overlap di punti, ma questo puo' dipendere anche dalla natura casuale di inizializzazione dell'algoritmo del k-means.

3.4 Conclusioni

Terminiamo l'analisi commentando i risultati ottenuti. Possiamo affermare infatti che la regione predominante in ogni titolo di studio risulta la Lombardia, seguita dalle regioni Campania e Lazio. Solo nel caso del diploma di 3 anni il secondo posto e' dato alla regione Veneto. La totale predominazione della Lombardia probabilmente e' data dalla grande differenza nel numero di popolazione appartenente alla regione stessa.

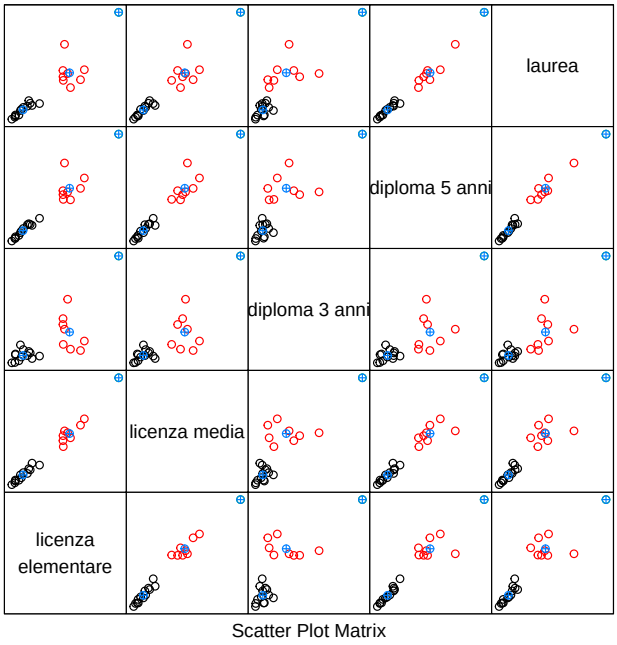

Figura 3.13: Scatterplot k-means 1

Appendici

Appendice A

Variabili aleatorie in R

R offre diversi strumenti per lavorare con le varaibili aleatorie, sia discrete che continue. Per ciascuna delle distribuzioni di probabilita' principali, R fornisce:

- la funzione di probabilita' ;
- la funzione di distribuzione;
- le funzioni quantili;
- un generatore di numeri pseudocasuali che simula la distribuzione.

Tra le distribuzioni, quella piu' interessante sotto il punto dell'analisi, risulta essere la **distribuzione normale**. Per questo motivo in quest'appendice mostreremo alcune operazioni effettuabili tramite variabili aleatorie in R, applicandole alla **distribuzione normale**.

A.1 La distribuzione normale

La funzione di distribuzione normale, chiamata anche di Gauss o Gaussiana e' una delle piu' importanti funzioni di distribuzione, ed e' anche quella pi'u usata. Essa riveste un ruolo importante anche nella statistica. La funzione di distribuzione Gaussiana si basa su due parametri: la media e la varianza. Se indichiamo con μ la media e con σ^2 la varianza, allora possiamo scrivere la funzione di distribuzione nel seguente modo:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad x \in \mathbb{R}, \quad \mu \in \mathbb{R}, \quad \sigma > 0$$

le funzioni R associate a questa distribuzione sono

dnorm: calcola la densita';

pnorm: calcola la funzione di distribuzione;

qnorm: Calcola la funzione quantile;

rnorm: genera numeri pseudo-casuali secondo questa distribuzione.

Utilizzeremo queste funzioni e ne osservermo i risultati.

A.1.1 Densita' di probabilita'

```
> a=-10
> b=10
> curve(dnorm(x, mean=-5, sd=1), from=a, to=b, col="steelblue",xlab="", ylab="",
...      main = expression(paste(mu, "= -5, -1, 1, 5")))
> curve(dnorm(x, mean=-1, sd=1), from=a, to=b, col="tomato", add=TRUE)
> curve(dnorm(x, mean=1, sd=1), from=a, to=b, col="yellowgreen", add=TRUE)
> curve(dnorm(x, mean=5, sd=1), from=a, to=b, col="tan1", add=TRUE)
```

Il codice produce il grafico di figura A.1. Vediamo che cambiare il parametro μ ha l'effetto di spostare la curva in senso orizzontale sull'asse delle ascisse.

```
> a=-10
> b=10
> curve(dnorm(x, mean=0, sd=1), from=a, to=b, col="steelblue", ylab="",
...      xlab="", main=expression(paste(sigma, "= 1, 1.5, 2, 2.5, 3")))
> curve(dnorm(x, mean=0, sd=1.5), from=a, to=b, col="tomato", add=TRUE)
> curve(dnorm(x, mean=0, sd=2), from=a, to=b, col="yellowgreen", add=TRUE)
> curve(dnorm(x, mean=0, sd=2.5), from=a, to=b, col="tan1", add=TRUE)
> curve(dnorm(x, mean=0, sd=3), from=a, to=b, col="blue", add=TRUE)
```

Il codice produce il plot visibile in figura A.2. Notiamo che aumentare il parametro σ ha l'effetto di appiattire e allargare la curva.

A.1.2 Funzione di distribuzione

```
> a=-10
> b=10
> curve(pnorm(x, mean=-5, sd=1), from=a, to=b, col="steelblue", ylab="",
...      xlab="", main = expression(paste(mu, "= -5, -1, 1, 5")))
> curve(pnorm(x, mean=-1, sd=1), from=a, to=b, col="tomato", add=TRUE)
> curve(pnorm(x, mean=1, sd=1), from=a, to=b, col="yellowgreen", add=TRUE)
> curve(pnorm(x, mean=5, sd=1), from=a, to=b, col="tan1", add=TRUE)
```

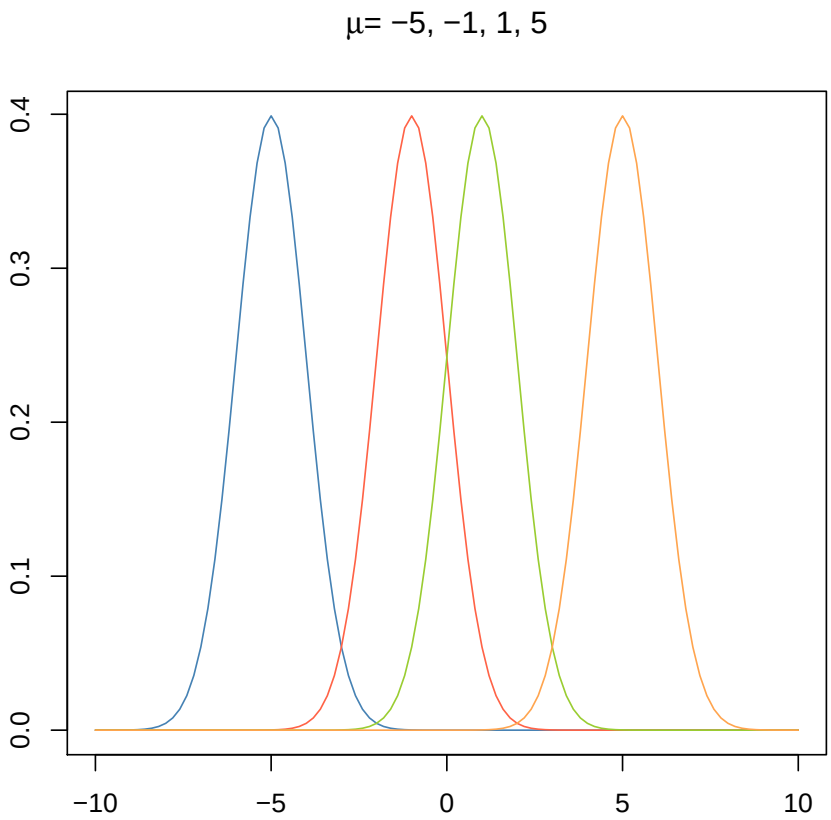

Figura A.1: densita' normali con diversi valori di μ

```
> a=-10
> b=10
> curve(pnorm(x, mean=0, sd=1), from=a, to=b, col="steelblue", ylab="",
...       xlab="", main=expression(paste(sigma, "= 1, 1.5, 2, 2.5, 3")))
> curve(pnorm(x, mean=0, sd=1.5), from=a, to=b, col="tomato", add=TRUE)
> curve(pnorm(x, mean=0, sd=2), from=a, to=b, col="yellowgreen", add=TRUE)
> curve(pnorm(x, mean=0, sd=2.5), from=a, to=b, col="tan1", add=TRUE)
> curve(pnorm(x, mean=0, sd=3), from=a, to=b, col="blue", add=TRUE)
```

A.1.3 Calcolo dei quantili

La funzione `qnorm` permette di calcolare i quantili della distribuzione normale. I quantili suddividono la popolazione in parti uguali, fornendo i valori

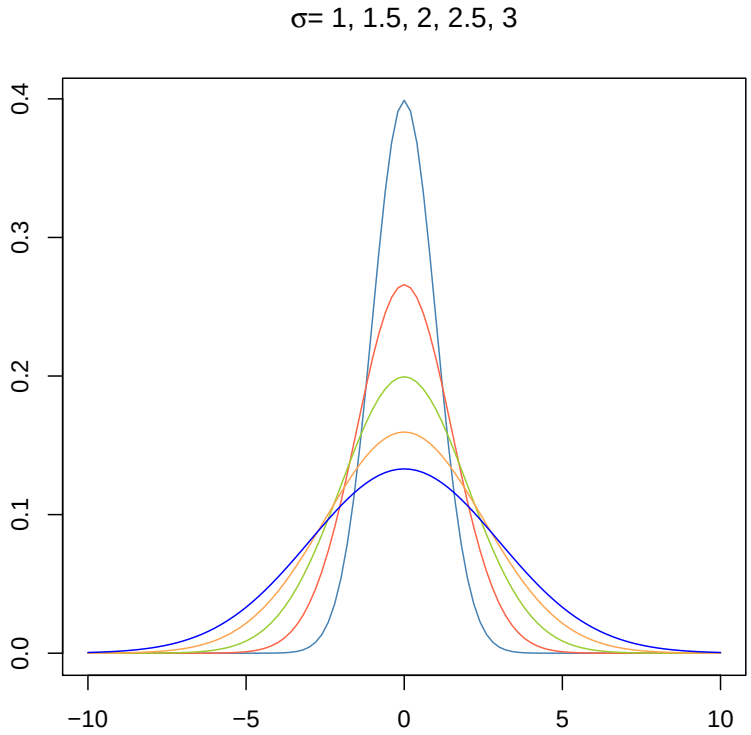

Figura A.2: Densita' normali con diversi valori di σ

che fanno da confine fra una parte e la successiva. I terzili ad esempio, dividono la popolazione in tre parti, i quintili in 5, i ventili in venti, ecc.
Di particolare interesse sono i **quartili** e i **percentili**. I quartili sono un caso particolare dei percentili e si ottengono dividendo l'insieme dei dati ordinati in quattro parti uguali. Il primo quartile Q_1 e' un valore tale che il 25% dei dati ordinati e' minore o uguale di Q_1 ed e' detto anche 25-esimo percentile. Il secondo quartile Q_2 e' un valore tale che il 50% dei dati ordinati e' minore o uguale di Q_2 ed e' detto anche 50-esimo percentile. Il secondo quartile Q_2 coincide con la mediana. Il terzo quartile Q_3 e' un valore tale che il 75% dei dati ordinati e' minore o uguale a Q_3 ed e' detto anche 75-esimo percentile.

L'output della funzione `qnorm` e' z, il 100-esimo percentile. Facciamo un esempio in R con una normale con $\mu = 5$ e $\sigma = 1.5$:

```
> soglie <- c(0, .25, .5, .75, 1)
> qnorm(soglie, mean=5, sd=1.5)
[1]      -Inf 3.988265 5.000000 6.011735       Inf
```

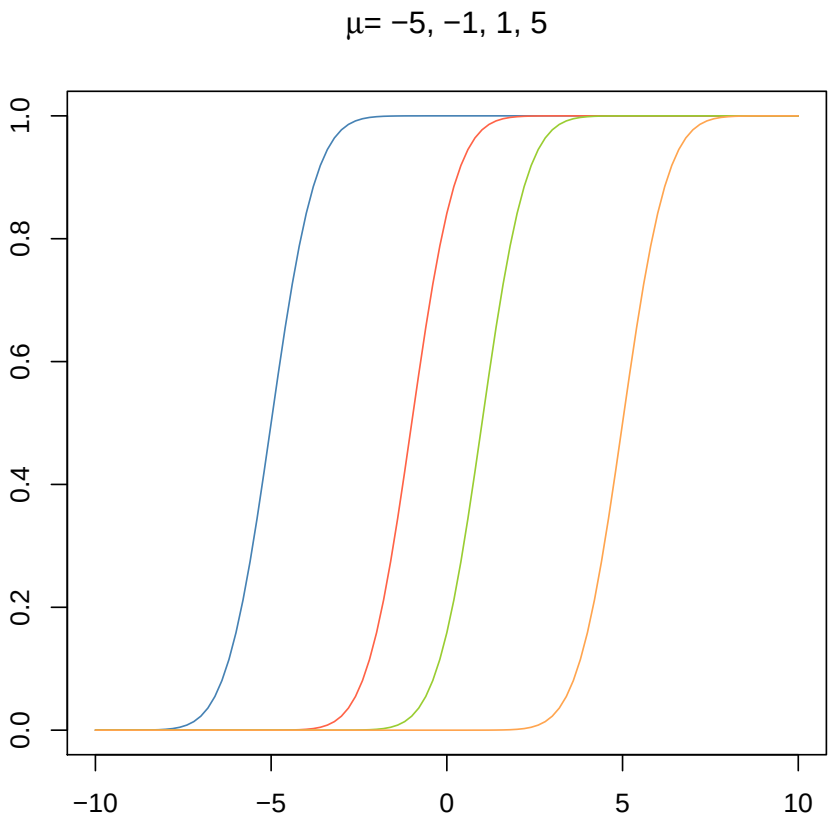

Figura A.3: Distribuzioni normali con diversi valori di μ

A.1.4 Simulazione

La simulazione e' quel processo che tende a ricreare artificialmente un evento reale, sulla base di alcuni parametri di regolarizzazione. Sfruttiamo la simulazione di una variabile aleatoria normale con lo scopo di creare delle popolazioni di taglia diversa. Utlizzeremo, poi, queste popolazioni per due scopi: per confrontare la distribuzione simulata con quella teorica, e per calcolare gli **intervalli di confidenza** nella sezione successiva.

Popolazione di 1000 elementi

Utilizziamo la funzione `rnorm` per creare una popolazione di 1000 individui.

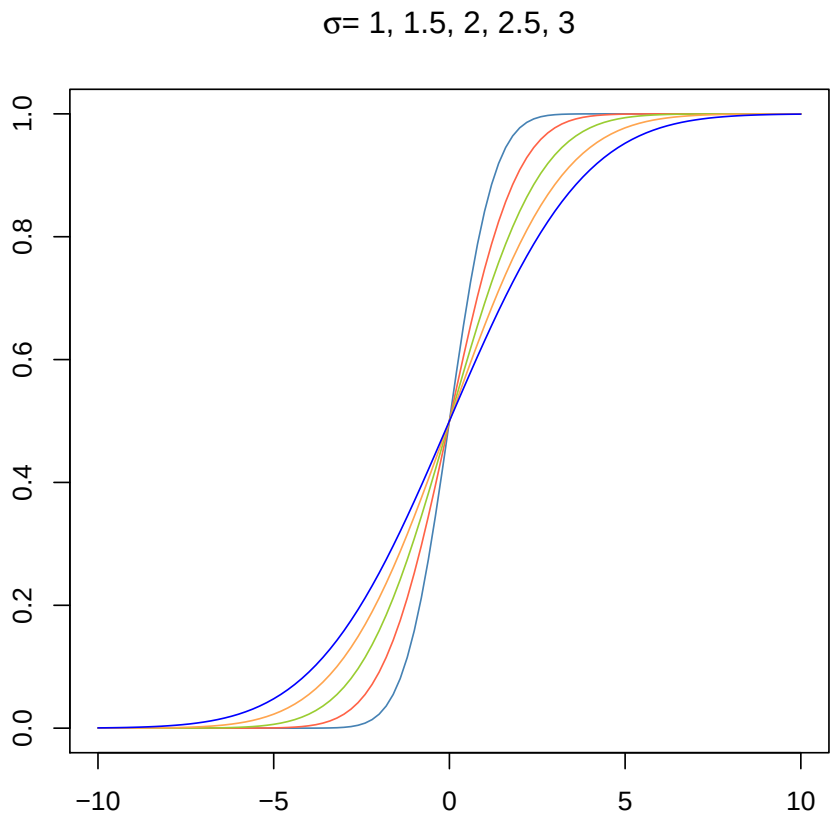

Figura A.4: Distribuzioni normali con diversi valori di σ

```
> popolazione1 <- rnorm(1000, mean=36, sd=6)
```

Osserviamo il confronto fra la distribuzione teorica e quella simulata

```
> par(mfrow = c(1,2))
> curve(dnorm(x, mean=36, sd=6), from=0, to=84,
...       xlab="", ylab="", main="Teorica", las=1)
> hist(popolazione1, breaks=50, xlab="", ylab="",
...      main = "Simulata", las=1)
```

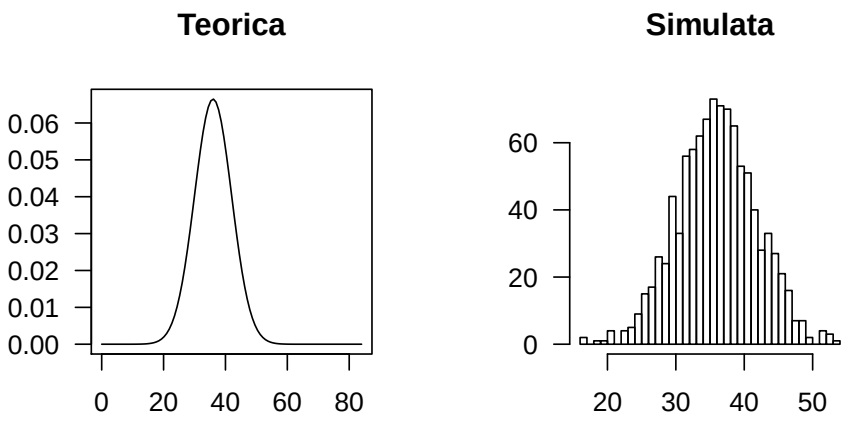

Figura A.5: Confronto fra la densita' della variabile aleatoria teorica e simulata, $\mu = 36$, $\sigma = 6$, 1000 individui

Popolazione di 50000 elementi

```
> popolazione2 <- rnorm(50000, mean=36, sd=6)
```

Osserviamo il confronto fra la distribuzione teorica e quella simulata

```
> par(mfrow = c(1,2))
> curve(dnorm(x, mean=36, sd=6), from=0, to=84,
...      xlab="", ylab="", main="Teorica", las=1)
> hist(popolazione2, breaks=50, xlab="", ylab="",
...     main = "Simulata", las=1)
```

Si nota subito che aumentando la numerosita' del campione la distribuzione simulata si avvicina a quella teorica.

A.2 Intervalli di confidenza

Alla stima puntuale di un parametro non noto di una popolazione (singolo valore reale) si preferisce spesso calcolare un intervallo di valori, detto **intervallo di confidenza**,ossia si cerca di determinare in base ai dati del campione, due limiti (uno inferiore ed uno superiore) entro i quali sia compreso il parametro non noto con un certo coefficiente di confidenza (detto anche grado di fiducia o **grado di confidenza**). Sia $X_1, X_2, \ldots, X_n$. un campione casuale di ampiezza n estratto da una popolazione con funzione

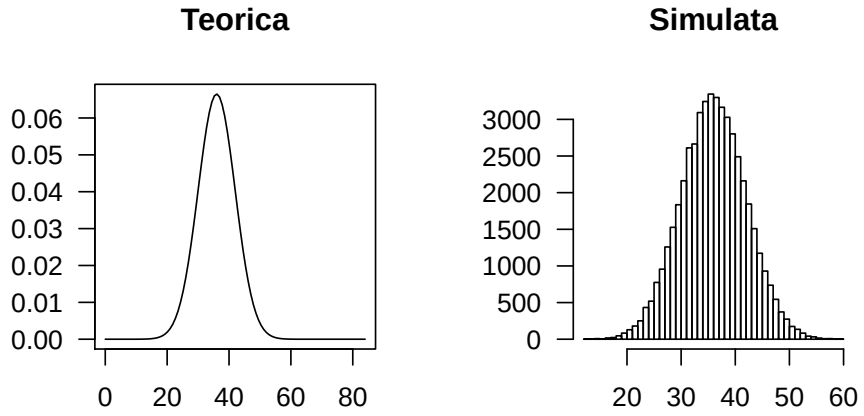

Figura A.6: Confronto fra la densita' della variabile aleatoria teorica e simulata, $\mu = 36$, $\sigma = 6$, 50000 individui

di probabilita' (densita' nel caso assolutamente continuo) $f(x, \theta)$, con θ il parametro da stimare, vogliamo trovare due valori $\underline{C}_n$ e $\overline{C}_n$ tali che

$$P(\,\underline{C}_n < \theta < \overline{C}_n\,) < 1 - \alpha$$

dove

- $\underline{C}_n = g_1(X_1, X_2, \ldots, X_n)$ e $\overline{C}_n = g_2(X_1, X_2, \ldots, X_n)$ sono due statistiche che soddisfano la condizone $\underline{C}_n < \overline{C}_n$ e sono chiamate rispettivamente limite inferiore e superiore dell'intervallo di confidenza;

- $1 - \alpha,$ con $(0 < \alpha < 1)$, e' detto **coefficiente di confidenza**.

In generale esistono numerosi intervalli di confidenza dello stesso grado 1???α per un parametro non noto θ della popolazione.
Un metodo per la costruzione di tali intervalli e' il **metodo pivotale**. Questo metodo consiste nel creare una variabile aleatoria di pivot $\gamma(X_1, X_2, \ldots, X_n; \theta)$ che dipende dal campione casuale e dal parametro non noto θ: la funzione di distribuzione di questa variabile non contiene il parametro da stimare.
Passiamo di seguito a calcolare l'intervallo di confidenza dei parametri di una Gaussiana.

A.2.1 Intervallo di confidenza per μ con σ^2 nota

Vogliamo stimare la media μ di una popolazione distribuita in modo normale con varianza σ^2. Fissiamo

$$1 - \alpha = 0.95$$
$$n = 50$$

Utilizziamo il metodo pivotale, considerando la variabile aleatoria di pivot

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

che e' distribuita normalmente con media nulla e varianza unitaria. Scrivendo

$$P\left[\overline{X}_n - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

ricaviamo l'intervallo di confdenza di grado $1 - \alpha$ per μ:

$$\left[\overline{X}_n - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}};\ \overline{X}_n + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Utilizzando R, possiamo estrarre un campione della popolazione con le seguenti istruzioni:

```
> n <- 50
> x <- sample(popolazione1, n, replace=TRUE)
```

e quindi calcoliamo l'intervallo di confidenza

```
> alpha <- 0.05
> media <- mean(x)
> z <- qnorm(1-alpha/2, mean=0, sd=1)
> b <- sd(popolazione1)/sqrt(n)
> low <- media - z * b
> high <- media + z * b
```

Otteniamo:
Intervallo di confidenza: $[34.991,\ 38.185]$;
Lunghezza dell'intervallo $= \overline{C}_n - \underline{C}_n = 3.194$;
z=1.96.
I comandi che seguono producono il grafico in figura A.7 che ci permette di visualizzare le regioni di accettazione e rifiuto:

```
> par(mar = c(2, .5, 2, .5))
> curve(dnorm(x,mean=0,sd=1),from=-3, to=3, axes=FALSE, xlab="", ylab="",
...main=expression(paste("Densita  normale intervallo di confidenza 1-",
...                      alpha,"=0.95")),
```

```
...ylim = c(0, .5))
> text(0,0.05, expression (1-alpha))
> text(0,0.1,"Regione di accettazione")
> axis(1,c(-4,-z,0,z,4), c("",abbreviate(-z,5),0,abbreviate(z,4),""))
> vals<-seq(-3,-z, length =100)
> x1<-c(-4,vals ,-z,-4)
> y1<-c(0,dnorm(vals),0,0)
> polygon (x1,y1,density=20,angle=-45)
> vals<-seq(z,4, length=100)
> x1<-c(z,vals ,4,1)
> y1<-c(0,dnorm(vals),0,0)
> polygon (x1,y1,density =20,angle=45)
> abline(h=0)
> text(-2.45,0.07, expression (alpha/2))
> text(-2.45,0.12,"Regione di rifiuto ")
> text(2.45,0.07, expression (alpha/2))
> text(2.45,0.12,"Regione di rifiuto ")
> box()
```

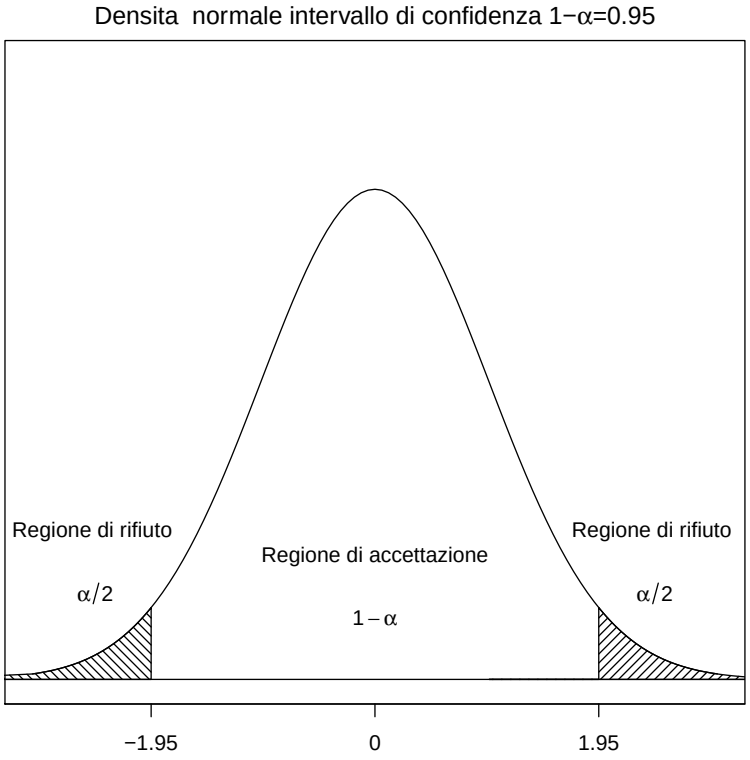

Figura A.7: Intervalli di confidenza per μ - σ nota: $1 - \alpha = 0.95$

A.2.2 Intervallo di confidenza per μ con σ^2 non nota

Per determinare un intervallo di confidenza di grado $1 - \alpha$ per la media non conoscendo la varianza della popolazione normale in esame, utilizzando il metodo pivotale consideriamo la seguente variabile aleatoria:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

che e' distribuita con legge di Student con $n - 1$ gradi di liberta' . Scrivendo

$$P\left[\overline{X} - t_{\alpha/2,\,n-1}\,\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2,\,n-1}\,\frac{S}{\sqrt{n}}\right] = 1 - \alpha$$

diciamo che nel $100(1 - \alpha)\%$ dei campioni la media μ e' compresa fra gli estremi

$$\left[\overline{X} - t_{\alpha/2,\,n-1}\,\frac{S}{\sqrt{n}}\,;\ \overline{X} + t_{\alpha/2,\,n-1}\,\frac{S}{\sqrt{n}}\right]$$

Calcoliamo tali estremi in R:

```
> n <- 50
> x <- sample(popolazione1, n, replace=TRUE)
> alpha <- 0.05
> media <- mean(x)
> ta <- qt(1-alpha/2, df=n-1)
> b <- sd(x)/sqrt(n)
> low <- media - ta * b
> high <- media + ta * b
```

Otteniamo:
Intervallo di confidenza: $[\,33.561,\ 36.288\,]$
Lunghezza dell'intervallo $= \overline{C}_n - \underline{C}_n = 2.727$
$t_{\alpha/2,\,n-1}$=2.01

I comandi che seguono producono il grafico in figura A.8 che ci permette di visualizzare le regioni di accettazione e rifiuto:

```
> par(mar = c(2, .5, 2, .5))
> curve(dt(x,df=(n-1)),from=-4, to=4, axes=FALSE, xlab="", ylab="",
...main=expression(paste("Densita Student intervallo di confidenza 1-",alpha,"=(
> text(0,0.05, expression (1-alpha))
> text(0,0.1,"Regione di accettazione")
> axis(1,c(-4,-ta,0,ta,4), c("",abbreviate(-ta,5),0,abbreviate(ta,4),""))
> vals<-seq(-3,-ta, length =100)
> x1<-c(-4,vals ,-ta,-4)
> y1<-c(0,dt(vals, n-1),0,0)
> polygon (x1,y1,density=20,angle=-45)
> vals<-seq(ta,4, length=100)
> x1<-c(ta,vals ,4,1)
> y1<-c(0,dt(vals,n-1),0,0)
> polygon (x1,y1,density =20,angle=45)
> abline(h=0)
> text(-3,0.07, expression (alpha/2))
> text(-3,0.12,"Regione di rifiuto ")
> text(3,0.07, expression (alpha/2))
> text(3,0.12,"Regione di rifiuto ")
> box()
```

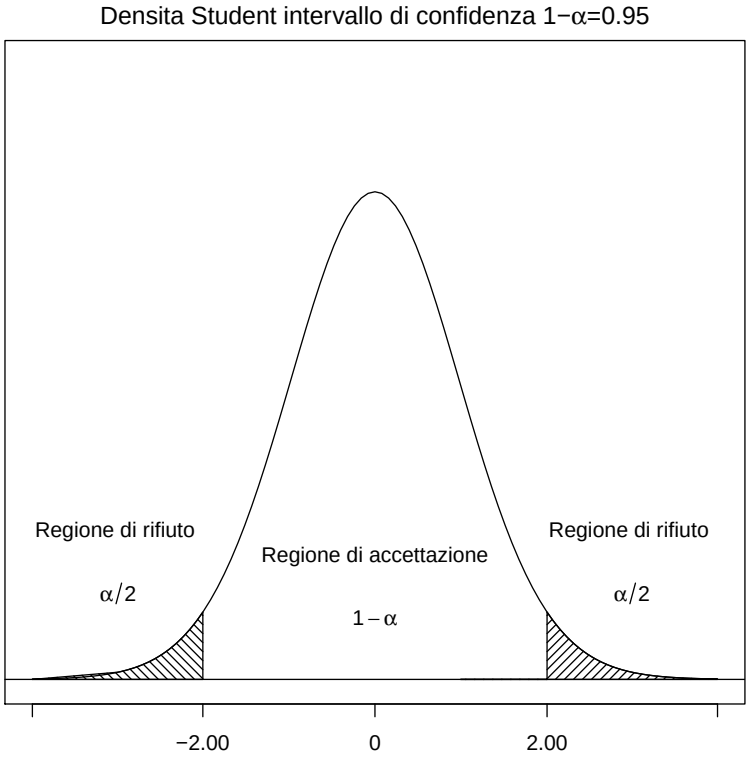

Figura A.8: Intervalli di confidenza per μ - σ non nota: $1 - \alpha = 0.95$

A.2.3 Intervallo di confidenza per σ^2 con μ nota

Passiamo ora al calcolo dell'intervallo di confidenza per il parametro σ, partendo dal caso in cui il parametro μ e' fissato. Utilizziamo il metodo pivotale considerando la seguente varaibile aleatoria di pivot:

$$V_n = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{(n-1)S_n^2}{\sigma^2} + \left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \right)^2$$

che, in quanto somma di quadrati di variabili aleatorie normali standard, ha distribuzione con legge chi-quadrato con n gradi di liberta' . Scriviamo

$$P\left(\chi^2_{1-\alpha/2,\,n} < V_n < \chi^2_{\alpha/2,\,n} \right) = 1 - \alpha$$

sostituendo V_n abbiamo

$$P\left(\chi^2_{1-\alpha/2,\,n} < \frac{(n-1)S_n^2}{\sigma^2} + \left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \right)^2 < \chi^2_{\alpha/2,\,n} \right) = 1 - \alpha$$

che risolta rispetto a σ^2 ci da

$$P\left(\frac{(n-1)S_n^2 + n(\overline{X}_n - \mu)^2}{\chi^2_{\alpha/2,\,n}} < \sigma^2 < \frac{(n-1)S_n^2 + n(\overline{X}_n - \mu)^2}{\chi^2_{1-\alpha/2,\,n}}\right) = 1 - \alpha$$

Abbiamo quindi i valori per i due estremi dell'intervallo di confidenza:

$$\underline{C}_n = \frac{(n-1)S_n^2 + n(\overline{X}_n - \mu)^2}{\chi^2_{\alpha/2,\,n}}$$
$$\overline{C}_n = \frac{(n-1)S_n^2 + n(\overline{X}_n - \mu)^2}{\chi^2_{1-\alpha/2,\,n}}$$

Ora che sappiamo come fare, calcoliamo l'intervallo di confidenza utilizzando R:

```
> n <- 50
> x <- sample(popolazione1, n, replace=TRUE)
> alpha <- 0.05
> mediaCamp <- mean(x)
> mediaPop <- mean(popolazione1)
> varianza <- var(x)
> chiLow <- qchisq(alpha/2, df=n)
> chiHigh <- qchisq(1-alpha/2, df=n)
> numeratore <- (n-1) * varianza + n*( (mediaCamp - mediaPop)**2)
> low <- numeratore / chiHigh
> high <- numeratore / chiLow
```

Otteniamo:
Intervallo di confidenza: $[\,56.053,\ 25.395\,]$
Lunghezza dell'intervallo $= \overline{C}_n - \underline{C}_n = 30.658$
$\chi^2_{1-\alpha/2,\,n} = 71.42$
$\chi^2_{\alpha/2,\,n} = 32.357$

Il seguente codice genera il grafico mostrato in figura A.9

```
> a = chiLow - chiLow/10
> b = chiHigh + chiHigh/10
> par(mar = c(2, .5, 2, .5))
> curve(dchisq(x,df=n),from=a, to=b, axes=FALSE, xlab="", ylab="",
...main=expression(paste("Densita' ",{chi^2} ,
...                      " intervallo di confidenza 1-",alpha,"=0.95")),
...ylim = c(0, .1))
> text(0,0.05, expression (1-alpha))
> text(0,0.1,"Regione di accettazione")
```

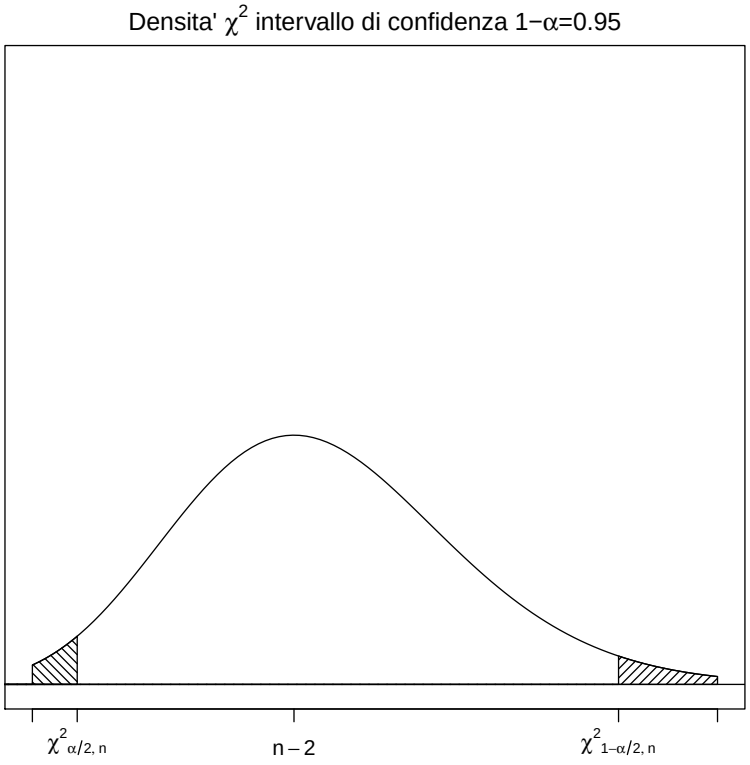

Figura A.9: Intervalli di confidenza per σ - μ nota: $1 - \alpha = 0.95$

```
> axis(1,c(a,chiLow,n-2,chiHigh,b), c("",expression({chi^2}[list(alpha/2,n)]),
...  expression(n-2),expression({chi^2}[list(1-alpha/2,n)]),""))
> vals<-seq(a,chiLow, length =100)
> x1<-c(a,vals ,chiLow,0)
> y1<-c(0,dchisq(vals,n),0,0)
> polygon (x1,y1,density=20,angle=-45)
> vals<-seq(chiHigh,b, length=100)
> x1<-c(chiHigh,vals ,b,1)
> y1<-c(0,dchisq(vals,n),0,0)
> polygon (x1,y1,density =20,angle=45)
> abline(h=0)
> text(-2.45,0.07, expression (alpha/2))
> text(-2.45,0.12,"Regione di rifiuto ")
> text(2.45,0.07, expression (alpha/2))
> text(2.45,0.12,"Regione di rifiuto ")
> box()
```

A.2.4 Intervallo di confidenza per σ^2 con μ non nota

Per calcolare l'intervallo di confidenza di grado $1-\alpha$ per il parametro σ^2 non conoscendo la media, utilizziamo il metodo del pivot considerando la seguente variabile aleatoria pivot:

$$Q_n = \frac{(n-1)S_n^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

che, ancora una volta, e' distribuita con legge chi-quadrato con $n-1$ gradi di liberta' . Poniamo

$$P\left(\chi^2_{1-\alpha/2,n-1} < Q_n < \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$$

ovvero

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

che risolvendo rispetto a σ^2:

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

ci da gli estremi dell'intervallo di confidenza di grado $1-\alpha$:

$$\underline{C}_n = \frac{(n-1)S_n^2}{\chi^2_{1-\alpha/2,n-1}} \qquad \overline{C}_n = \frac{(n-1)S_n^2}{\chi^2_{\alpha/2,n-1}}$$

Utilizziamo le formule per calcolare gli estremi:

```
> n <- 50
> x <- sample(popolazione1, n, replace=TRUE)
> alpha <- 0.05
> media <- mean(x)
> varianza <- var(x)
> chiLow <- qchisq(alpha/2, df=n-1)
> chiHigh <- qchisq(1-alpha/2, df=n-1)
> low <- ((n-1) * varianza ) / chiLow
> high <- ((n-1) * varianza) / chiHigh
```

Otteniamo:
Intervallo di confidenza: $[\,27.132,\ 60.38\,]$
Lunghezza dell'intervallo $= \overline{C}_n - \underline{C}_n = 33.248$
$\chi^2_{1-\alpha/2,\,n} = 70.222$
$\chi^2_{\alpha/2,\,n} = 31.555$

Il seguente codice genera il grafico mostrato in figura A.10

```
> a = chiLow - chiLow/10
> b = chiHigh + chiHigh/10
> par(mar = c(2, .5, 2, .5))
> curve(dchisq(x,df=n),from=a, to=b, axes=FALSE, xlab="", ylab="",
...main=expression(paste("Densit&agrave  ",{chi^2} ,
...                      " intervallo di confidenza 1-",alpha,"=0.95")),
...ylim = c(0, .1))
> text(0,0.05, expression (1-alpha))
> text(0,0.1,"Regione di accettazione")
> axis(1,c(a,chiLow,n-2,chiHigh,b), c("",expression({chi^2}[list(alpha/2,n)]),
...  expression(n-2),expression({chi^2}[list(1-alpha/2,n)]),""))
> vals<-seq(a,chiLow, length =100)
> x1<-c(a,vals ,chiLow,0)
> y1<-c(0,dchisq(vals,n),0,0)
> polygon (x1,y1,density=20,angle=-45)
> vals<-seq(chiHigh,b, length=100)
> x1<-c(chiHigh,vals ,b,1)
> y1<-c(0,dchisq(vals,n),0,0)
> polygon (x1,y1,density =20,angle=45)
> abline(h=0)
> text(-2.45,0.07, expression (alpha/2))
> text(-2.45,0.12,"Regione di rifiuto ")
> text(2.45,0.07, expression (alpha/2))
> text(2.45,0.12,"Regione di rifiuto ")
> box()
```

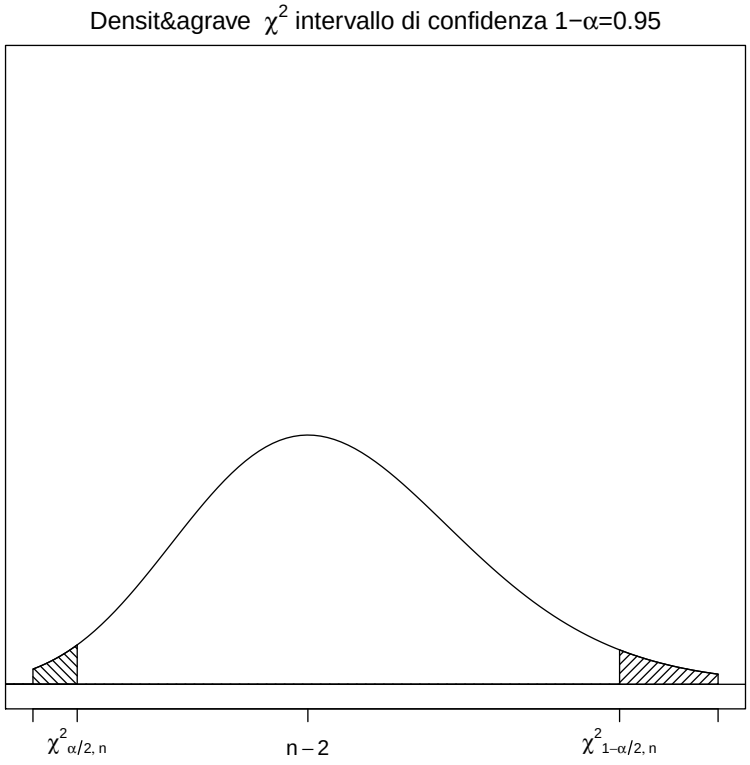

Figura A.10: Intervalli di confidenza per σ - μ non nota: $1 - \alpha = 0.95$

Elenco delle figure