Università degli Studi di Salerno

Dipartimento di Informatica
Corso di Laurea Triennale in Informatica

Progetto Calcolo Probabilità Statistica Matematica
(CPSM)

Indagine Statistica sulle Morti in incidenti stradali

Tozza Gennaro Carmine
Matricola: 0512120382

Anno Accademico 2024-2025

Indice

Capitolo 1

Introduzione

1.1 Problematica

Gli **incidenti stradali** costituiscono una delle principali emergenze di sanità pubblica, in quanto responsabili ogni anno di un elevato numero di decessi, in particolare tra i giovani, oltre al drammatico impatto umano e psicologico sulle loro famiglie.

1.2 Scopo del progetto

Il progetto consiste nel realizzare un'indagine statistica [1]sugli incidenti stradali verificatisi sulla rete stradale del territorio nazionale, tra il 2010 e il 2023 verbalizzati da un'autorità di Polizia o dai Carabinieri, avvenuti su una strada aperta alla circolazione pubblica e che hanno causato morti (entro il 30° giorno) con il coinvolgimento di almeno un veicolo.

La rilevazione è condotta correntemente dall'Istat, con la compartecipazione dell'ACI (Automobile Club d'Italia) e di numerosi Enti pubblici istituzionali.

Per l'analisi dei dati è stato scelto l'ambiente di calcolo statistico **R**.
R fornisce un'ampia varietà di tecniche statistiche (modellazione lineare e non lineare, test statistici classici, analisi delle serie temporali, classificazione, ...) e grafiche ed è altamente estensibile.

Uno dei punti di forza di R è la facilità con cui possono essere prodotti grafici ben progettati e di qualità per la pubblicazione, compresi simboli matematici e formule se necessario.

[1] https://siqual.istat.it/SIQual/visualizza.do?id=7777778&refresh=true&language=IT

Per semplificare l'analisi, si è scelto di lavorare non sull'intero dataset, ma su un sottoinsieme filtrato di dati, relativo alle **morti per incidenti stradali che riguardano solo i conducenti di età compresa tra i 21 e i 24 anni.**

Per approfondire l'analisi con dati dettagliati e specifici, è possibile consultare il dataset direttamente sul sito dell'ISTAT al seguente link: https://esploradati.istat.it/databrowser/#/it/dw/categories/IT1,Z0810HEA,1.0/HEA_ROAD/IT1,41_270_DF_DCIS_MORTIFERITISTR1_1,1.0

Il dataset in formato **CSV(Comma-Separated Values)** già filtrato è stato ottenuto dalla fonte ISTAT tramite il link indicato, assicurando così l'affidabilità dei dati.

Il formato scelto (CSV) permette un'agevole manipolazione dei dati, essendo compatibile con la maggior parte dei software statistici e dei fogli di calcolo, ottimizzando l'analisi e la visualizzazione delle informazioni.

Successivamente, utilizzando il software R, sono stati applicati ulteriori criteri di selezione per estrarre esclusivamente i dati rilevanti ai fini dell'analisi.

```r
# inclusione librerie
library("tidyverse")
require("tidyverse")
library("dplyr")
library(moments)

# viene caricato il dataset
dati <- read.csv("dati_istat.csv")

# filtraggio dati
dati <- dati %>% select(Intersezione, TIME_PERIOD,
    Osservazione)
```

Intersezione	**'10**	**'11**	**'12**	**'13**	**'14**	**'15**	**'16**	**'17**	**'18**	**'19**	**'20**	**'21**	**'22**	**'23**
Incrocio	55	45	42	23	34	24	28	25	23	13	15	16	21	10
Rotatoria	7	4	2	5	5	1	5	3	1	2	1	2	–	3
Rettilineo	92	102	89	82	96	95	66	70	60	72	66	69	69	74
Curva	58	51	48	46	45	44	41	36	38	42	23	40	27	38
Dosso/Pend.	2	7	2	4	2	3	5	1	2	3	2	4	3	1
Galleria	1	1	1	1	–	1	1	1	–	–	3	1	–	–
Totale	215	210	184	161	182	168	146	136	124	132	110	132	120	126

Capitolo 2

Tabelle delle frequenze

```r
dati <- dati %>%
  filter(Intersezione != "Totale")

# 1. Frequenze assolute per intersezione dell'incidente
freq_assolute <- aggregate(Osservazione ~ Intersezione, data=dati, sum)
colnames(freq_assolute) <- c("Intersezione", "Frequenza_Assoluta")
print(freq_assolute)

# 2. Frequenze relative per intersezione dell'incidente
totale <- sum(freq_assolute$Frequenza_Assoluta)
freq_assolute$Frequenza_Relativa <- freq_assolute$Frequenza_Assoluta / totale
print(freq_assolute)

# 3. Frequenze cumulate assolute
freq_assolute <- freq_assolute[order(-freq_assolute$Frequenza_Assoluta),]
freq_assolute$Frequenza_Cumulata_Assoluta <- cumsum(freq_assolute$Frequenza_Assoluta)
print(freq_assolute)

# 4. Frequenze cumulate relative
freq_assolute$Frequenza_Cumulata_Relativa <- freq_assolute$Frequenza_Cumulata_Assoluta / totale
print(freq_assolute)

# Per esportare i risultati
#write.csv(freq_assolute, "tabella_frequeza.csv", row.names = FALSE)
```

L'output generato dal codice R riportato è sintetizzato nella seguente tabella:

Tipo intersezione	**Freq. Assoluta**	**Freq. Relativa**	**Freq. Cum. Assoluta**	**Freq. Cum. Relativa**
Rettilineo	1102	0.5135	1102	0.5135
Curva	577	0.2689	1679	0.7824
Incrocio	374	0.1743	2053	0.9567
Dosso/Pendenza/Strettoia	41	0.0191	2094	0.9758
Rotatoria	41	0.0191	2135	0.9949
Galleria	11	0.0051	2146	1.0000

L'analisi di frequenza, fornisce una chiara mappatura della distribuzione degli incidenti stradali analizzati, per un totale di 2.146 casi.

Dai dati emerge in modo netto che la tipologia di strada più a rischio è il rettilineo, che da solo rappresenta oltre la metà degli incidenti totali (1.102 casi, pari al 51,4%). Questo dato è di particolare rilevanza, in quanto concentra la maggior parte del rischio in un contesto spesso percepito come meno pericoloso.

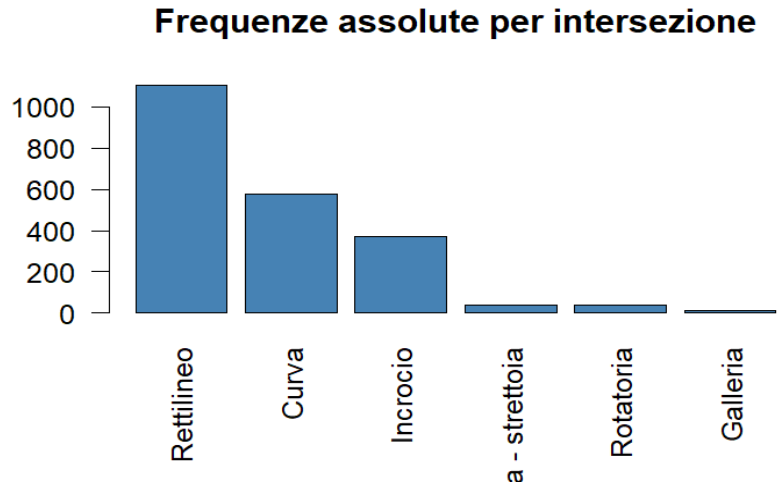
L'analisi cumulativa rafforza questa evidenza. Le prime due categorie, Rettilineo e Curva, coprono complessivamente il 78,2% di tutti gli incidenti. Aggiungendo gli incroci, si arriva a spiegare oltre il 95% del fenomeno (95,7%).

Le restanti tipologie di infrastruttura (dosso/pendenza/strettoia, rotatoria, galleria) hanno un'incidenza marginale, contribuendo complessivamente per meno del 5% al totale degli eventi.

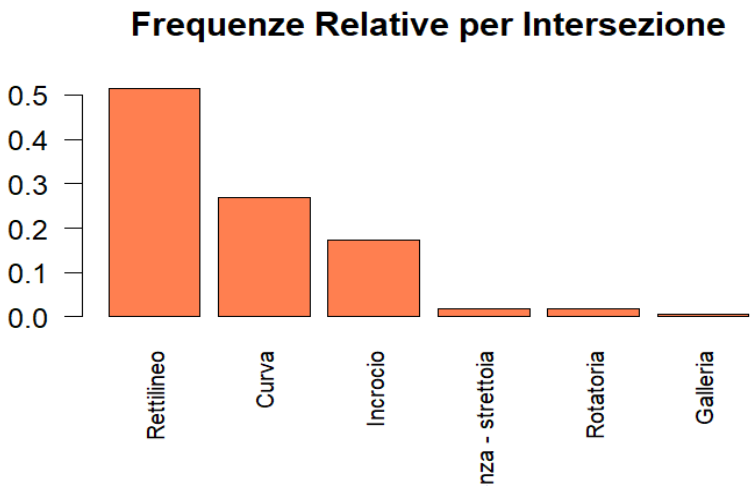
Capitolo 3

Rappresentazione dei dati mediante grafici

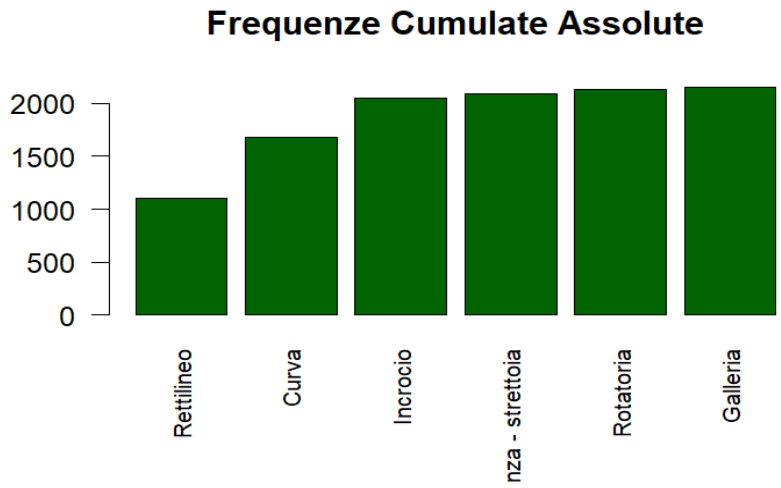
```r
# 1. Grafico frequenza assolute
barplot(freq_assolute$Frequenza_Assoluta,
        names.arg = freq_assolute$Intersezione,
        main = "Frequenze assolute per intersezione",
        xlab = "",
        ylab = "",
        col = "steelblue",
        las = 2) # Etichette verticali
```



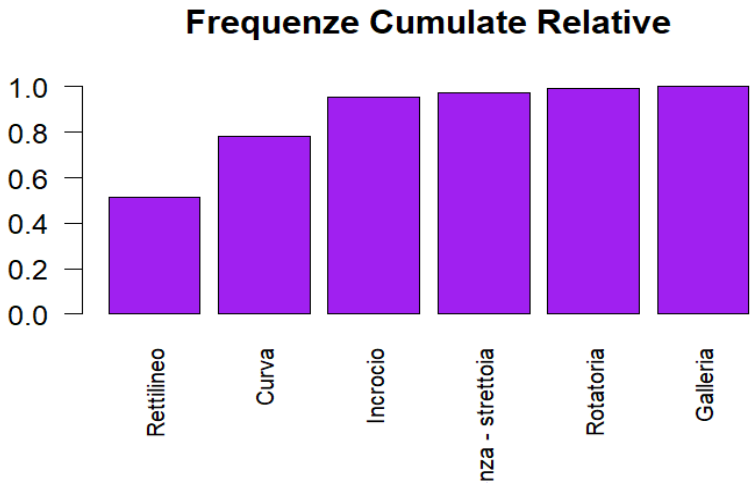
```r
# 2. Grafico frequenze relative
barplot(freq_assolute$Frequenza_Relativa,
        names.arg = freq_assolute$Intersezione,
        main = "Frequenze Relative per Intersezione",
        xlab = "",
        ylab = "",
        col = "coral",
        las = 2,
        cex.names = 0.8)
```




```r
# 3. Grafico frequenze cumulate assolute
barplot(freq_assolute$Frequenza_Cumulata_Assoluta,
        names.arg = freq_assolute$Intersezione,
        main = "Frequenze Cumulate Assolute",
        xlab = "",
        ylab = "",
        col = "darkgreen",
        las = 2,
        cex.names = 0.8)
```




```r
# 4. Grafico frequenze cumulate relative
barplot(freq_assolute$Frequenza_Cumulata_Relativa,
        names.arg = freq_assolute$Intersezione,
        main = "Frequenze Cumulate Relative",
        xlab = "",
        ylab = "",
        col = "purple",
        las = 2,
        cex.names = 0.8)
```

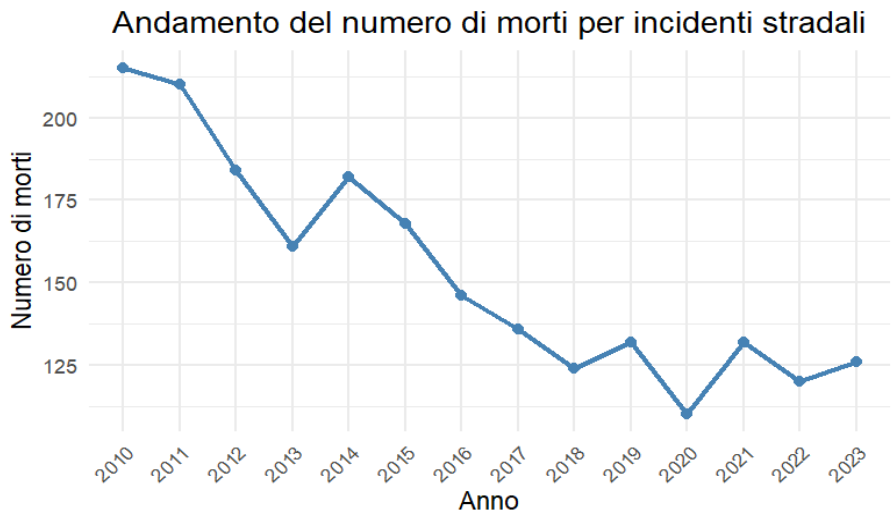


Altri grafici:

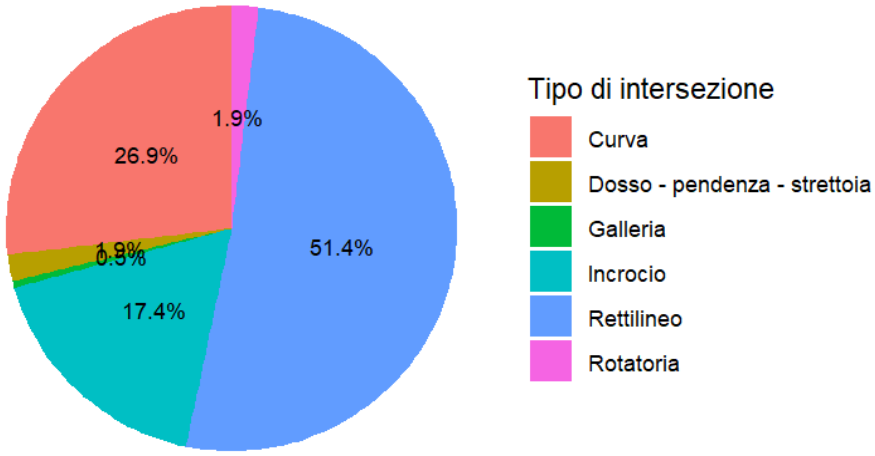
```r
# Converti TIME_PERIOD in fattore per mantenere l'ordine originale
dati$TIME_PERIOD <- factor(dati$TIME_PERIOD, levels = unique(dati$TIME_PERIOD))

# Calcola il numero totale di morti per anno
morti_per_anno <- dati %>%
  group_by(TIME_PERIOD) %>%
  summarise(Totale_Morti = sum(Osservazione, na.rm = TRUE))

# Crea il grafico a linee con tutti gli anni visibili
ggplot(morti_per_anno, aes(x = TIME_PERIOD, y = Totale_Morti, group = 1)) +
  geom_line(color = "steelblue", size = 1) +
  geom_point(color = "steelblue", size = 2) +
  labs(title = "Andamento del numero di morti per incidenti stradali",
       x = "Anno",
       y = "Numero di morti") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45, hjust = 1, size = 8)) +  # Riduci dimensione testo
  scale_x_discrete(breaks = levels(morti_per_anno$TIME_PERIOD))  # Mostra tutti i valori
```



```r
# Crea il grafico a torta
ggplot(freq_assolute, aes(x = "", y = Frequenza_Assoluta,
    fill = Intersezione)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0) +
  labs(title = "",
       fill = "Tipo di intersezione") +
  theme_void() +
  theme(plot.title = element_text(hjust = 0.5, size = 14,
      face = "bold"),
        legend.position = "right") +
  geom_text(aes(label = paste0(round(Frequenza_Assoluta/sum(
      Frequenza_Assoluta)*100, 1), "%")),
            position = position_stack(vjust = 0.5),
            size = 3)
```



Capitolo 4

Indici di posizione

In questo capitolo verranno calcolati gli **indici di posizione** sulla variabile numerica `Osservazione`.

Questa variabile rappresenta il numero di morti registrato per una specifica combinazione di tipo di intersezione e anno, limitatamente alla fascia di età dei conducenti tra 21 e 24 anni.

Gli indici di posizione(o di tendenza generale) sono misure che consentono di sintetizzare i dati osservati con un solo valore numerico che sia rappresentativo dei dati stessi. Gli indici di posizone più adoperati sono tre:

- media
- mediana
- moda

4.1 Media campionaria

La **media campionaria** è la somma di tutte le osservazioni divisa per il numero di esse. Fornisce una misura del valore centrale della distribuzione.

```
1 media_generale <- mean(dati$Osservazione, na.rm = TRUE)
2 print(paste("Media campionaria generale delle osservazioni:"
    , round(media_generale, 2)))
```

Output:

```
"Media campionaria generale delle osservazioni: 27.51"
```

Questo valore indica che, in media, per ogni specifica combinazione di tipo di intersezione e anno considerata nel dataset filtrato, si sono registrati circa 27.51 decessi.

4.2 Mediana campionaria

La **mediana** è un numero che precede tanti dati quanti ne segue, ossia il valore centrale dei dati riordinati in senso crescente.

```
mediana_generale <- median(dati$Osservazione, na.rm = TRUE)
print(paste("Mediana generale delle osservazioni:", mediana_
    generale))
```

Output:

```
"Mediana generale delle osservazioni: 15.5"
```

Il fatto che la mediana (15.5) sia inferiore alla media (27.51) suggerisce una distribuzione asimmetrica a destra.

4.3 Moda campionaria

La **moda** è il valore (o i valori, in caso di distribuzioni multimodali) che appare più frequentemente in un insieme di dati.

```
find_mode <- function(x) {
    u <- unique(x)
    tab <- tabulate(match(x, u))
    u[tab == max(tab)]
  }

find_mode(dati$Osservazione)
```

Output:

```
  1
```

L'output '1', indica che il numero di decessi più frequentemente osservato per una singola combinazione intersezione/anno è 1. Se ci fossero più valori con la stessa frequenza massima, sarebbero elencati tutti.

Capitolo 5

Indici di variabilità

Di seguito calcoliamo gli **indici di variabilità(o di dispersione)**, che descrivono la variabilità dei dati osservati e consentono di valutare l'informazione fornita dall'indice di posizione utilizzato, dando dei dati più accurati.

5.1 Varianza campionaria

La **varianza campionaria** (s^2) misura la dispersione media quadratica dei dati attorno alla media campionaria. È espressa nell'unità di misura dei dati al quadrato.

```
# Calcolo della varianza sulla variabile 'Osservazione'.
varianza_campionaria <- var(dati$Osservazione, na.rm = TRUE)
print(paste("Varianza campionaria delle osservazioni:",
    round(varianza_campionaria, 2)))
```

Output:

```
"Varianza campionaria delle osservazioni: 859.52"
```

Un valore elevato della varianza indica una notevole dispersione dei dati attorno alla media.

5.2 Deviazione standard campionaria

La **deviazione standard campionaria** (s) è la radice quadrata della varianza campionaria. Fornisce una misura della dispersione media dei dati attorno alla media, espressa nella stessa unità di misura dei dati originali, rendendola più interpretabile della varianza.

```r
# Calcolo della deviazione standard sulla variabile 'Osservazione'.
dev_std_campionaria <- sd(dati$Osservazione, na.rm = TRUE)
print(paste("Deviazione standard campionaria delle osservazioni:", round(dev_std_campionaria, 2)))
```

Output (radice quadrata della varianza stimata):

```
"Deviazione standard campionaria delle osservazioni: 29.61"
```

Questo valore indica che, mediamente, i singoli conteggi di decessi si discostano dalla media campionaria (27.51) di circa 29.61 unità.

5.3 Scarto medio assoluto

Lo **scarto medio assoluto** è dato dalla media aritmetica degli scarti assoluti dalla media campionaria. Come la deviazione standard, misura la dispersione media, ma è meno sensibile ai valori anomali perché non eleva al quadrato gli scarti.

```r
# Calcolo dello scarto medio assoluto dalla media per 'Osservazione'.
media_oss <- mean(dati$Osservazione, na.rm = TRUE)
scarto_medio_assoluto <- mean(abs(dati$Osservazione - media_oss), na.rm = TRUE)
print(paste("Scarto medio assoluto (dalla media) delle osservazioni:", round(scarto_medio_assoluto, 2)))
```

Output (stima):

```
"Scarto medio assoluto (dalla media) delle osservazioni: 25.13"
```

In media, le osservazioni si discostano (in valore assoluto) dalla media di circa 25.13 decessi.

5.4 Ampiezza del campo di variazione

L'**ampiezza del campo di variazione (o semplicemente "range")** è la differenza tra il valore massimo e il valore minimo osservato nel dataset. È una misura di variabilità semplice ma molto sensibile ai valori estremi.

```r
# Calcolo del minimo, massimo e ampiezza del campo di
    variazione per 'Osservazione'.
min_oss <- min(dati$Osservazione, na.rm = TRUE)
max_oss <- max(dati$Osservazione, na.rm = TRUE)
ampiezza_variazione <- max_oss - min_oss

print(paste("Valore minimo delle osservazioni:", min_oss))
print(paste("Valore massimo delle osservazioni:", max_oss))
print(paste("Ampiezza del campo di variazione delle
    osservazioni:", ampiezza_variazione))
```

Output:

```
"Valore minimo delle osservazioni: 1"
"Valore massimo delle osservazioni: 102"
"Ampiezza del campo di variazione delle osservazioni: 101"
```

5.5 Coefficiente di variazione

Il **coefficiente di variazione (CV)** è una misura di variabilità relativa, data dal rapporto tra la deviazione standard e la media (in valore assoluto).

```r
# Calcolo del coefficiente di variazione per 'Osservazione'.
media_oss <- mean(dati$Osservazione, na.rm = TRUE)
dev_std_oss <- sd(dati$Osservazione, na.rm = TRUE)
coeff_variazione <- (dev_std_oss / abs(media_oss)) * 100 #
    abs() per media se potesse essere negativa
print(paste("Coefficiente di variazione delle osservazioni:"
    , round(coeff_variazione, 2), "%"))
```

Output(in percentuale):

```
"Coefficiente di variazione delle osservazioni: 107.6 %"
```

Un CV del 107.6% indica una variabilità molto elevata rispetto alla media. Questo è coerente con il fatto che la deviazione standard (29.61) è addirittura leggermente superiore alla media (27.51), suggerendo una notevole eterogeneità nei conteggi dei decessi.

Capitolo 6

Indici di forma

Gli **indici di forma** misurano caratteristiche relative alla forma della distribuzione dati. I più usati sono l'indice di asimmetria e l'indice di curtosi.

6.1 Indice di asimmetria

L'**indice di asimmetria (skewness)** misura il grado di simmetria di una distribuzione di dati attorno alla sua media.

- Un valore di **skewness** > 0 indica un'asimmetria positiva (coda a destra): la distribuzione ha una coda che si estende maggiormente verso i valori più alti. La media è tipicamente maggiore della mediana.
- Un valore di **skewness** < 0 indica un'asimmetria negativa (coda a sinistra): la distribuzione ha una coda che si estende maggiormente verso i valori più bassi. La media è tipicamente minore della mediana.
- Un valore di **skewness** ≈ 0 indica una distribuzione approssimativamente simmetrica (come la distribuzione normale).

```
1 # Calcolo dell'indice di asimmetria per 'Osservazione'.
2 indice_asimmetria <- skewness(dati$Osservazione, na.rm =
      TRUE)
3 print(paste("Indice di asimmetria (Skewness) delle
      osservazioni:", round(indice_asimmetria, 2)))
```

Output:

```
"Indice di asimmetria (Skewness) delle osservazioni: 0.9"
```

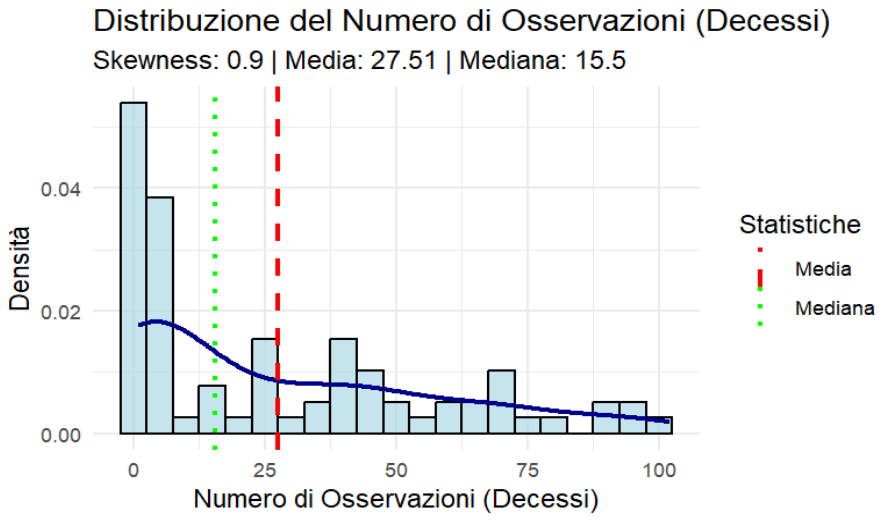
Un valore di 0.9 indica una moderata asimmetria positiva (coda a destra). Questo è coerente con l'osservazione precedente che la media (27.51) è maggiore della mediana (15.5), suggerendo che ci sono alcune osservazioni con un numero

di decessi relativamente alto che "stirano" la coda destra della distribuzione.

Per visualizzare graficamente questa asimmetria, possiamo generare un istogramma con una curva di densità sovrapposta e linee verticali per la media e la mediana.

```
# Creazione del grafico
skewness_plot <- ggplot(dati, aes(x = Osservazione)) +
  geom_histogram(aes(y = ..density..), binwidth = 5, fill = "lightblue", color = "black", alpha = 0.7) +
  geom_density(color = "darkblue", linewidth = 1) +
  geom_vline(aes(xintercept = media_oss, color = "Media"), linetype = "dashed", linewidth = 1) +
  geom_vline(aes(xintercept = mediana_generale, color = "Mediana"), linetype = "dotted", linewidth = 1) +
  labs(title = "Distribuzione del Numero di Osservazioni (Decessi)",
       subtitle = paste0("Skewness: ", round(indice_asimmetria, 2),
                         " | Media: ", round(media_oss, 2),
                         " | Mediana: ", round(mediana_generale, 2)),
       x = "Numero di Osservazioni (Decessi)",
       y = "Densita") +
  scale_color_manual(name = "Statistiche", values = c("Media" = "red", "Mediana" = "green")) +
  theme_minimal()

print(skewness_plot)
```



6.2 Indice di curtosi

L'**indice di curtosi** misura l'appiattimento di una distribuzione, in particolare la concentrazione dei dati attorno al valore centrale e la "pesantezza" delle code, rispetto a una distribuzione normale.

- **Eccesso di curtosi > 0** (Curtosi > 3): Distribuzione leptocurtica. È più appuntita al centro e ha code più pesanti (maggior probabilità di valori estremi) rispetto a una normale.
- **Eccesso di curtosi < 0** (Curtosi < 3): Distribuzione platicurtica. È più piatta al centro e ha code più leggere rispetto a una normale.
- **Eccesso di curtosi ≈ 0** (Curtosi ≈ 3): Distribuzione normocurtica. Ha una forma simile a quella di una distribuzione normale in termini di appiattimento.

```
1 # Calcolo dell'indice di curtosi (eccesso di curtosi) per '
    Osservazione'.
2 indice_curtosi <- kurtosis(dati$Osservazione, na.rm = TRUE)
3 print(paste("Indice di curtosi (eccesso di curtosi) delle
    osservazioni:", round(indice_curtosi, 2)))
4 # La curtosi "classica" sarebbe approssimativamente: round(
    indice_curtosi, 2) + 3
```

Output:

```
"Indice di curtosi (eccesso di curtosi) delle osservazioni: 2.64"
```

Un eccesso di curtosi di circa 2.64 (corrispondente a una curtosi "classica" di circa $2.64 + 3 = 5.64$) indica una distribuzione marcatamente **leptocurtica.**

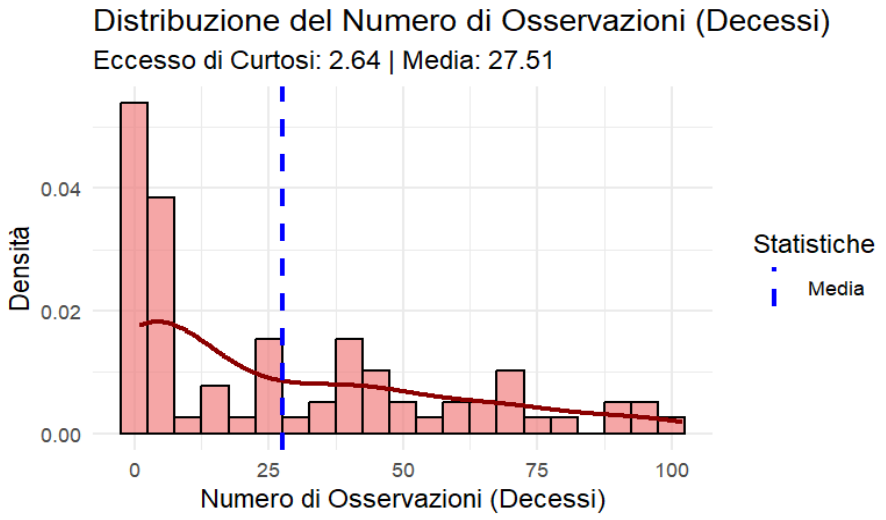
Questo significa che la distribuzione del numero di decessi per combinazione intersezione/anno presenta un picco più pronunciato attorno alla moda/mediana e code più "pesanti" rispetto a una distribuzione normale.

Ciò implica una maggiore frequenza di valori vicini al centro, ma anche una maggiore probabilità di osservare valori estremi (molto alti, data l'asimmetria positiva) rispetto a quanto ci si aspetterebbe da una distribuzione normale.

Graficamente si ha

```r
# Creazione del grafico per la curtosi
kurtosis_plot <- ggplot(dati, aes(x = Osservazione)) +
  geom_histogram(aes(y = ..density..), binwidth = 5, fill = "lightcoral", color = "black", alpha = 0.7) +
  geom_density(color = "darkred", linewidth = 1) +
  geom_vline(aes(xintercept = media_oss, color = "Media"), linetype = "dashed", linewidth = 1) +
  labs(title = "Distribuzione del Numero di Osservazioni (Decessi)",
       subtitle = paste0("Eccesso di Curtosi: ", round(indice_curtosi, 2),
                         " | Media: ", round(media_oss, 2)),
       x = "Numero di Osservazioni (Decessi)",
       y = "Densita") +
  scale_color_manual(name = "Statistiche", values = c("Media" = "blue")) +
  theme_minimal()

print(kurtosis_plot)
```



Capitolo 7

Percentili campionari

Un **percentile** k-esimo di un campione di dati è un valore che è maggiore di una percentuale k dei dati e minore della restante percentuale.

Il 25-esimo percentile si dice primo **quartile**, il 50-esimo percentile si dice mediana campionaria o secondo quartile, il 75-esimo percentile si dice terzo quartile e si denotano con

$$Q_1, Q_2, Q_3$$

.
La differenza tra il terzo e il primo quartile è chiamata **scarto interquartile.**

```r
# Calcolo dei quartili per la variabile 'Osservazione'
quartili_osservazione <- quantile(dati$Osservazione, probs = c(0.25, 0.50, 0.75), na.rm = TRUE)
Q1 <- quartili_osservazione[1]
Q2_mediana <- quartili_osservazione[2] # Coincide con la mediana gia calcolata
Q3 <- quartili_osservazione[3]
IQR_osservazione <- Q3 - Q1

print(paste("Primo Quartile (Q1) delle osservazioni:", round(Q1, 2)))
print(paste("Secondo Quartile (Q2) delle osservazioni:", round(Q2_mediana, 2)))
print(paste("Terzo Quartile (Q3) delle osservazioni:", round(Q3, 2)))
print(paste("Scarto Interquartile (IQR) delle osservazioni:", round(IQR_osservazione, 2)))

print("Sommario statistico della variabile Osservazione:")
summary_stats <- summary(dati$Osservazione, na.rm = TRUE)
print(summary_stats)
```

Output:

```
"Primo Quartile (Q1) delle osservazioni: 2"
"Secondo Quartile (Q2) delle osservazioni: 15.5"
"Terzo Quartile (Q3) delle osservazioni: 45"
"Scarto Interquartile (IQR) delle osservazioni: 43"

"Sommario statistico della variabile Osservazione:"
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00    2.00   15.50   27.51   45.00  102.00
```

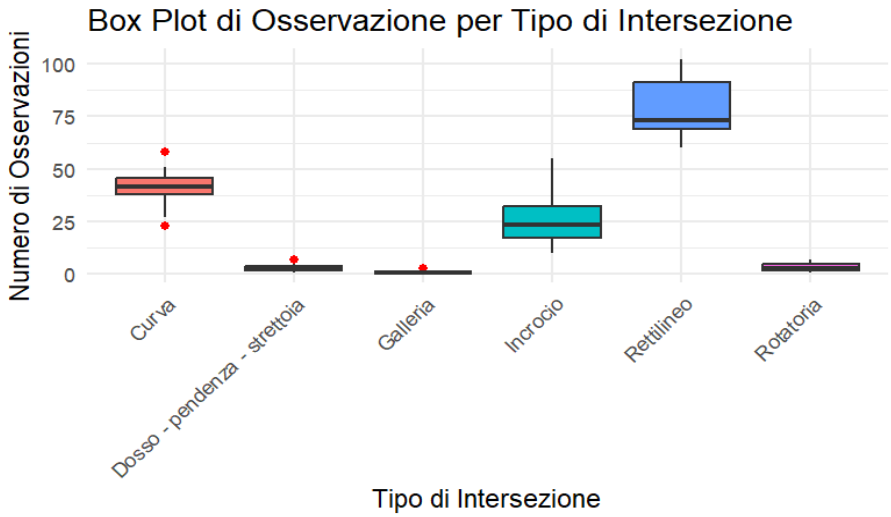
Dall'output del sommario, il 25% delle osservazioni (combinazioni intersezione/anno) ha un numero di decessi inferiore o uguale a 2, il 50% inferiore o uguale a 15.5, e il 75% inferiore o uguale a 45. Lo scarto interquartile indica che il 50% centrale dei dati si distribuisce in un intervallo di ampiezza 43.

7.1 Box Plot

Uno strumento grafico utile per la visualizzare alcuni degli indici rappresentativi dei dati è il **box plot.**

Si ottiene sovrapponendo ad una linea orizzontale che va dal minore al maggiore dei dati, un rettangolo che va dal primo al terzo quartile, con una linea verticale che lo divide al livello del secondo quartile.

```
1  # Box plot di Osservazione per tipo di Intersezione
2  Assicurarsi che la colonna Intersezione sia un fattore per
       un ordinamento corretto (se necessario)
3  dati$Intersezione <- factor(dati$Intersezione, levels = c("
       Rettilineo", "Curva", ...)) # opzionale
4  boxplot_oss_intersezione <- ggplot(dati, aes(x =
       Intersezione, y = Osservazione, fill = Intersezione)) +
5  geom_boxplot(outlier.colour = "red", na.rm = TRUE) +
6  labs(title = "Box␣Plot␣di␣Osservazione␣per␣Tipo␣di␣
       Intersezione",
7  x = "Tipo␣di␣Intersezione",
8  y = "Numero␣di␣Osservazioni") +
9  theme_minimal() +
10 theme(axis.text.x = element_text(angle = 45, hjust = 1), #
       Ruota etichette asse x
11 legend.position = "none") # Rimuove la legenda se i colori
       sono distintivi
12
13 print(boxplot_oss_intersezione)
```



Il grafico permette di confrontare la distribuzione del numero di decessi per ciascun tipo di intersezione. Si possono notare differenze nelle mediane, nella variabilità e nella presenza di [1]outlier tra le categorie. Ad esempio, i rettilinei potrebbero mostrare una mediana più alta e una maggiore dispersione.

[1]Si dice outlier un dato anomalo, ossia molto distante dagli altri dati. Vengono solitamente rappresentati come punti individuali.

7.2 Disuguaglianza di Chebyshev

La **disuguaglianza di Chebyshev** fornisce un limite inferiore alla proporzione di dati che si trovano entro un certo numero (k) di deviazioni standard dalla media campionaria (s), indipendentemente dalla forma della distribuzione. La disuguaglianza afferma che almeno $1 - 1/k^2$ dei valori si trova nell'intervallo $[\bar{x} - ks, \bar{x} + ks]$, per $k > 1$.

```r
# Disuguaglianza di Chebyshev
# media_generale e dev_std_campionaria sono stati calcolati precedentemente
# media_generale = 27.51, dev_std_campionaria = 29.61
# N_validi = sum(!is.na(dati$Osservazione)) # Numero di osservazioni non NA (79)
N_validi <- sum(!is.na(dati$Osservazione))

# Per k = 2 deviazioni standard
k2 <- 2
lim_inf_k2 <- media_generale - k2 * dev_std_campionaria
lim_sup_k2 <- media_generale + k2 * dev_std_campionaria
prop_teorica_k2 <- 1 - 1/k2^2
prop_osservata_k2 <- sum(dati$Osservazione >= lim_inf_k2 & dati$Osservazione <= lim_sup_k2, na.rm = TRUE) / N_validi

cat(paste0("Disuguaglianza di Chebyshev per k = ", k2, ":\n"))
cat(paste0("  Intervallo (media +/- ", k2, " * dev_std): [", round(lim_inf_k2, 2), ", ", round(lim_sup_k2, 2), "]\n"))
cat(paste0("  Proporzione minima teorica di dati nell'intervallo (almeno): ", round(prop_teorica_k2*100, 2), "%\n"))
cat(paste0("  Proporzione osservata di dati nell'intervallo: ", round(prop_osservata_k2*100, 2), "%\n\n"))

# Per k = 3 deviazioni standard
k3 <- 3
lim_inf_k3 <- media_generale - k3 * dev_std_campionaria
lim_sup_k3 <- media_generale + k3 * dev_std_campionaria
prop_teorica_k3 <- 1 - 1/k3^2
prop_osservata_k3 <- sum(dati$Osservazione >= lim_inf_k3 & dati$Osservazione <= lim_sup_k3, na.rm = TRUE) / N_validi

cat(paste0("Disuguaglianza di Chebyshev per k = ", k3, ":\n"))
cat(paste0("  Intervallo (media +/- ", k3, " * dev_std): [", round(lim_inf_k3, 2), ", ", round(lim_sup_k3, 2), "]\n"))
```

```r
cat(paste0("  Proporzione minima teorica di dati nell'
    intervallo (almeno): ", round(prop_teorica_k3*100, 2), "
    %\n"))
cat(paste0("  Proporzione osservata di dati nell'intervallo:
    ", round(prop_osservata_k3*100, 2), "%\n"))
```

Output:

```
Disuguaglianza di Chebyshev per k = 2:
  Intervallo (media +/- 2 * dev_std): [-31.7, 86.72]
  Proporzione minima teorica di dati nell'intervallo (almeno): 75%
  Proporzione osservata di dati nell'intervallo: 93.59%

Disuguaglianza di Chebyshev per k = 3:
  Intervallo (media +/- 3 * dev_std): [-61.3, 116.33]
  Proporzione minima teorica di dati nell'intervallo (almeno): 88.89%
  Proporzione osservata di dati nell'intervallo: 100.00%
```

Per $k = 2$, la disuguaglianza di Chebyshev garantisce che almeno il 75% dei dati si trovi nell'intervallo $[-31.7, 86.72]$. Poiché il numero di osservazioni non può essere negativo, l' intervallo effettivo considerato per i dati è $[1, 86.72]$ (essendo 1 il minimo osservato). La proporzione osservata di dati in questo intervallo è del 93.59%, che soddisfa ampiamente il limite teorico.

Per $k = 3$, almeno l'88.89% dei dati dovrebbe trovarsi nell' intervallo $[-61.3, 116.33]$ (effettivamente $[1, 102]$ per i nostri dati, dato che il massimo è 102). In questo caso, il 100% delle osservazioni valide ricade in questo intervallo, confermando la disuguaglianza.

Capitolo 8

Dati Bivariati

In questa sezione analizzeremo la relazione tra due variabili quantitative: l'Anno di riferimento e il Numero Totale di Morti annuali per incidenti stradali (limitatamente alla fascia di età 21-24 anni, conducenti). Useremo i dati aggregati in `morti_per_anno`.

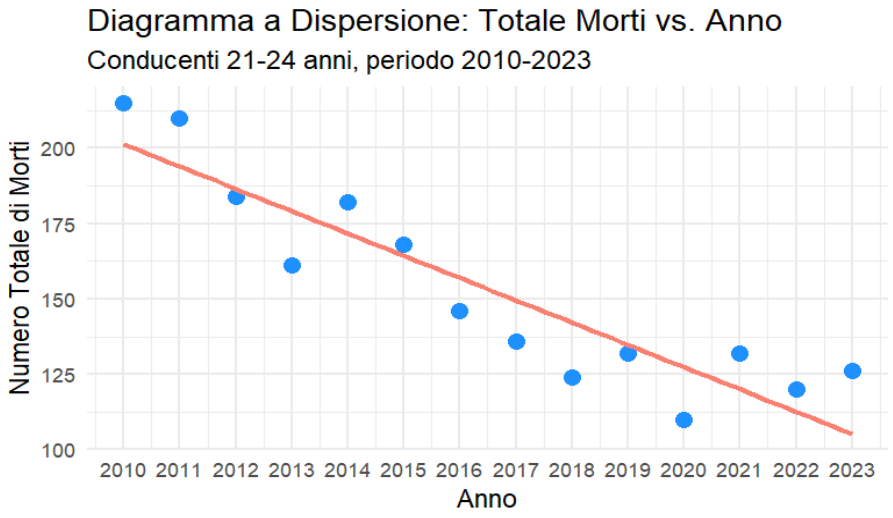
8.1 Diagramma a dispersione

Un diagramma a dispersione (scatter plot) mostra la relazione tra due variabili quantitative.

```
# Converti gli anni in numerici per la regressione
if (!is.numeric(morti_per_anno$TIME_PERIOD)) {
  morti_per_anno$TIME_PERIOD_Numeric <- as.numeric(as.
      character(morti_per_anno$TIME_PERIOD))
} else {
  morti_per_anno$TIME_PERIOD_Numeric <- morti_per_anno$TIME_
      PERIOD
}

# Crea lo scatter plot
scatter_plot_annuale <- ggplot(morti_per_anno, aes(x = TIME_
    PERIOD_Numeric, y = Totale_Morti)) +
  geom_point(color = "dodgerblue", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "salmon") +
       # Aggiunge retta di regressione (opzionale qui, ma
      utile)
  labs(title = "Diagramma␣a␣Dispersione:␣Totale␣Morti␣vs.␣
      Anno",
       subtitle = "Conducenti␣21-24␣anni,␣periodo␣2010-2023"
           ,
       x = "Anno",
       y = "Numero␣Totale␣di␣Morti") +
```

```
16    scale_x_continuous(breaks = seq(min(morti_per_anno$TIME_
          PERIOD_Numeric, na.rm=T),
17                                    max(morti_per_anno$TIME_
                                          PERIOD_Numeric, na.rm=T
                                          ), by = 1)) + #
                                          Assicura che gli anni
                                          siano interi
18    theme_minimal()
19
20  print(scatter_plot_annuale)
```



Il diagramma a dispersione mostra il numero totale di morti per incidenti stradali (nella fascia di età 21-24 anni, conducenti) per ciascun anno dal 2010 al 2023. Ogni punto blu rappresenta un anno. La linea color salmone rappresenta la **retta di regressione** lineare, che suggerisce l'andamento generale dei dati nel tempo. Visivamente, si osserva una tendenza generale alla diminuzione del numero di morti nel corso degli anni considerati, sebbene con alcune fluttuazioni.

8.2 Coefficiente di correlazione campionario

Il **coefficiente di correlazione campionario** (r) è una misura normalizzata della forza e della direzione della relazione lineare tra due variabili quantitative. Varia tra -1 e +1.

- $r = +1$: perfetta correlazione lineare positiva.
- $r = -1$: perfetta correlazione lineare negativa.

- $r \approx 0$: assenza di correlazione lineare (o relazione lineare molto debole).
- Valori vicini a +1 o -1 indicano una forte relazione lineare, mentre valori vicini a 0 indicano una relazione lineare debole o assente.

```
# Calcolo del coefficiente di correlazione tra Anno e Totale Morti
correlazione_annuale <- cor(morti_per_anno$TIME_PERIOD_Numeric, morti_per_anno$Totale_Morti, use = "complete.obs", method = "pearson")
print(paste("Coefficiente di correlazione campionario tra Anno e Totale Morti:", round(correlazione_annuale, 4)))
```

Output:

```
"Coefficiente di correlazione campionario tra Anno e Totale Morti: -0.9132"
```

Un coefficiente di correlazione campionario di circa -0.9132 indica una **forte relazione lineare negativa** tra l'anno e il numero totale di morti.

Ciò conferma in modo più standardizzato l'osservazione precedente: all'avanzare degli anni, si è osservata una marcata tendenza alla diminuzione del numero di decessi per incidenti stradali tra i conducenti di età compresa tra 21 e 24 anni nel periodo 2010-2023.