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, e di gravi conseguenze in termini di disabilità temporanee e permanenti, oltre al drammatico impatto umano e psicologico sulle vittime e 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 e di numerosi Enti pubblici istituzionali, è a carattere totale e a cadenza mensile (inserita tra le rilevazioni di interesse pubblico nel Programma Statistico Nazionale - PSN - IST00142).

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

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

formule se necessario.

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) 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.

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
# inclusione librerie
library("tidyverse")
require("tidyverse")
library("dplyr")

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

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

```
28
29 # 4. Frequenze cumulate relative
30 freq_assolute$Frequenza_Cumulata_Relativa <- freq_assolute$
       Frequenza_Cumulata_Assoluta / totale
31 print(freq_assolute)
```

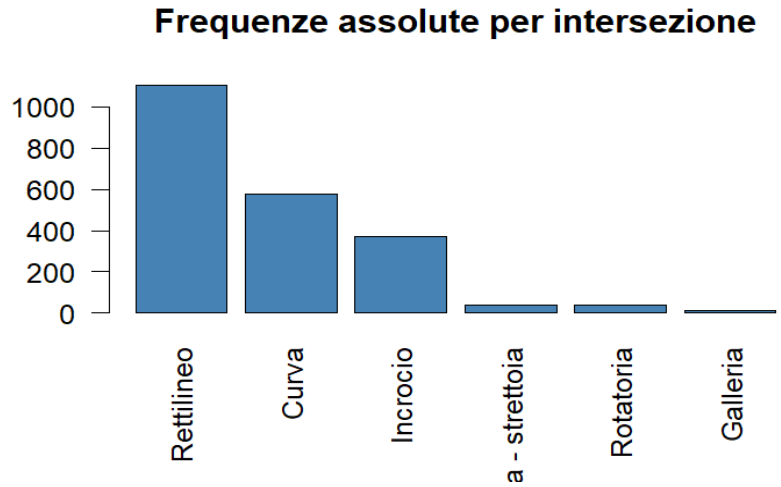
L'output generato dal codice R riportato nel Listing è sintetizzato nella seguente tabella (si noti che i valori esatti dipendono dalla gestione dei dati mancanti "–" nel file CSV originale; i valori qui riportati sono quelli della tabella fornita nel prompt):

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

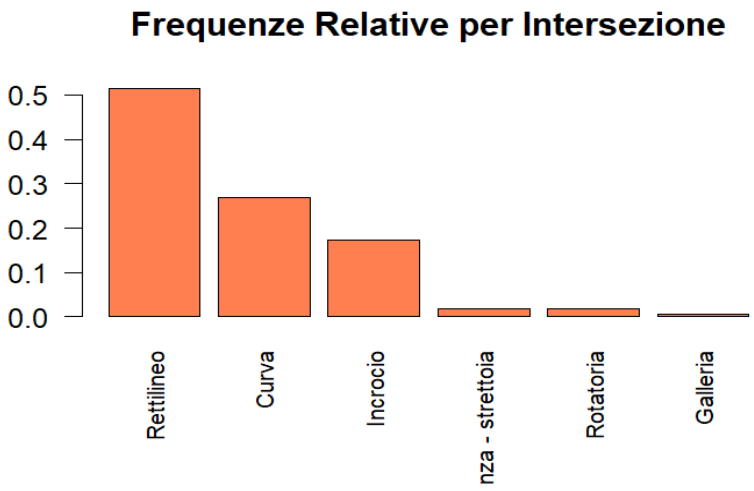
Capitolo 3

Rappresentazione dei dati mediante grafici

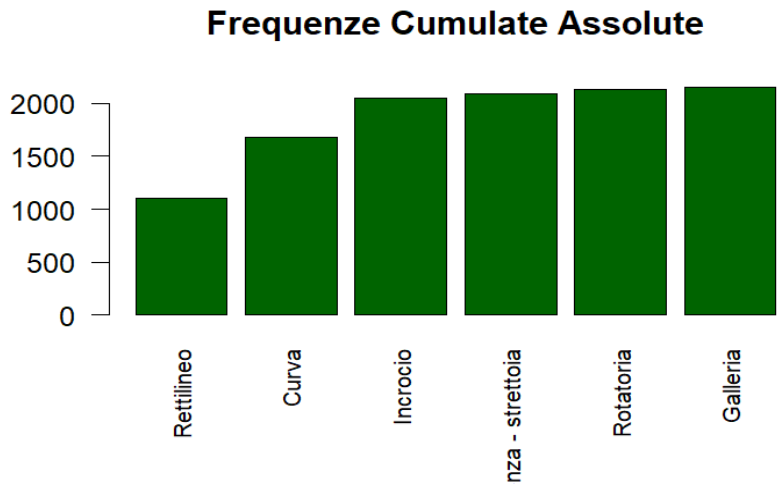
```r
# 1. Grafico frequenza assolute
barplot(freq_assolute$Frequenza_Assoluta,
        names.arg = freq_assolute$Intersezione,
        main = "Frequenze assolute per intersezione",
        xlab = "",
        ylab = "",
        col = "steelblue",
        las = 2) # Etichette verticali
```



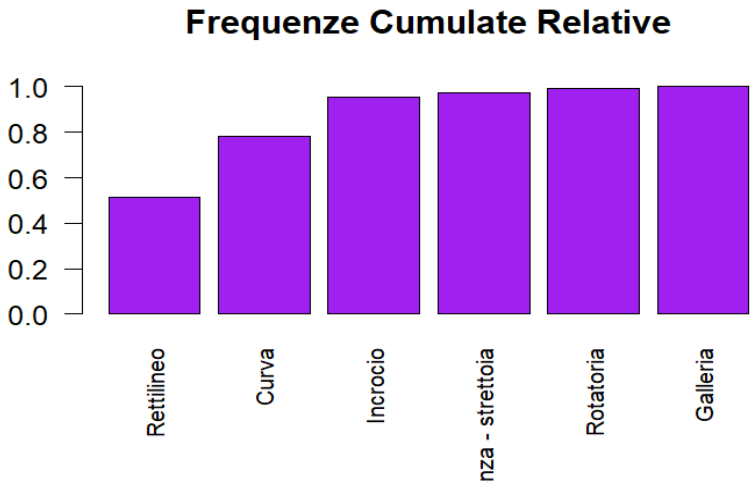
```r
# 2. Grafico frequenze relative
barplot(freq_assolute$Frequenza_Relativa,
        names.arg = freq_assolute$Intersezione,
        main = "Frequenze Relative per Intersezione",
        xlab = "",
        ylab = "",
        col = "coral",
        las = 2,
        cex.names = 0.8)
```




```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)
```



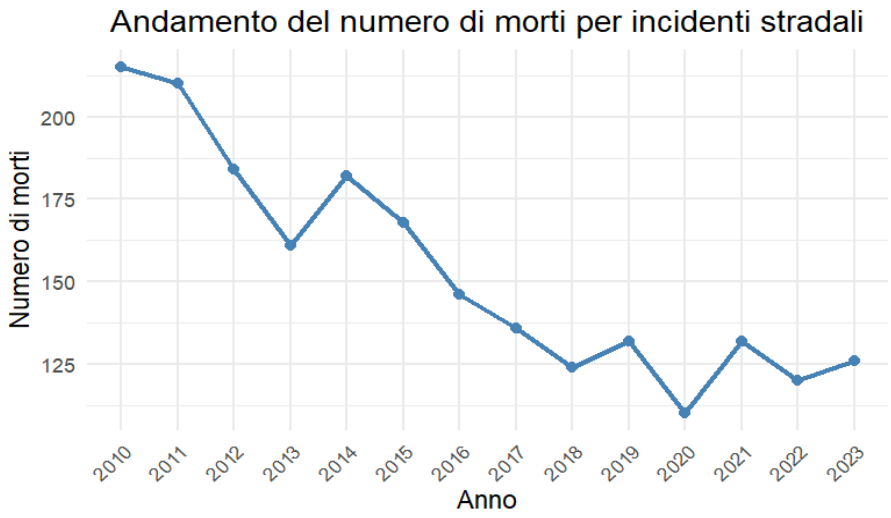

```r
# 4. Grafico frequenze cumulate relative
barplot(freq_assolute$Frequenza_Cumulata_Relativa,
        names.arg = freq_assolute$Intersezione,
        main = "Frequenze Cumulate Relative",
        xlab = "",
        ylab = "",
        col = "purple",
        las = 2,
        cex.names = 0.8)
```



```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
```



Capitolo 4

Indici di posizione

In questo capitolo verranno calcolati gli indici di posizione sulla variabile numerica `Osservazione` del dataset `dati`.

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.

4.1 Media campionaria

La media campionaria è la somma di tutte le osservazioni divisa per il numero di osservazioni. 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 il valore centrale di un insieme di dati ordinati. Divide i dati in due metà: il 50% delle osservazioni ha un valore inferiore o uguale alla mediana, e il 50% ha un valore superiore o uguale. È meno sensibile ai valori estremi rispetto alla media.

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

Output:

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

Metà delle combinazioni intersezione/anno hanno registrato 15.5 morti o meno, e l'altra metà 15.5 morti o più. 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.

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

find_mode(dati)
```

Output:

Capitolo 5

Indici di variabilità

Di seguito calcoliamo gli indici di variabilità, 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 (valore basato sulla media e mediana fornite, è una stima):

```
"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 (Mean Absolute Deviation, MAD) dalla media è la media delle deviazioni assolute (cioè, senza segno) dei dati dalla loro media. 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
```

```
5
6 print(paste("Valore minimo delle osservazioni:", min_oss))
7 print(paste("Valore massimo delle osservazioni:", max_oss))
8 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).

```
1 # Calcolo del coefficiente di variazione per 'Osservazione'.
2 media_oss <- mean(dati$Osservazione, na.rm = TRUE)
3 dev_std_oss <- sd(dati$Osservazione, na.rm = TRUE)
4 coeff_variazione <- (dev_std_oss / abs(media_oss)) * 100 #
    abs() per media se potesse essere negativa
5 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

6.1 Indice di asimmetria

6.2 Indice di curtosi