Università degli Studi di Milano - Bicocca

Foundations of Probability and Statistics

Progetto finale

Analisi degli incidenti stradali con lesioni a persone in Italia

Microdati ISTAT, 2017

Testa Deborah 817343, Villa Greta 800491, Zayeva Nataliya 867981

8 febbraio 2021

Abstract

L'informazione statistica sull'incidentalità stradale è prodotta dall'Istat sulla base di una rilevazione di tutti gli incidenti stradali verbalizzati da un'autorità di polizia, verificatisi sul territorio nazionale nell'arco di un anno solare, che hanno causato lesioni alle persone (deceduti entro il trentesimo giorno oppure feriti). I dati sono stati raccolti da Organi pubblici a competenza locale e fanno riferimento all'anno 2017. Lo scopo di questa analisi è quindi comprendere le circostanze nelle quali si verificano tali incidenti e le loro caratteristiche principali, inoltre si è cercato di individuare quali condizioni sono associate a un maggior numero di soggetti feriti e deceduti. Dopo una prima parte introduttiva di presentazione e pulizia del dataset, si è proceduto con l'analisi descrittiva univariata e bivariata. Sono state quindi analizzate sia le variabili quantitative, relative al numero di soggetti lesi o deceduti, sia le variabili categoriche, estraendo informazioni riguardanti, ad esempio, la tipologia dei mezzi coinvolti, i fattori ambientali e le caratteristiche legate a genere, età e tipologia di patente conseguita dal conducente. Si è poi proceduto analizzando il grado di connessione sia tra le variabili categoriche sopracitate sia tra queste e il numero di persone decedute o rimaste ferite. Gli strumenti statistici utilizzati per quest'ultima parte di analisi sono l'indice V di Cramer e la tecnica ANOVA a una via.

1 Dataset e Pre-Processing

Il dataset utilizzato è reperibile al sito Istat nella sezione 'microdati ad uso pubblico' al link `http://www.istat.it/it/archivio/87539` e si riferisce ai dati raccolti nell'anno 2017. Il dataset è composto da 174933 osservazioni e 117 colonne. Durante il pre-processing sono state unite le colonne riguardanti i morti entro 24 ore e i morti entro 30 giorni nella colonna "Totale morti". Inoltre si è deciso di rimuovere le colonne non considerate rilevanti al fine della nostra indagine, come ad esempio "trimestre di riferimento" e "organo di rilevazione" e le colonne che non portavano nessun tipo di informazione come "anno di riferimento dei dati" (essendo sempre 2017). Si è poi osservata la presenza di valori nulli all'interno del dataset; tali valori mancanti forniscono comunque un'informazione nel contesto considerato, quindi le righe corrispondenti a tali osservazioni non sono state rimosse.
I dati raccolti si riferiscono ad incidenti che coinvolgono una oppure due veicoli principali "Veicolo A" e "Veicolo B", non ci sono dati relativi ad altri veicoli coinvolti. Nella nostra analisi abbiamo quindi spesso suddiviso il dataset tra osservazioni che riportano un dato non mancante in "tipo veicolo B", ovvero relative ad incidenti con almeno due veicoli coinvolti, e osservazioni che riportano un dato mancante in "Tipo veicolo B", ovvero relative ad incidenti con un solo veicolo coinvolto o senza informazioni riguardo ad eventuali altri veicoli.
Per questo report abbiamo analizzato i dati relativi alle seguenti 24 colonne del dataset iniziale e abbiamo creato la colonna 'Totale Lesioni' che unisce i dati relativi sia a feriti che ai deceduti.
Le prime 11 colonne si riferiscono ai fattori esterni e alle circostanze generali in cui è avvenuto l'incidente e al numero totale di persone rimaste lese.

- Localizzazione incidente
- Tipo di strada
- Pavimentazione
- Fondo stradale
- Segnaletica
- Condizioni meteorologiche
- Natura incidente
- Intersezione o non intersezione
- Totale morti
- Totale feriti
- Totale lesioni

Le successive 14 colonne si riferiscono alle caratteristiche principali dei conducenti e dei veicoli coinvolti nell'incidente.

- Tipo veicolo 'a'
- Veicolo 'a': circostanze/inconvenienti di circolazione
- Immatricolazione veicolo 'a'
- Veicolo 'a': età conducente
- Veicolo 'a': sesso conducente
- Veicolo 'a': patente conducente
- Veicolo 'a': anno rilascio patente conducente
- Tipo veicolo 'b'
- Veicolo 'b': circostanze/inconvenienti di circolazione
- Immatricolazione veicolo 'b'
- Veicolo 'b': età conducente
- Veicolo 'b': sesso conducente
- Veicolo 'b': patente conducente
- Veicolo 'b': anno rilascio patente conducente

2 Analisi Univariata

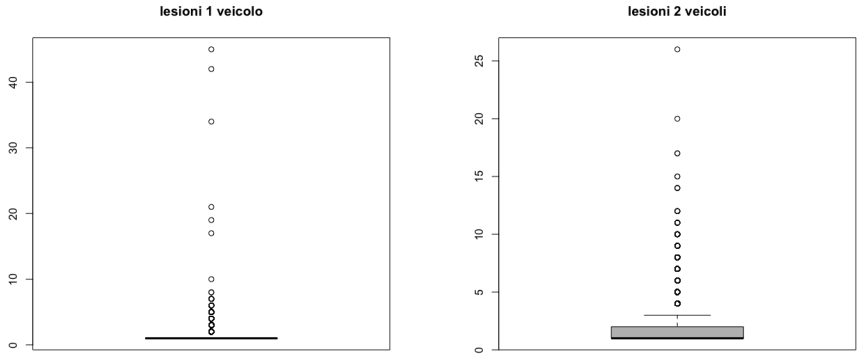
2.1 Variabili Quantative

L'analisi si è inizialmente focalizzata sulle variabili quantitative di cui abbiamo osservato le distribuzioni mediante i principali indicatori di posizione e di variabilità, di seguito riportati.

```
 totale morti      totale feriti       lesioni
Min.   : 0.00000   Min.   : 0.000   Min.   : 1.00
1st Qu.: 0.00000   1st Qu.: 1.000   1st Qu.: 1.00
Median : 0.00000   Median : 1.000   Median : 1.00
Mean   : 0.01931   Mean   : 1.411   Mean   : 1.43
3rd Qu.: 0.00000   3rd Qu.: 2.000   3rd Qu.: 2.00
Max.   :16.00000   Max.   :45.000   Max.   :45.00
```

Variabili	**Varianza**	**Deviazione Standard**
totale morti	0.0228017	0.1510028
totale feriti	0.7767274	0.8813239
lesioni	0.7783003	0.8822158

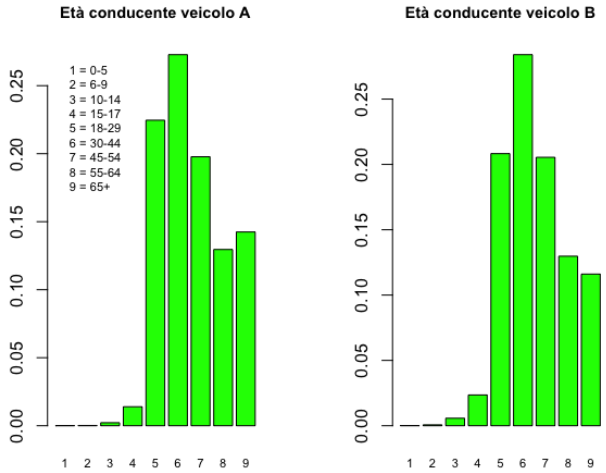
Possiamo osservare che per la variabile "Totale morti" la mediana, il primo e il terzo quartile coincidono e sono pari a zero. Inoltre si ha che la deviazione standard è pari a 0.15, mentre quella relativa ai soggetti rimasti feriti e al totale delle persone lesionate è pari a 0.88; da cui si può affermare una maggiore variabilità nella distribuzione di queste ultime.



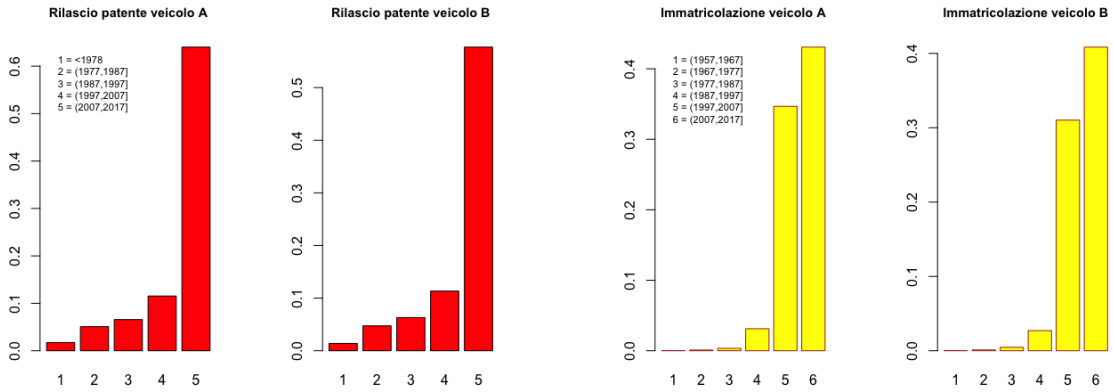
Si osservino i boxplot per la variabile "Totale Lesioni", riferiti rispettivamente ad incidenti con un solo veicolo coinvolto e ad incidenti con due veicoli coinvolti. Si può notare che nel caso di due veicoli coinvolti la distribuzione è asimmetrica positiva, mentre nel caso di un solo veicolo coinvolto la mediana coincide con il primo e terzo quartile. In entrambi i casi notiamo un'elevata presenza di outliers.

2.2 Variabili Categoriche Ordinali

Si è successivamente proceduto con l'analisi delle variabili categoriche ordinali, analizzando i grafici delle frequenze relative.

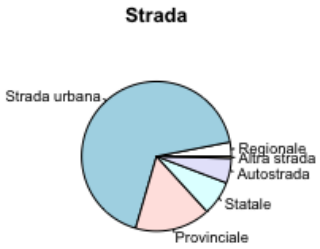


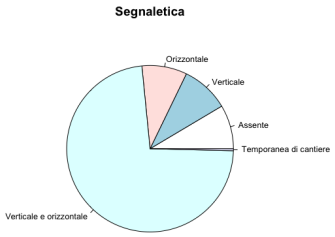
Si considerino i dati relativi all'età dei conducenti dei veicoli coinvolti: come possiamo aspettarci, non c'è differenza tra veicolo A e veicolo B. Inoltre, si osservano alcune rilevazioni di conducenti con meno di 18 anni: questo può essere spiegato dal fatto che le biciclette siano considerate veicoli e che la patente AM può essere conseguita a partire dai 14 anni. Il picco si verifica in corrispondenza del gruppo 30-44 anni.

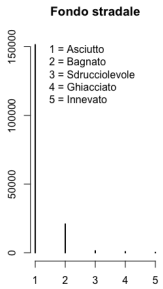


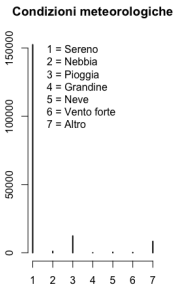
Per quanto riguarda gli anni in cui sono state rilasciate le patenti dei conducenti e l'immatricolazione dei loro veicoli, anche in questo caso non c'è significativa differenza tra veicolo A e veicolo B. La maggior parte delle patenti è stata rilasciata nei 10 anni precedenti rispetto all'anno in cui sono stati registrati gli incidenti, nonostante la maggior parte dei conducenti avesse più di 30 anni (come osservato precedentemente): questo indica che la poca esperienza alla guida, più che l'età del conducente, favorisce l'eventualità di incidenti. Osserviamo che la quasi totalità dei veicoli è stata immatricolata nei 20 precedenti, con una maggiore prevalenza negli ultimi 10 anni: possiamo aspettarci che questo sia dovuto al fatto che la maggior parte dei veicoli che circolano sul territorio non abbia più di 20 anni.

2.3 Variabili Categoriche Nominali

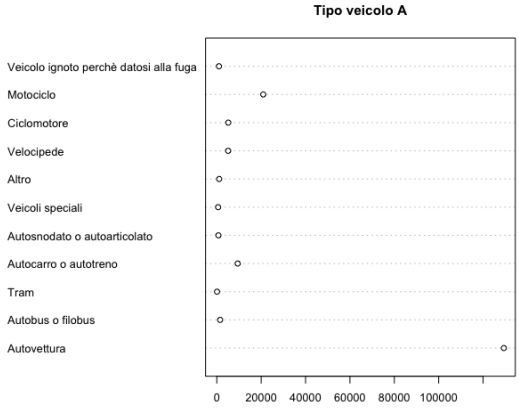


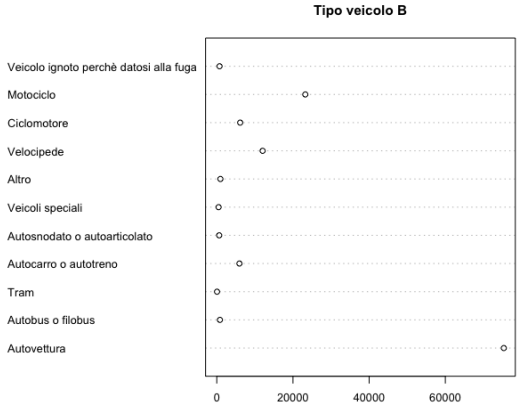


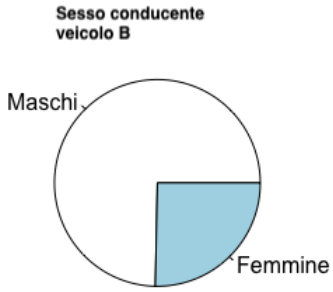




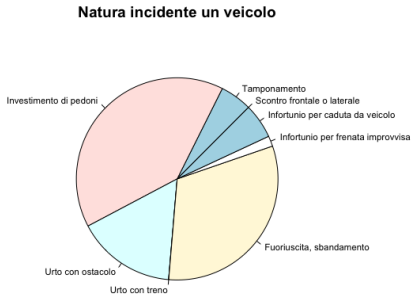
L'analisi prosegue con lo studio delle variabili categoriche nominali. Analizzando le caratteristiche della strada nelle figure soprastanti, si nota che quasi due terzi degli incidenti sono avvenuti in strade urbane a doppio senso, nella quasi totalità dei casi la strada era pavimentata e la segnaletica orizzontale o verticale era presente. Di conseguenza, l'eventuale non pavimentazione della strada (presenza di buche o tratti sterrati) o la mancanza di segnaletica non hanno una forte incidenza sull'eventualità di incidenti. Inoltre si è verificato che il fondo stradale era prevalentemente asciutto, in concordanza con le condizioni meteorologiche in prevalenza di tempo sereno.

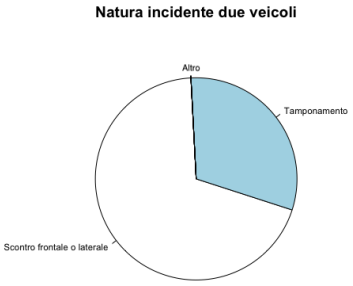


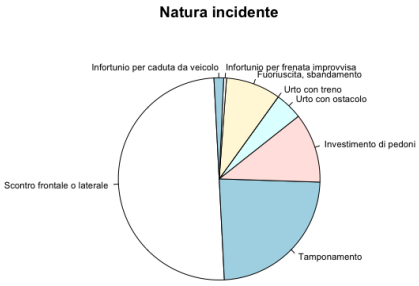




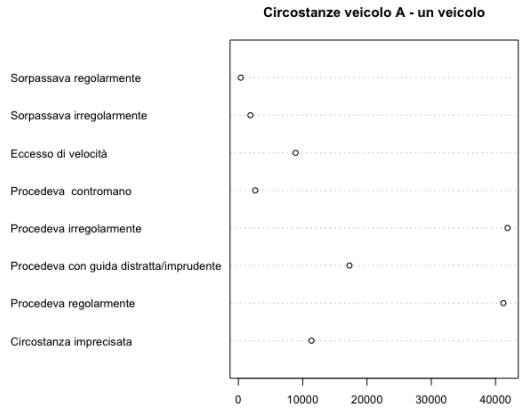
Per quanto riguarda il genere dei conducenti notiamo che i due terzi sono maschi e un terzo femmine sia per il veicolo A che per il veicolo B. I veicoli sono per la maggior parte autovetture, consistentemente con il fatto che questa tipologia di veicoli è quella più frequente; in "Veicolo B" vediamo un aumento dei velocipedi (biciclette).

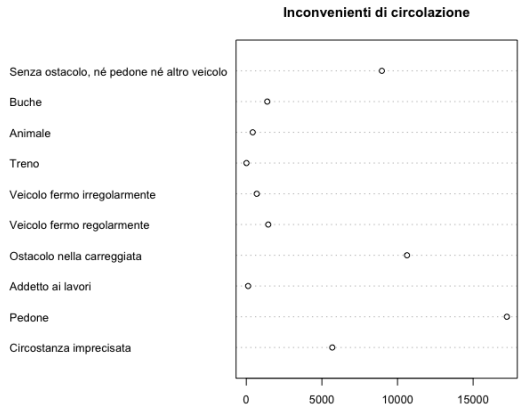




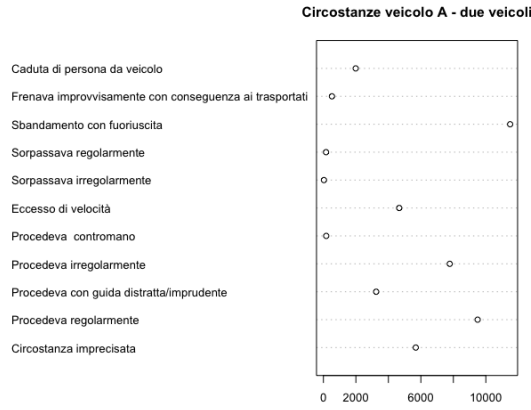


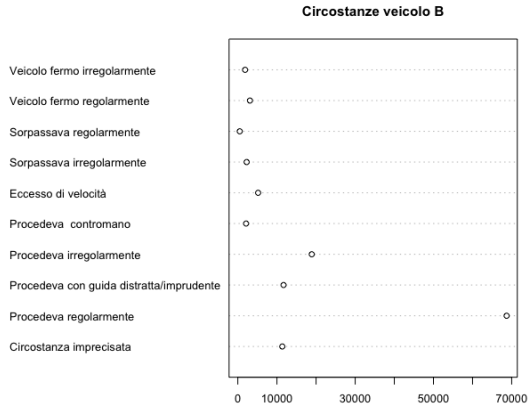
Passando alla natura dell'incidente, abbiamo prima analizzato il totale delle osservazioni e poi separatamente i casi con un solo veicolo coinvolto e con due veicoli coinvolti. La metà degli incidenti totali è avvenuta a causa di scontri frontali o laterali, un altro quarto è avvenuto per tamponamento; tra le cause rimanenti troviamo fuoriuscite/sbandamenti e investimenti. Questo risultato è dovuto al fatto che il dataset in questione comprende solo incidenti che hanno causato danni a persone, e gli scontri tra veicoli, soprattutto se frontali, hanno spesso questo spiacevole esito. Nel caso di due veicoli coinvolti, le cause si riducono principalmente a due: scontro frontale o laterale per quasi i due terzi dei casi e tamponamento per la restante parte. Nel caso di un solo veicolo invece, tra le maggiori cause troviamo: fuoriuscita/sbandamento, investimento di pedoni e urto con ostacolo.





Nel caso di un solo veicolo coinvolto inoltre, le circostanze più frequenti sono: "precedeva regolarmente" e "precedeva irregolarmente" (modalità che riguarda il rispetto della segnaletica ma non comprende eccessi di velocità); si nota poi una consistente parte di incidenti in cui il conducente precedeva con guida distratta. In questo caso gli "inconvenienti di circolazione" più frequenti riguardano la presenza di pedoni e ostacoli nella carreggiata, anche se una larga parte di incidenti è avvenuta senza cause esterne.

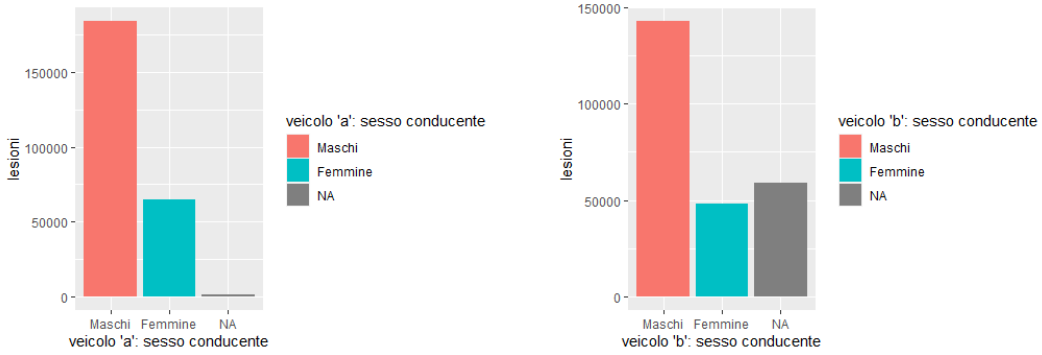




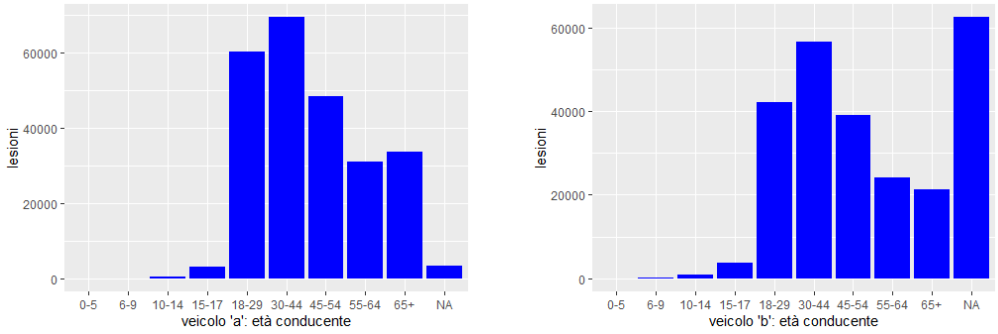
Nel caso di due veicoli coinvolti, le circostanze più frequenti sono: "sbandamento con fuoriuscita dalla corsia", "precedeva regolarmente" e "precedeva irregolarmente"; il secondo veicolo coinvolto nella maggior parte dei casi procedeva regolarmente.

3 Analisi Bivariata

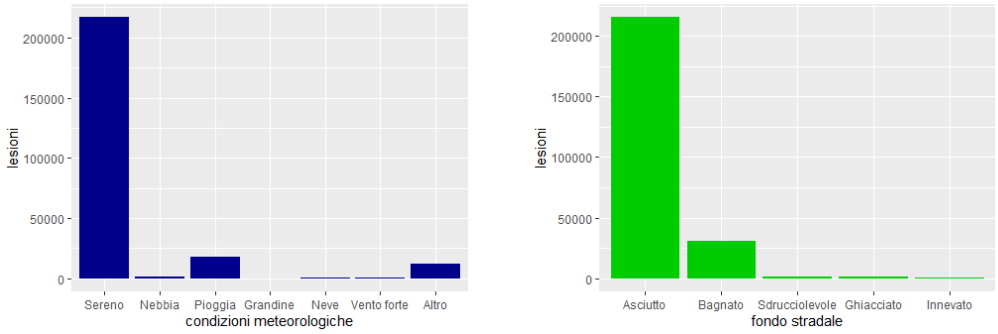
Si è proceduto con l'analisi bivariata delle variabili categoriali per valutare l'esistenza di relazioni tra di loro o per descriverne l'andamento. Il punto di partenza è stata la costruzione delle tabelle di contingenza per poi rappresentarle con i grafici che ci hanno permesso di studiare le caratteristiche degli incidenti che hanno portato a un maggior numero di lesioni.



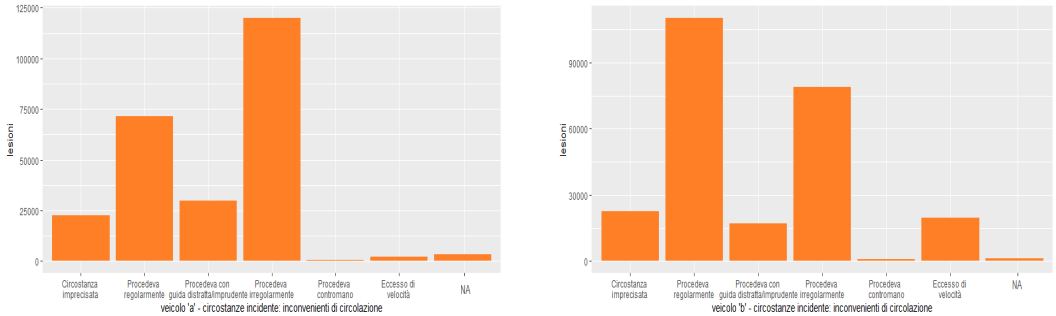
I risultati sono visualizzati nei barplot che permetto di sintetizzare le varie caratteristiche. Nelle rappresentazioni soprastanti possiamo osservare che il maggior numero di lesioni è stato procurato da conducenti di sesso maschile, sia per il veicolo A che B. Come precedentemente discusso la presenza di valori nulli nelle colonne relative il veicolo B è dovuta al fatto che non tutti gli incidenti vedono il coinvolgimento di due veicoli.



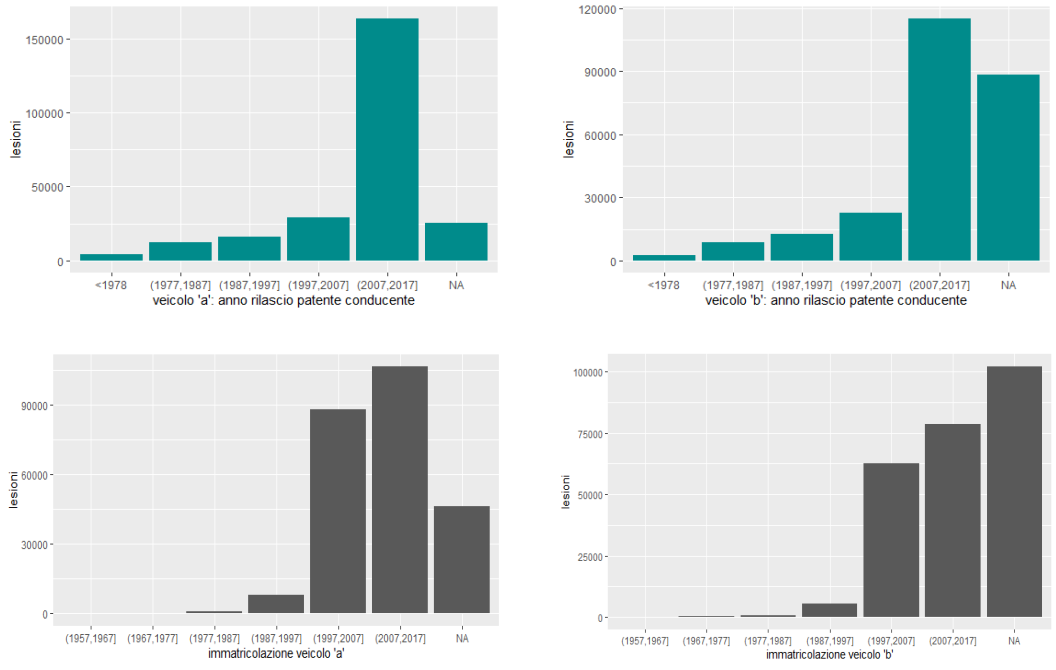
Considerando i conducenti dei mezzi coinvolti negli incidenti, si è verificato che la fascia d'età tra i 30-44 anni è quella associata ad un maggior numero di soggetti lesionati; inoltre si registra una quota significativa anche in corrispondenza della fascia tra i 18 e i 29 anni. Confrontando i grafici relativi al veicolo A e al veicolo B possiamo osservare un maggior numero di di valori mancanti nel secondo caso, poiché non in tutti gli incidenti è stato coinvolto un secondo veicolo. Questa considerazione sarà valida anche per le analisi successive.



Si è poi analizzato in che condizioni meteorologiche e di fondo stradale si è registrato il maggior numero di soggetti lesionati. Dai due grafici precedenti vediamo che il maggior numero di persone lesionate è stato osservato in condizioni di tempo sereno e con fondo stradale asciutto. Anche se c'è una significativa quantità di persone lesionate in tempo di pioggia e su strada bagnata, questo risultato minore rispetto al tempo sereno può essere dovuto alla maggior attenzione del guidatore durante un tempo non favorevole e al fatto che i giorni di pioggia sono meno rispetto ai giorni sereni.



Per quanto riguarda invece le circostanze degli incidenti possiamo osservare una discrepanza tra il veicolo A e il veicolo B. Se nel primo caso la maggioranza di lesioni è avvenuta a causa di andamento irregolare, nel secondo caso sono avvenuti in condizioni regolari di guida, questo è probabilmente dovuto al fatto che il Veicolo A sia il maggior responsabile dell'incidente. Si potrebbe pensare che la maggior parte degli incidenti con feriti sia dovuto ad eccesso di velocità, ma dai grafici si vede che per il veicolo A il numero è minimo e solo per il veicolo B è leggermente più alto, posizionandosi al quarto posto tra tutte le circostanze degli incidenti.



Osservando i grafici relativi all'immatricolazione dei veicoli e al rilascio della patente del conducente si osserva un maggior numero di lesioni nel caso di veicoli immatricolati tra il 2008 e il 2017 e nel caso di soggetti che hanno conseguito la patente da meno di 10 anni: si può quindi supporre che tra questi siano presenti persone neopatentate, ma essendo gli intervalli molto ampi si va a perdere una parte dell'informazione. Questo risultato è consistente con quanto atteso: la minor esperienza alla guida favorisce l'eventualità di incidenti. Si osserva una sostanziale parte anche di veicoli immatricolati tra il 1997-2007 sia per il veicolo A che per il

veicolo B: questo può essere dovuto a macchine datate che necessitano di una migliore manutenzione e che possono presentare dei problemi improvvisi che potrebbero essere tra le cause degli incidenti. Inoltre i veicoli meno recenti hanno sistemi di sicurezza peggiori e questo può portare a un maggior numero di persone lese.

4 Correlazione

4.1 V di Cramer

In questa parte finale di analisi è stata calcolata la statistica descrittiva V di Cramer che misura l'associazione tra due variabili categoriche, la cui formula è:

$$V = \sqrt{\frac{\chi^2}{\min(k-1, r-1)n}}$$

Di seguito vengono riportati alcuni dei valori della statistica V di Cramer calcolati:

Prima variabile	**Seconda variabile**	**V di Cramer**
natura incidente	tipo veicolo 'a'	*0.1232*
veicolo 'a': patente conducente	natura incidente	*0.0681*
condizioni meteorologiche	fondo stradale	*0.4989*
pavimentazione	fondo stradale	*0.0728*
veicolo 'a' - circostanze	natura incidente	*0.6416*
veicolo 'b' o inconvenienti di circolazione	natura incidente	*0.6457*
natura incidente	intersezione o non intersezione	*0.2000*
intersezione o non intersezione	veicolo 'a' - circolazione	*0.2749*
intersezione o non intersezione	veicolo 'b' o inconvenienti di circolazione	*0.2827*
veicolo 'a' - circolazione	veicolo 'b' o inconvenienti di circolazione	*0.4739*

In particolare è interessante notare che, come ci si aspettava, sussiste una significativa connessione tra gli inconvenienti di circolazione in entrambi i veicoli (A e B), e la natura dell'incidente. Si noti anche che tra le variabili riferite rispettivamente alle condizioni meteorologiche e alla condizione del fondo stradale sussiste un moderato livello di connessione. Una buona connessione sussiste anche tra gli inconvenienti di circolazione rispettivamente del veicolo A e del veicolo B. I restanti valori della statistica ottenuti risultano invece essere deboli.

4.2 ANOVA

ANOVA è una procedura statistica che consente di individuare se vi siano delle differenze tra più gruppi di dati. E' stata considerata la variabile 'Totale Lesioni' rispetto ad ogni variabile categorica. Per interpretare tale risultato bisogna tenere conto che l'ipotesi nulla è che le varianze siano uguali fra di loro, e che dunque la variabile indipendente non produca effetti su quella dipendente. La probabilità che sia vera l'ipotesi nulla è indicata dal valore P-value. Nei casi riportati di seguito si ha un valore di p-value inferiore alla soglia di significatività pari al 0.01, ciò indica una significativa differenza nella varianza tra i gruppi e quindi si può affermare con il 99% di confidenza che il numero di persone con lesioni varia significativamente rispetto alla variabile categorica considerata.

```
Analysis of Variance Table

Response: lesioni
                    Df Sum Sq Mean Sq F value    Pr(>F)
`natura incidente`   7   3845  549.32  726.28 < 2.2e-16 ***
Residuals       174925 132305    0.76
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Analysis of Variance Table

Response: lesioni
                           Df Sum Sq Mean Sq F value    Pr(>F)
`condizioni meteorologiche`  6     58  9.6629   12.42 4.851e-14 ***
Residuals                174926 136092  0.7780
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Analysis of Variance Table

Response: lesioni
                      Df Sum Sq Mean Sq F value    Pr(>F)
df$`fondo stradale`     4    109 27.1931  34.966 < 2.2e-16 ***
Residuals          174928 136042  0.7777
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Variance Table

Response: lesioni
                                  Df Sum Sq Mean Sq F value    Pr(>F)
`intersezione o non intersezione`  5    308  61.698   79.45 < 2.2e-16 ***
Residuals                     174927 135842   0.777
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Variance Table

Response: lesioni
                       Df Sum Sq Mean Sq F value    Pr(>F)
df$`tipo veicolo 'a'`  10   2406 240.619   314.7 < 2.2e-16 ***
Residuals          174922 133744   0.765
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Variance Table

Response: lesioni
                       Df Sum Sq Mean Sq F value    Pr(>F)
df$`tipo veicoli 'b'`  10   7950  795.00  956.86 < 2.2e-16 ***
Residuals          126360 104985    0.83
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Variance Table

Response: lesioni
                                    Df Sum Sq Mean Sq F value    Pr(>F)
df$`veicolo 'a': sesso conducente`   1     24  23.700  30.344 3.624e-08 ***
Residuals                       173955 135868   0.781
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Variance Table

Response: lesioni
                                    Df Sum Sq Mean Sq F value    Pr(>F)
df$`veicolo 'a': patente conducente`  9    718  79.733  101.18 < 2.2e-16 ***
Residuals                       166662 131339   0.788
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Variance Table

Response: lesioni
                                    Df Sum Sq Mean Sq F value    Pr(>F)
df$`veicolo 'b': patente conducente`  9   2660 295.534  328.02 < 2.2e-16 ***
Residuals                       116570 105026   0.901
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Nei seguenti casi non si è potuta rifiutare l'ipotesi nulla poiché si sono ottenuti valori di p-value non significativi, superiori alla soglia di significatività pari al 0.01.

```
Analysis of Variance Table

Response: lesioni
                                    Df Sum Sq Mean Sq F value Pr(>F)
df$`veicolo 'b': sesso conducente`   1      2 1.70116  1.8964 0.1685
Residuals                       125644 112708 0.89704
```

```
Analysis of Variance Table

Response: lesioni
                      Df Sum Sq Mean Sq F value  Pr(>F)
df$pavimentazione      1      4  3.5432  4.5526 0.03287 *
Residuals         174931 136147  0.7783
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5 Conclusioni

In conclusione dell'analisi condotta riportiamo le considerazioni più significative:

- la media delle persone lesionate è 1,43
- quasi due terzi degli incidenti è avvenuta su strada urbana a doppio senso, ma la segnaletica e la pavimentazione non sembrano influenzare l'eventualità di incidenti
- le condizioni meteorologiche associate ad un maggior numero di incidenti e, di conseguenza, ad un maggior numero di lesioni sono relative ad un tempo sereno
- la maggior parte degli incidenti coinvolge conducenti di autovetture di età compresa tra i 30 e 44 anni con poca esperienza alla guida (patente da meno di 10 anni); tali circostanze portano ad un maggior numero di lesioni, come visto anche nell'analisi bivariata
- la maggior parte degli incidenti che coinvolgono un veicolo riguarda l'investimento di pedoni o fuoriuscite dalla corsia e sbandamenti
- la maggior parte degli incidenti che coinvolgono due veicoli riguarda scontri frontali o laterali e tamponamenti, inoltre la porzione più consistente dei conducenti del principale veicolo coinvolto (A) procedevano irregolarmente, mentre per il secondo veicolo procedevano regolarmente.

Dai risultati dei test V di Cramer possiamo osservare che i risultati più significativi sono stati ottenuti associando le circostanze di circolazione del veicolo A e del veicolo B con la natura dell'incidente.
Infine, mediante la procedura statistica ANOVA si è osservato che le variabili categoriche che influenzano in modo significativo il numero di soggetti lesionati sono:

- condizioni meteorologiche
- fondo stradale
- natura incidente
- sesso del conducente del veicolo 'a'
- intersezione o non intersezione
- tipologia di patente conducente 'a'
- tipologia di patente conducente 'b'
- tipo veicolo 'a'
- tipo veicolo 'b'