Relazione di campionamento

D'alto Jacopo, Mattazzi Anna Chiara, Quercini Luca, Spinelli Sonia

DESCRIZIONE DEL DATASET

Il dataset oggetto di esame contiene i prezzi degli alloggi Airbnb a Londra nei finesettimana.

La popolazione totale è composta da 5379 alloggi, per ciascuno dei quali sono registrate alcune variabili come il prezzo a notte, la posizione, la valutazione della pulizia, il livello di soddisfazione degli ospiti, il numero di persone ospitabili, il numero di camere totali, il tipo di camera.

La variabile di interesse è `realSum`, che indica il prezzo di listino totale dell'affitto. Il parametro di studio è la sua media. La variabile viene stratificata, in un primo momento, per la variabile qualitativa `room_type`, ovvero il tipo di alloggio (stanza privata, stanza condivisa e appartamento intero), e, successivamente, per `person_capacity`, variabile quantitativa discreta che conta il numero di persone che l'Airbnb può ospitare, da 2 a 6. Si suppone, infatti, che il prezzo dell'Airbnb sia correlato con il tipo di sistemazione e il numero di ospiti. Le due variabili distinguono le unità sperimentali in gruppi, le cui numerosità sono eplicitate nella *figura 1*.

```
> table(room_type)
room_type
Entire home/apt    Private room     Shared room
           2418            2934              27
> table(person_capacity)
person_capacity
   2    3    4    5    6
3316  466 1016  207  374
```

Figura 1:Numerosità per livelli delle variabili `room_type` e `person_capacity`

```
> summary(realSum)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   54.33   174.51   268.12   364.39   438.27 12937.27
> mediaTot
[1] 364.3897
> stddevTot
[1] 437.7425
```

Figura 2: Summary su tutta la popolazione della variabile `realSum`.

Osservando la *figura 2*, si deduce che, sull'intera popolazione, la variabile `realSum` assume valori in un range molto vasto. In particolare, i dati si concentrano nel range [0-1000], come si nota dal grafico in *figura 3*. La mediana è pari a €268.12, mentre il prezzo medio per alloggio è di €364.39. L'asimmetria positiva si osserva specialmente dalla *figura 4*, dove si registra un picco per alloggi che hanno prezzo compreso tra 100 e 200 €. Poichè i dati si discostano molto dal valore medio e la standard deviation risulta essere molto alta, €437.74, si deduce la presenza di outliers alti.

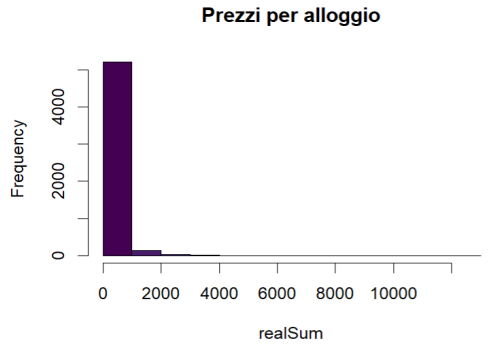

Figura 3: Istogramma della distribuzione di `realSum`.

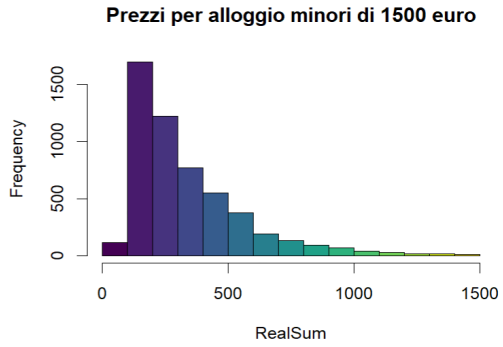

Figura 4: Istogramma di `realSum per prezzi minori di €1500`.

SCELTA DELLA TAGLIA CAMPIONARIA

Si effettua un campionamento casuale semplice senza reinserimento per diverse taglie campionarie, da 250 a 2500 con passo pari a 50.

Il valore della media della variabile `realSum` dipende dalla taglia campionaria e dal campione stesso. Come si nota dai picchi del grafico in *figura 5*, la distorsione è dovuta principalmente dalla presenza di outliers nel campione, ma tendenzialmente, sembra essere tanto più marcata quanto la numerosità è bassa.

A partire da queste informazioni, però, non si può concludere su quale sia la taglia migliore. Al fine di scegliere con più cognizione, oltre alla media, è stata considerata la standard deviation, osservata sulla variabile di interesse al variare di n.

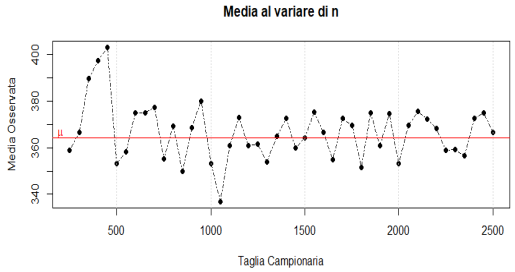

Figura 5: Medie di `realSum`

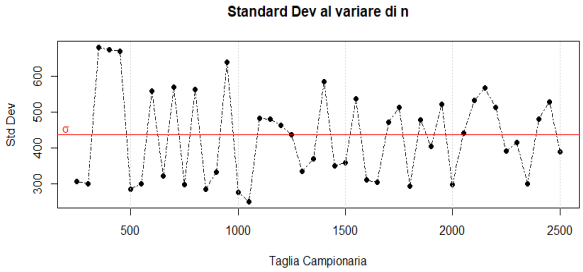

Figura 6: Deviazione standard di `realSum`

Come si può osservare dalla *figura 6*, per taglie campionarie basse la deviazione standard assume valori estremi.

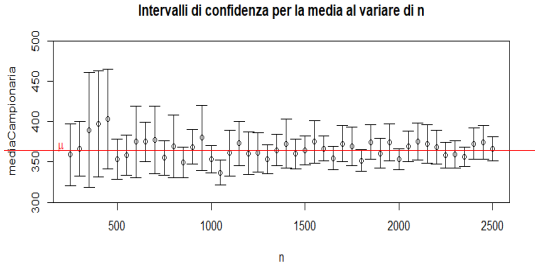

Figura 7: Intervalli di confidenza di `realSum`

Da 1000 in poi, invece, questa si stabilizza intorno al valor medio reale. A conferma di ciò, nella *figura 7*, si nota che da 1000 in poi l'ampiezza degli intervalli di confidenza con fiducia al 95% si restringe. Si può ipotizzare che, scegliendo 1000 come taglia campionaria, potremmo ottenere una rappresentatività del campione simile a quella di uno di taglia più alta, contenendo i costi.

CAMPIONAMENTO STRATIFICATO

ALLOCAZIONE PROPORZIONALE

Si procede con la stratificazione per la variabile `room_type` con allocazione proporzionale. Per un campione di numerosità 1000, lo schema campionario ha probabilità di inclusione del primo ordine pari a 18%. Media e standard deviation sono stimate per strato, in particolare, per l'ultima è utilizzato lo stimatore corretto S, radice di S^2.

Si osserva dalla *figura 8* che per `Entire home/apt` la media è quasi il doppio rispetto alle altre due: ragionevolmente un intero appartamento o un intera casa, in media, costa di più di una stanza. D'altra parte sembrerebbe che una stanza condivisa abbia un costo maggiore rispetto a una privata.

Ciò può essere spiegato dalla bassa taglia campionaria per lo strato `Shared Room` e dalla possibile presenza di outliers.

	Entire home/apt	Private room	Shared room
Numerosità	450.00000	545.00000	5.00000
Percentuale campionata	18.61042	18.57532	18.51852
Media	517.86783	215.64758	297.32590
Standard Deviation	353.15738	167.39208	197.03801

Figura 8: Output stratificazione proporzionale per `room_type`

	2	3	4	5	6
Numerosità	616.0000	87.00000	189.00000	38.00000	70.00000
Percentuale campionata	18.5766	18.66953	18.60236	18.35749	18.71658
Media	252.4475	329.49435	510.74086	533.77376	834.73271
Standard Deviation	167.2502	215.44794	291.51886	367.71825	566.46991

Figura 9: Output stratificazione proporzionale per `person_capacity`

Si effettua lo stesso campionamento per la variabile `person_capacity` (*figura 9*).

	stime	varianze	sd
Room_type	351.9134	1683863590	41034.91
Person_capacity	359.2218	1524280727	39042.04

Figura 10: Confronto tra stratificazioni

Per le due stratificazioni si stima la media e la varianza nel campione. Entrambe sono sottostime per il valore vero di €364.39, ma la media per `person_capacity` è meno distorta.

Inoltre, paragonando le standard deviation, si ottiene un errore inferiore nella seconda stratificazione.

ALLOCAZIONE DI NEYMAN

Si propone, quindi, un'allocazione di Neyman stratificando per `person_capacity`. La deviazione standard in ciascuno strato è richiesta per calcolare la taglia campionaria di ogni livello. Al fine di non ricorrere al valore reale tramite i dati censuari, è stimata dal campione stratificato in allocazione proporzionale.
La probabilità di inclusione del primo ordine è diversa per ogni strato. Per il campionamento osservato, la percentuale campionaria è maggiore all'aumentare della capacità della sistemazione. Si può notare una correlazione tra `realSum` e `person_capacity`: all'aumentare della capacità dell'alloggio il prezzo medio aumenta.

	2	3	4	5	6
Numerosità	448.00000	81.00000	239.00000	61.0000	171.00000
Percentuale campionaria	13.51025	17.38197	23.52362	29.4686	45.72193
Media	240.92538	331.36675	511.20439	546.0278	780.08057
Stardard Deviation	133.15705	164.86222	315.15683	299.3959	706.07635

Figura 11: Output stratificazione di Neyman per `person_capacity`

Dal confronto in *figura 12*, è evidente che con le due allocazioni non ci sono forti distorsioni nelle stime della media. In *figura 13* sono paragonate le deviazioni standard dei due campionamenti. Con Neyman si nota che l'errore è minore, ad eccezione dell'ultimo strato, il più campionato, ma anche il più soggetto a eventuali outliers.

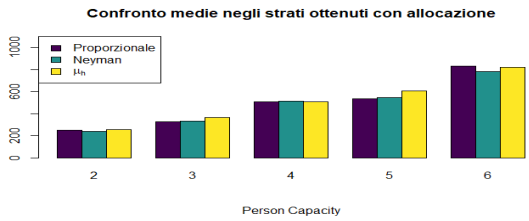

Figura 12: Medie di `realSum` stratificato per `person_capacity`

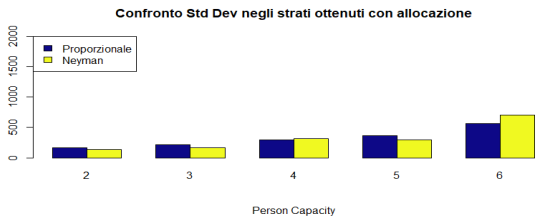

Figura 13: Deviazione standard di `realSum` stratificato per `person_capacity`

	stime	varianze	sd
Proporzionale	359.2218	1524280727	39042.04
Neyman	349.0402	941597066	30685.45

Figura 14: Confronto tra stratificazioni

Dai risultati delle stratificazioni in *figura 14*, si ottengono sottostime di μ. Paragonando le due allocazioni, quella proporzionale presenta una sottostima migliore, ma un errore, rappresentato dalla deviazione standard, maggiore, mentre con Neyman si ha una stima più distorta, però con varianza minore.