Bonanomi Paolo, matricola n°1166678

Romeo Silvestri, matricola n°1169828

Maurizio Nicolaio, matricola n°1172560

REPORT COVID-19 – MODELLI STATISTICI APPLICATI

Partendo dal dataset sull'andamento del Covid-19, si è deciso di analizzare l'andamento del virus nel periodo di riferimento Gennaio-Febbraio 2020 in tre diversi paesi asiatici: Giappone, Corea del Sud e Hong Kong.
L'obiettivo dell'analisi consiste nello studio dell'impatto del Covid-19 in alcuni paesi rappresentativi vicini alla Cina nei mesi iniziali dell'epidemia al fine di riuscire a comprendere l'evolversi del virus nelle diverse classi d'età e genere dei soggetti e la conseguente risposta dei suddetti stati.
Dal dataset originario si prendono in considerazione le variabili: country, gender, age, symptom_onset, death e recovered.
L'evento d'interesse per lo studio corrisponde al decesso dei soggetti presi in analisi, di conseguenza risulta essere un evento che non si verifica nella totalità dei casi e non sempre viene osservato.
Al fine quindi di analizzare le osservazioni presenti come dati di sopravvivenza, viene stabilito che la variabile symptom_onset è volta a rappresentare la data di entrata nel campione di ciascun soggetto, death rappresenta l'evento (o la data in cui avviene l'evento) e recovered indica infine la guarigione del soggetto e la conseguente uscita dal campione (o la relativa data in cui avviene).
In particolare si assume che alla variabile symptom_onset si attribuisca un valore indicativo anche nel caso in cui il soggetto non manifesti sintomi e risulti quindi asintomatico; tale assunzione viene proposta per potere assegnare a tutti i soggetti del dataset una specifica data di entrata nello studio.
Per semplificare la trattazione si decide inoltre di creare un nuovo dataset con variabili costruite a partire da quelle originarie:
- id : codice identificativo del soggetto
- giorni : numero di giorni intercorsi tra l'entrata del soggetto nello studio, coincidente con symptom_onset, e la manifestazione dell'evento o l'eventuale uscita dallo studio, rappresentate a seconda dei casi da death, recovered o dalla data comune di fine studio (29 febbraio 2020 alle ore 23:59)
- age : età del soggetto alla data in cui è stato riportato il caso di coronavirus
- country : paese in cui viene registrato il soggetto
- status : variabile dicotomica che assume il valore 1 nel caso in cui il soggetto abbia sperimentato l'evento e il valore 0 nel caso in cui il soggetto sia uscito dallo studio a causa dell'avvenuta guarigione o per la fine dello studio stesso
- gender : genere del soggetto

E' da evidenziare che l'uscita di un soggetto dallo studio viene qui trattata come un caso di censura a destra.

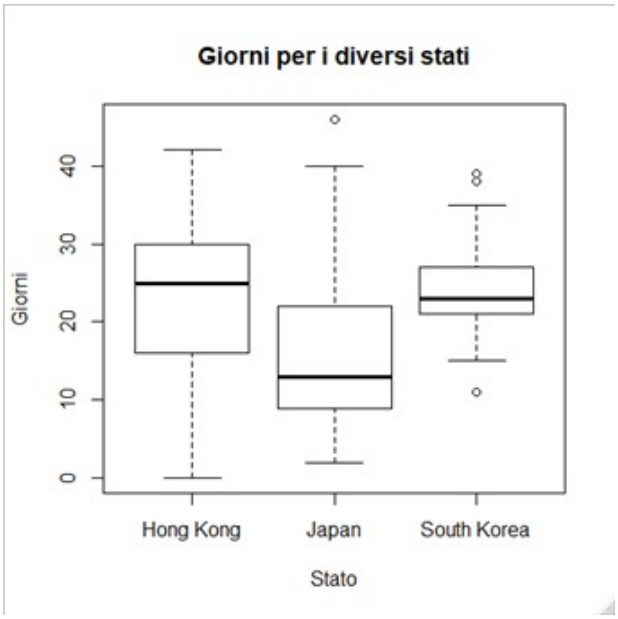
DATI MANCANTI

Il dataset originario presenta numerosi dati mancanti per diverse variabili prese in considerazione.

A tal proposito si sceglie di applicare delle procedure inferenziali basate sulla teoria della probabilità per riuscire ad assegnare un valore ad ogni variabile per ciascun soggetto su cui non è stato precedentemente riportato il dato.
Le assegnazioni vengono effettuate in maniera separata per ognuno dei tre paesi, in modo tale da tenere in considerazione le differenti specificità.
Per la variabile gender si assegna casualmente un valore (male o female) a seconda della probabilità di estrarre uno dei soggetti tra quelli su cui la variabile è stata già osservata; nello specifico per ciascun paese la probabilità di estrazione sarà ponderata in base alla frequenza relativa di ognuno dei due generi sul totale dei soggetti osservati.
L'assenza di valori per le variabili age e symptom_onset è trattata simultaneamente tramite la funzione mice() del software R.
La funzione mice() implementa delle procedure di tipo bayesiano per la creazione di valori sostitutivi per dati mancanti multivariati e sfrutta la specificazione completamente condizionale in cui ogni variabile incompleta è imputata da un modello separato.
Si nota in aggiunta che per lo stato di Hong Kong due soggetti non presentano una data per la variabile death nonostante la manifestazione dell'evento e lo stesso vale per un ulteriore soggetto con la variabile recovered.
Per ovviare al problema si ricavano le corrispondenti date visionando i siti web dei giornali riportati alla variabile link dello stesso dataset.

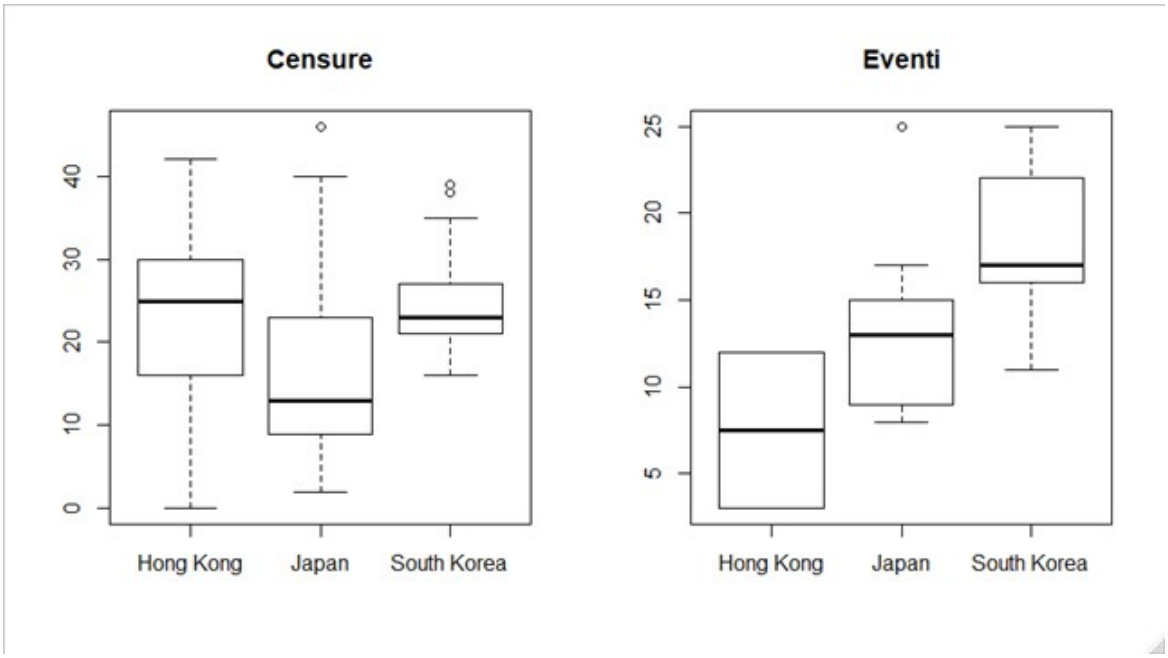
ANALISI DI SOPRAVVIVENZA

Si vuole effettuare un'analisi esplorativa per visualizzare il numero di giorni in cui le unità statistiche sono sotto osservazione prima di sperimentare l'evento o uscire dallo studio.
A tal fine si costruiscono tre boxplot: il primo indica i giorni divisi per ciascuno stato, mentre gli altri riguardano i giorni necessari prima di sperimentare l'evento o la censura (guarigione o fine studio).



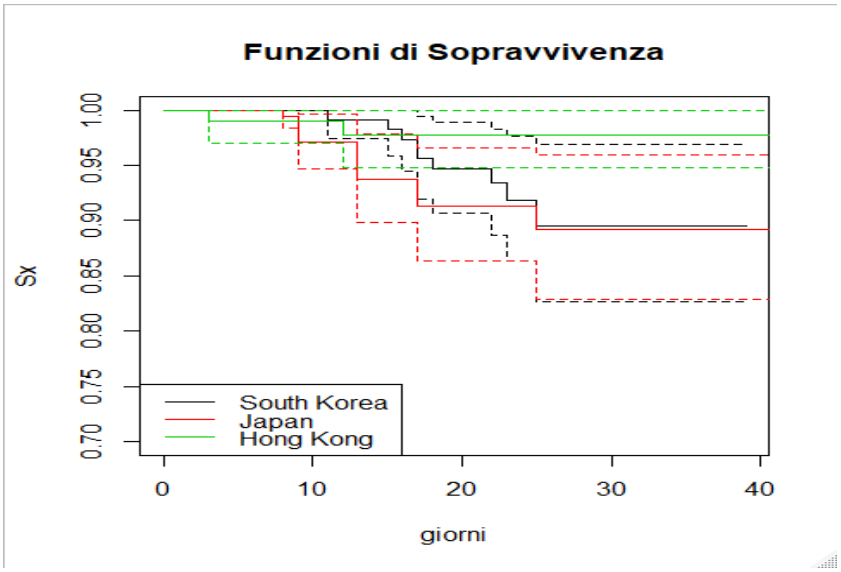
Dai boxplot si nota che le distribuzioni dei giorni dei tre stati sono diverse fra loro.
In particolare Hong Kong presenta una leggera asimmetria a sinistra mentre Giappone e Corea del Sud presentano un'asimmetria a destra, molto più marcata nel Giappone.
Si nota, inoltre, che i giorni mediani del Giappone sono molto inferiori rispetto alle controparti asiatiche.

Quindi questo significa che in Giappone le unità che entrano nello studio escono molto più velocemente rispetto agli altri due stati o in qualità di censure o perché si è verificato l'evento.
Si procede quindi a costruire i boxplot dei tre paesi relativi sia alle censure che agli eventi.

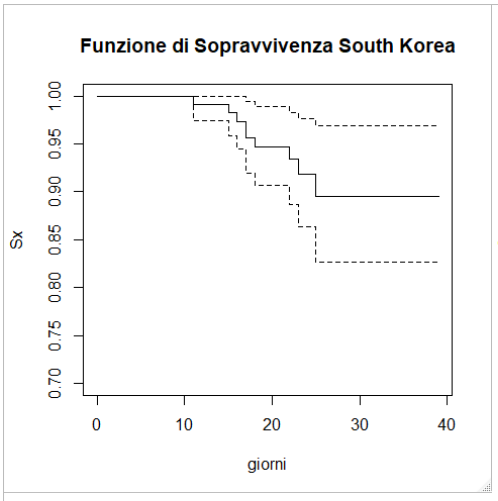


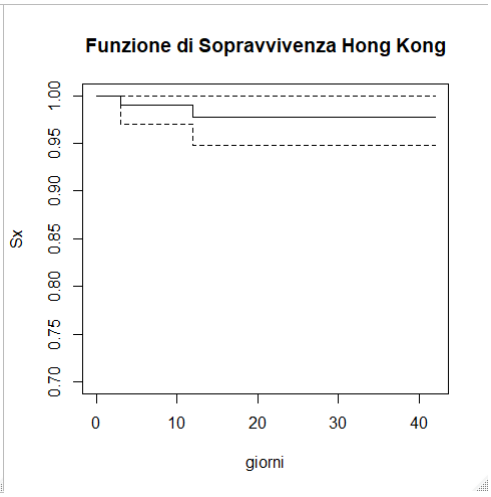
Si può notare che le distribuzioni delle censure sono molto simili alle precedenti dato che il numero di eventi è molto basso.
Dalle distribuzioni degli eventi invece possiamo notare che gli eventi ad Hong Kong si verificano molto presto rispetto alle altre due, inoltre, possiamo notare che gli eventi in Corea del Sud si verificano molto più lentamente rispetto ad Hong Kong e al Giappone, questo potrebbe significare che il sistema sanitario della Corea del Sud è più efficiente rispetto alle altre due.
Tuttavia, si deve tenere conto che il numero di eventi è basso e quindi bisognerà fare attenzione nelle analisi future.

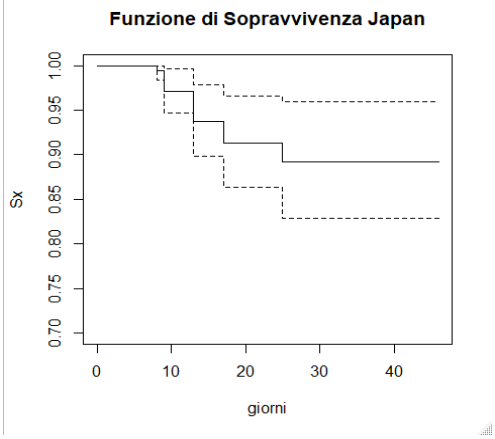
FUNZIONI DI SOPRAVVIVENZA PER I VARI PAESI



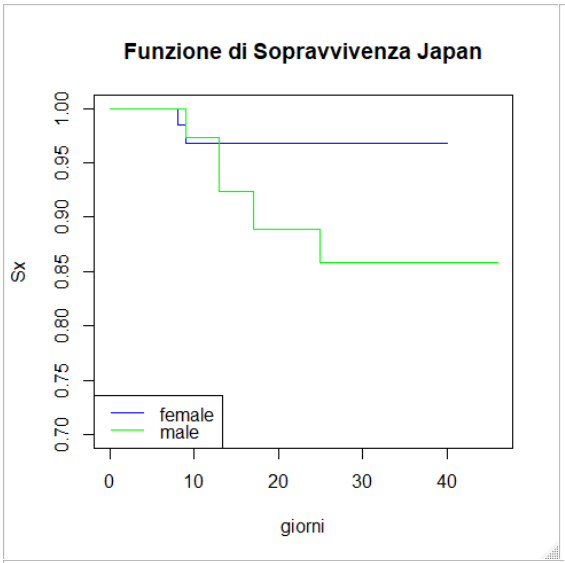
Si ottengono le funzioni di sopravvivenza attraverso lo stimatore di Kaplan-Meier e i rispettivi intervalli di confidenza sono calcolati con un livello di significatività pari a 0.95.
Da questo grafico si può notare una differenza molto marcata tra la funzione di sopravvivenza di Hong Kong e le controparti asiatiche.
La differenza è, probabilmente, dovuta al minor numero di eventi morte verificatisi nel paese in confronto agli altri due, ma questo potrebbe essersi verificato anche per via del minor numero di unità statistiche.
Le funzioni di sopravvivenza di Giappone e Corea del Sud appaiono invece sufficientemente simili.
Per vedere se si può assumere che ci siano differenze tra le varie curve si applica il test dei ranghi logaritmici, con ipotesi nulla nella quale le tre funzioni di sopravvivenza vengono supposte uguali.
Il p value del test è pari a 0.1: non possiamo quindi affermare che le tre funzioni di sopravvivenza siano differenti, ma il p value risulta comunque poco elevato e se ne terrà in considerazione in fase di conclusioni.

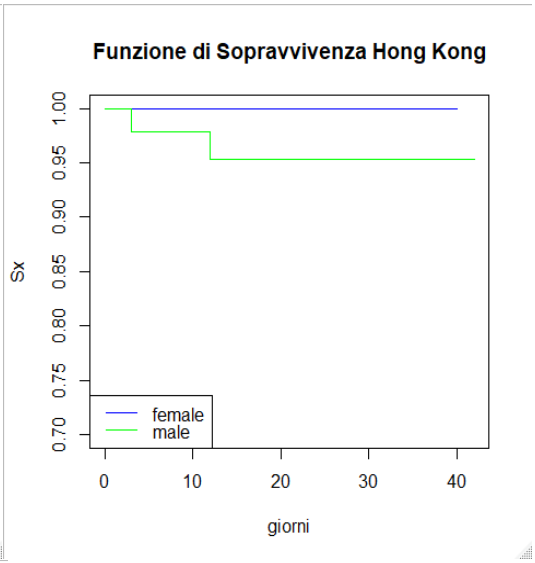


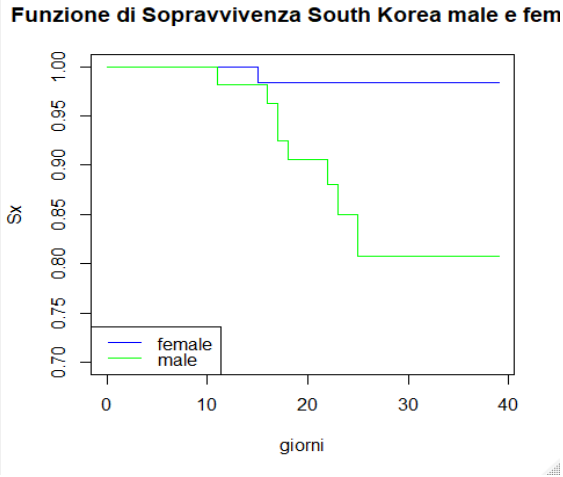




CONFRONTO DELLE FUNZIONI DI SOPRAVVIVENZA IN FUNZIONE DEL GENERE

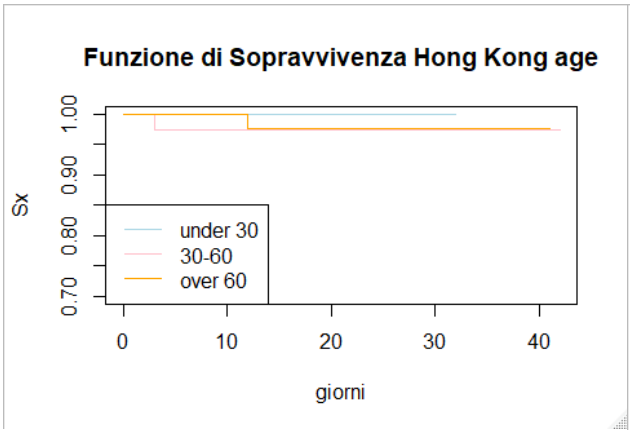


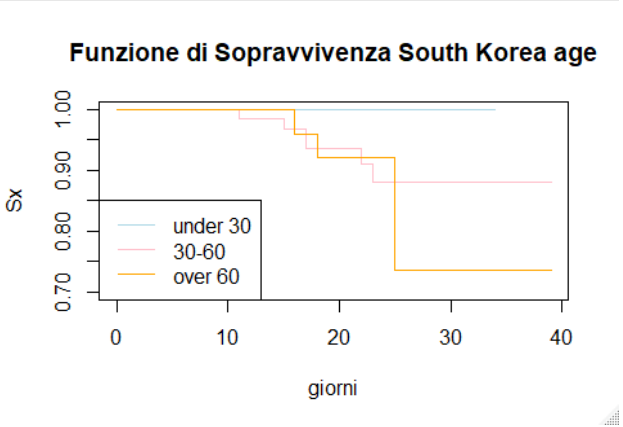


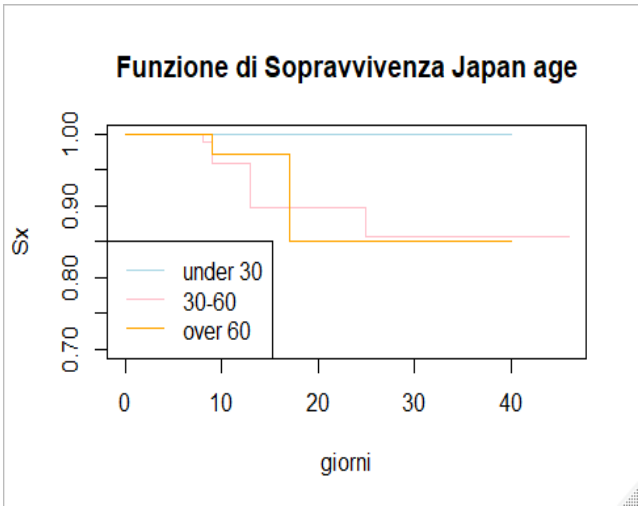


Si cerca di capire se ci sia una differenza nel comportamento della popolazione maschile e di quella femminile nei vari stati.
In contrasto con l'esempio precedente in cui i vari stati si distinguevano per numero di unità statistiche, in questo caso il numero di maschi e quello di femmine nei singoli paesi è distribuito senza differenze significative come nel caso precedente tranne per il Giappone. Per inciso le femmine nella Corea del Sud rappresentano il 52,6% della popolazione, ad Hong Kong sono il 51% ed in Giappone il 36% circa.
In questo caso vediamo una differenza sostanziale tra le due funzioni di sopravvivenza stratificate per genere in Corea del Sud e in Giappone, ma vista la percentuale molto diversa tra maschi e femmine, in questo secondo caso, potrebbe essere dovuta al numero di unità statistiche.
Anche in questo caso si applica il test dei ranghi logaritmici per vedere se all'interno dei singoli stati si registrano degli scarti nelle funzioni di sopravvivenza in base al genere. Il p value del test ad Hong Kong è 0.1, in Giappone vale 0.2 e in Corea del Sud 0.01.
Si rifiuta l'ipotesi nulla sull'uguaglianza delle funzioni di sopravvivenza solo per la Corea del Sud, come si era previsto.

CONFRONTO DELLE FUNZIONI DI SOPRAVVIVENZA IN FUNZIONE DELL' ETÀ







Per semplificare l'analisi e renderla più comprensibile si decide di raggruppare le età in tre classi: una che raccoglie gli under 30, una che comprende i soggetti tra i 30 e i 60 anni di età e infine una classe per gli over 60.
Si segnala che le età della popolazione giapponese sono state registrate solo come multipli di 5, di conseguenza questo metodo di raggruppamento dovrebbe essere in grado di evitare, almeno in parte, l'errore generato da questo tipo di raccolta dati.
In questo caso ci si aspetta una differenza nella funzione di sopravvivenza tra le varie classi di età nei vari paesi in quanto è stato dimostrato a livello medico che il Covid-19 sia più dannoso per determinate classe di età.
Risulta di particolare interesse valutare se nelle due classi di età più anziane si verifichi o meno questa ipotesi, in quanto dai grafici risulta piuttosto chiara l'evidenza che la classe più giovane non abbia subito particolarmente gli effetti del coronavirus.
Utilizziamo come consuetudine il test dei ranghi logaritmici: si nota però che non si rifiuta l'ipotesi nulla di differenza delle funzioni di sopravvivenza in nessuno dei tre stati asiatici.
Il p value del test della Corea del Sud è infatti 0.2, il p value per il Giappone è 0.1 e quello per Hong Kong è 0.9.
Questa risposta è particolarmente stupefacente in quanto va in contrasto con quello che ci aspettavamo in partenza.
Andiamo perciò a fare i confronti a coppie di età per verificare dove ci sia particolare

similitudine. Nei test a coppie di classi di età si verifica una differenza significativa (p value =0.04) tra la classe under 30 e la classe over 60 nella Corea del Sud. Anche in Giappone questo test risulta appena significativo se fatto tra la classe di età under 30 e la over 60 (p value=0.08).
Ad Hong Kong le differenze tra le funzioni di sopravvivenza sono non significative per tutti i confronti fra le classi di età, come d'altronde si può notare dal grafico.

CONSIDERAZIONI CONCLUSIVE

Si può quindi notare che i tre stati non si comportano in maniera particolarmente diversa di fronte all'emergenza covid attraverso l'analisi dei dati di sopravvivenza.
La Corea del Sud sembra essere lo stato in grado di rispondere in modo particolarmente efficiente in termini di giorni passati tra inizio sintomi e decesso dei soggetti per questi primi mesi di epidemia, anche se risulta svantaggiata la popolazione maschile e la classe over 60.
In Giappone non si rileva nulla di particolarmente significativo a livello di differenze per età o genere, con eccezione della classe over 60 che anche in questo caso si presenta come la più fragile.
Hong Kong invece risulta il paese con meno decessi, anche a fronte di un numero di unità statistiche inferiore rispetto alla Corea del Sud ed in particolare al Giappone.
Da sottolineare anche che nonostante Hong Kong presenti in termini assoluti meno casi di coronavirus rispetto agli altri due paesi, in termini relativi (basandosi sul rapporto numero di casi su numero di abitanti) risulta nettamente lo stato più colpito dal virus tra quelli presi in analisi.