Autores:

- Dr. Alvarez, Agustín: Profesor adjunto con dedicación exclusiva del Instituto de Ciencias, UNGS. mail: agalvarez@campus.ungs.edu.ar
- Mg. Fragalá Marina: Jefe de trabajos prácticos con dedicación exclusiva, Instituto de Ciencias, UNGS y Jefe de trabajos prácticos parcial de la FCEyN, U.B.A. mail: mfragala@campus.ungs.edu.ar
- Dra. Valdora, Marina: Profesora adjunta del Departamento de Matemáticas de la FCEyN de la UBA e Investigadora Asistente del Conicet. Mail: mvaldora@gmail.com

Autor responsable: Agustín Alvarez.

1. Introducción

Una de las preguntas más importantes cuando surge una nueva enfermedad infecciosa como la Covid 19 es poder saber cuan mortal es la enfermedad. O sea, saber qué proporción de personas va a morir entre aquellas que se contagian la enfermedad en un período de tiempo y una región determinada. A dicha proporción se la denomina Tasa de Letalidad sobre infectados (TLI). El poder calcular dicha tasa en pleno brote de la enfermedad tiene diversas dificultades. Una de ellas es: conocer el total de personas contagiadas, ya que puede haber muchos casos que no son detectados por no ser testeados, por ejemplo, los casos asintomáticos (en tal caso, el total de personas infectadas es desconocido y mayor a la cantidad de casos confirmados). Otra dificultad es el desfasaje temporal que hay entre el contagio de la enfermedad y el fallecimiento por esta causa. En este trabajo nos vamos a centrar en este segundo problema. La Tasa de Letalidad sobre casos confirmados (TLC), se suele definir como la proporción de personas con el virus confirmado que mueren a causa del virus. La TLC es lo que se informa diariamente (en realidad se aproxima) durante la epidemia de Covid 19. El número TLC que se informa por lo general subestima el verdadero valor ya que el cálculo que se hace diariamente consiste en dividir la cantidad de personas fallecidas registrados por Covid, hasta ese día, por la cantidad total de personas confirmadas hasta ese día. El sesgo puede ser grande y más alto aún si la cantidad acumulada de contagios crece velozmente. En pleno brote de la epidemia se pueden duplicar los casos en pocos días, sin embargo, entre los casos recientes solamente una baja proporción fallece en los primeros días desde la detección, y por lo tanto una alta proporción de estos casos recientes, que fallecerán por Covid, no son considerados en el numerador del cálculo de la TLC porque aún no fallecieron. Por ejemplo, supongamos para fijar ideas, la siguiente situación ficticia: si hasta ayer teníamos 80 casos confirmados

en total y hoy se suman 20 nuevos casos, contabilizando al día de hoy un total de 5 personas fallecidas entre los 100, se informará (hoy) una TLC de 5%. Si de estos últimos 20 casos supiéramos que efectivamente van a fallecer un total de 3 y que de los 80 anteriores aún falta morir una persona más, entonces la TLC en realidad sería del 9% y por lo tanto superior a la que se informa hoy. El objetivo será estimar cuántas personas faltan morir entre los casos confirmados de Covid hasta un determinado día para corregir la tasa que se informa dicho día. En este trabajo proponemos una definición formal de la TLC y un estimador de esta tasa que intenta corregir los problemas descriptos y resulta insesgado (sin sesgo).

En la Sección 2 detallamos la relación de este Capítulo con los contenidos de la escuela secundaria. En la Sección 3 introducimos el modelo y presentamos los estimadores propuestos. En la Sección 4 presentamos los resultados de un estudio de simulación computacional. En la Sección 5 ponemos a prueba los estimadores en el ejemplo de datos reales de la Argentina. En la Sección 6 damos recursos e ideas para la enseñanza. Finalmente, en la Sección 7 entregamos las conclusiones y las tareas pendientes. Como versión simplificada de este Capítulo puede leerse simplemente la introducción y la sección de Recursos para la enseñanza.

2. Relación con contenidos de la Escuela Secundaria

Este capítulo se relaciona con los siguientes Núcleos de Aprendizaje Prioritarios elaborados por el Ministerio de Educación de la Nación en el año 2012.

- La interpretación y uso de nociones básicas de estadística para estudiar fenómenos, comunicar resultados y tomar decisiones.
- El reconocimiento y uso de nociones de probabilidad para cuantificar la incertidumbre y argumentar en la toma de decisiones y/o evaluar la razonabilidad de inferencias.

En particular, se relaciona con los siguientes ejes:

- La interpretación y elaboración de información estadística en situaciones problemáticas que requieran:
  - identificar diferentes variables (cualitativas y cuantitativas), organizar los datos y construir gráficos adecuados a la información a describir (1ro y 2do año);
  - organizar datos para estudiar un fenómeno y/o tomar decisiones analizando el proceso de relevamiento de los datos y los modos de comunicar los resultados obtenidos (2do y 3er año).
- El análisis del problema/ fenómeno a explorar, lo que supone (para 3er y 4to año):

- delimitar las variables de estudio y la pertinencia de la muestra,
- seleccionar las formas de representar,
- comunicar los datos acordes a la situación en estudio.

3. Definiciones principales y propuestas de estimación

3.1. Definiciones

La letalidad sobre personas confirmadas puede variar en el tiempo y depende de muchos factores. Por ejemplo: durante la epidemia podría corregirse la definición de caso confirmado, la cantidad de testeos podría variar con el tiempo (a partir de un momento comenzar a testear masivamente), o podría aparecer algún tratamiento que morigere la letalidad.

Consideraremos:

- d un número entero y no negativo (0, 1, 2, …). El número 0 representa el día del primer contagio confirmado;
- $c(d)$ la cantidad de personas a los que se les confirma el virus durante el día d;
- $L(d)$ la probabilidad de que una persona confirmada durante el día d, muera a causa del virus;
- $M_{d,i}$ es una variable que toma valor 1 si la i-ésima persona confirmada del día d muere a causa del virus y 0 en caso contrario (se recupera). Notar que el índice i toma valores enteros de 1 a $c(d)$,
- $M_{d,i}(t)$ que vale 1 si la i-ésima persona confirmada del día d muere a causa del virus el día t o antes y 0 en caso contrario ($d \leq t$),
- $M_{t,f}$ la cantidad de casos confirmadas hasta el día t que finalmente morirán por el virus,
- $M(t)$ la cantidad de casos registrados de muertes a causa del virus hasta el día t inclusive (entre las personas confirmadas hasta el día t).

Notar que

$$M_{t,f} = \sum_{d=0}^{t} \sum_{i=1}^{c(d)} M_{d,i}$$

y

$$M(t) = \sum_{d=0}^{t} \sum_{i=1}^{c(d)} M_{d,i}(t).$$

Ampliemos por un momento el concepto de la variable $M_{d,i}$ y $M_{d,i}(t)$. En general, las variables que toman valor 1 si ocurre un evento de interés (en este caso la muerte de la *i*-ésima persona confirmada el día d para el caso de $M_{d,i}$) con cierta probabilidad *p* y toman valor 0 si no ocurre el evento de interés con probabilidad 1-*p,* se llaman variables aleatorias con distribución Bernoulli de parámetro *p*. Estas variables aleatorias son muy utilizadas en todo el ámbito de la Estadística. Son variables dicotómicas porque registran dos únicas posibilidades que suelen denotarse como éxito o fracaso, pero por comodidad en los cálculos se utilizan 1 y 0 respectivamente. Por lo tanto, la esperanza de $M_{d,i}$ se calcula como sigue:

$$E\left(M_{d,i}\right) = 1.P\left(M_{d,i} = 1\right) + 0.P\left(M_{d,i} = 0\right) = L(d).$$

Todas las tasas de letalidad que definiremos a continuación pueden considerarse tasas de letalidad acumuladas porque el período de tiempo que consideran abarca desde el inicio de la epidemia hasta un cierto día t. La Tasa de Letalidad entre casos confirmados TLC (que se informa cada día en Covid por ejemplo) hasta el día *t* está dada por

$$TLC(t) = \frac{M(t)}{\sum_{d=0}^{t} c(d)}.$$

Sin embargo, la letalidad acumulada hasta el día t es razonable definirla como la probabilidad de fallecer por Covid una persona que fue confirmada el día t o antes. También se puede interpretar como la proporción de personas que es de esperar que mueran entre los casos confirmados hasta el día t. De hecho, definimos la Tasa de Letalidad esperada entre las personas confirmadas hasta el día t del siguiente modo:

$$TLC_{t,e} = \mathrm{E}\left(\frac{M_{t,f}}{\sum_{d=0}^{t} c(d)}\right). \qquad (1)$$

Vale aclarar que $TLC_{t,e}$ es un valor poblacional (es un valor teórico) y no observable, mientras que $\mathrm{TLC}(t)$ es un valor observable.

Dado que

$$M_{t,f} = \sum_{d=0}^{t} \sum_{i=1}^{c(d)} M_{d,i}$$

y utilizando la propiedad de linealidad de la esperanza se deduce que

$$TLC_{t,e} = \frac{\sum_{d=0}^{t} c(d)L(d)}{\sum_{d=0}^{t} c(d)} = \sum_{d=0}^{t} \frac{c(d)}{\sum_{d=0}^{t} c(d)} L(d) = \sum_{d=0}^{t} \omega(d)L(d),$$

obteniendo una suma ponderada de las letalidades diarias, donde la ponderación de cada día es la proporción de personas confirmadas en dicho día respecto al total acumulado. Observar, como caso particular que si $L(d) = L$ entonces $TLC_{t,e} = L$.

Para poder estimar $TLC_{t,e}$ vamos a definir nuevas variables relativas al tiempo que demoran en morir las personas confirmadas. Nos interesaremos en el tiempo que demoran en morir, desde que se les detecta el virus, aquellas personas que finalmente mueren por la enfermedad. Sean:

- $T_{d,i}$ la cantidad de días que transcurren desde que se confirma el diagnóstico hasta que muere, la *i*-ésima persona infectada del día *d* y
- F_d a la distribución acumulada de estos tiempos $T_{d,i}$.

O sea, $F_d(k)$ representa la probabilidad de que una persona que se le confirmó el virus el día d y que va a morir por la enfermedad, muera en k días o menos.

Parece razonable suponer que la distribución F_d puede ir variando en el tiempo por diversas razones. Podría ocurrir por ejemplo, que un nuevo tratamiento o la detección más temprana, alarga el tiempo de sobrevida de las personas infectadas, o que el sistema de salud satura y se acorta el tiempo de sobrevida, entre otros motivos.

Definimos anteriormente para cada $d \leq t$ y para $1 \leq i \leq c(d)$ la variable aleatoria Bernoulli $M_{d,i}(t)$ que vale 1 si la *i*-ésima persona confirmada del día *d* muere por Covid el día t o antes y 0 en caso contrario. Tenemos entonces

$$P\big(M_{d,i}(t) = 1\big) = P\big(M_{d,i} = 1, T_{d,i} \leq t - d\big) = P\big(M_{d,i} = 1\big)P\big(T_{d,i} \leq t - d \big| M_{d,i} = 1\big),$$

luego

$$P\big(M_{d,i}(t) = 1\big) = L(d)F_d(t - d).$$

De aquí se desprende que, como las $M_{d,i}(t)$ tienen distribución Bernoulli,

$$\mathrm{E}\left(M_{d,i}(t)\right) = \mathrm{L}(d)F_d(t - d). \qquad (2)$$

3.2. Estimadores propuestos.

3.2.1. Nuestra Propuesta:

El desafío es poder predecir $M_{t,f}$ el mismo día t, es decir la cantidad de personas que finalmente morirán entre las personas confirmadas hasta el día t. La idea que proponemos es la siguiente. Sabemos que entre las confirmadas el día d, la proporción que morirán el día t o antes es de esperar que sea $F_d(t-d)$. Por ejemplo, si hace 3 días (día que llamamos d^*) hubo 100 nuevos casos confirmados y hoy sabemos que ya murieron 2 de ellos y además $F_{d^*}(3) = 1/5 = 0{,}2$ deducimos entonces que estas 2 personas fallecidas es de esperar que representen el 20% de los casos confirmados hace 3 días, que morirán por el virus. Por lo tanto 2 x 5 = 2 x $1/F_d(3)$ = 10 son las personas que finalmente se espera que mueran entre los 100 nuevos confirmados del día d^*.

Supongamos que $F_d(t-d)$ es un valor conocido para cada d y t. Luego predecimos $M_{t,f}$ con

$$\widehat{M}_{t,f} = \sum_{d=0}^{t} \left(\sum_{i=1}^{c(d)} M_{d,i}(t) \right) \frac{1}{F_d(t-d)}.$$

Finalmente estimamos la Tasa de Letalidad esperada del siguiente modo:

$$\widehat{TLC}_{t,e} = \frac{\widehat{M}_{t,f}}{\sum_{d=0}^{t} c(d)}. \tag{3}$$

Utilizando nuevamente la propiedad de linealidad de la esperanza y la ecuación (2) se demuestra que

$$E\left(\widehat{TLC}_{t,e}\right) = TLC_{t,e}.$$

Por lo tanto, el estimador $\widehat{TLC}_{t,e}$ resulta insesgado.

El estimador es utilizable tal como está planteado si se conocieran las distribuciones F_d. Poder aproximar las F_d es un problema de interés en sí mismo. En caso de obtener una buena aproximación, el estimador final resultará de reemplazar $F_d(t-d)$ por dicha aproximación en la ecuación (3).

3.2.2. Propuesta de Garske et al.

Vamos a comparar el estimador propuesto arriba con el estimador que introdujo Garske et al. (2009). En dicho trabajo no contemplan que las distribuciones F_d puedan variar en función de d y plantean una distribución constante $F_d = F$ cualquiera sea d. Del mismo modo tampoco tienen en cuenta que

las probabilidades $L(d)$ puedan ir variando con respecto al día de confirmación d, es decir $L(d) = L$, cualquiera sea d. Sabemos que $M(t)$ es la variable aleatoria que cuenta cuántas personas fallecieron por el virus desde el comienzo de la epidemia, hasta el día t incluido. El estimador que proponen se basa en la siguiente idea: con estas condiciones de que la letalidad $L(d)$ sea constantemente L y las distribuciones F_d sean constantemente F, se puede calcular fácilmente la esperanza de $M(t)$.

$$\mathrm{E}\big(M(t)\big) = \mathrm{E}\left(\sum_{d=0}^{t}\sum_{i=1}^{c(d)} M_{d,i}(t)\right) = L\sum_{d=0}^{t} c(d)F(t-d).$$

Luego, plantean como estimador de L al día t,

$$\widehat{L_{\mathrm{t}}} = \frac{M(t)}{\sum_{d=0}^{t} c(d)F(t-d)}.$$

Observar que $\mathrm{E}\big(\widehat{L}_{\mathrm{t}}\big) = \mathrm{L}$.

Adecuando su estimador a la posibilidad de cambio de las F_d y las $L(d)$, proponemos:

$$\widetilde{TLC}_{t,e} = \frac{M(t)}{\sum_{d=0}^{t} c(d)F_d(t-d)},$$

pudiendo calcular la esperanza del siguiente modo:

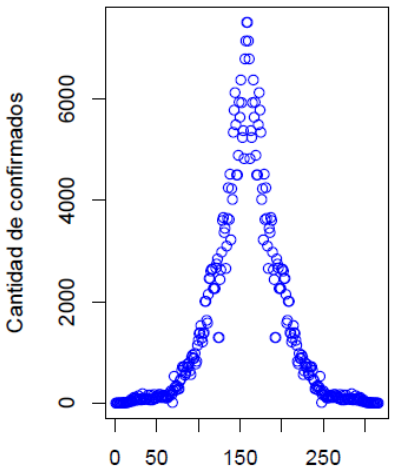
$$\mathrm{E}\big(\widetilde{TLC}_{t,e}\big) = \sum_{d=0}^{t} \frac{c(d)F_d(t-d)}{\sum_{d=0}^{t} c(d)F_d(t-d)}\mathrm{L}(d) = \sum_{d=0}^{t} \omega(d)L(d).$$

Podemos observar que el valor esperado del estimador es una media ponderada de las letalidades diarias, pero la ponderación es distinta a la dada por $TLC_{t,e}$. El estimador resultante no es necesariamente insesgado. Lo sería si $L(d) = L$ para todo d.

4. Estudio de simulación computacional

Hemos propuesto hasta aquí una manera de definir la letalidad sobre casos confirmados de una enfermedad y también hemos considerado un par de propuestas diferentes de cómo estimar dicho valor. Las propuestas, tanto de la definición como de la estimación se basan en un modelo. En nuestro caso el modelo son las variables aleatorias Bernoulli $M_{d,i}$, con parámetro $L(d)$, a través de las cuales se decide si el i-ésimo caso confirmado del día d muere o no. Además, nuestro modelo asume que para aquellas personas que efectivamente morirán y que fueron confirmadas el día d, el tiempo $T_{d,i}$ que demorarán en morir, es una variable aleatoria con distribución F_d. También hemos supuesto independencia tanto entre las $T_{d,i}$ como entre las $M_{d,i}$. Estas suposiciones que hemos hecho permiten

simular por medio de programas computacionales un escenario posible de cómo se desarrollaría una epidemia. Lo que uno necesita es fijar los valores $c(d)$ de la cantidad de personas que se confirman el día d, elegir las letalidades diarias $L(d)$ y fijar las distribuciones de tiempos F_d. Suele ser el primer paso en un estudio estadístico de estas características realizar una simulación para ver cómo se comporta el modelo y los estimadores que se desean analizar. Se suele estudiar en estos casos si en escenarios que parecen sencillos o razonables los estimadores dan buenos resultados, así como compararlos, para ver cuál funciona mejor en cada escenario. En este caso presentamos un ejemplo muy sencillo y realizamos un breve análisis gráfico de cómo resultan los estimadores. A modo pedagógico (luego se entenderá mejor) decidimos plantear un escenario en que la epidemia comienza y termina; llega un momento en que no hay más contagios y todas las personas que tuvieron la enfermedad: o ya fallecieron o ya se recuperaron. Para esto, basándonos en como venía la curva de contagios en Argentina, desde el 3 de marzo de 2020 hasta el 6 de agosto de 2020, planteamos $c(d)$ igual a la cantidad de confirmados el mismo día d. O sea, $c(0) = 1$ es la cantidad de casos confirmados el día 3 de marzo y $c(157) = 7513$ es la cantidad de casos confirmados el día 6 de agosto. Luego espejamos la cantidad de casos confirmados como si a partir del 7 de agosto comenzara a bajar y planteamos $c(158 + d) = c(157 - d)$ para d desde 0 hasta 157 (ver Figura 1).



5.

Figura 1: Curva de casos confirmados diarios asumida en la simulación

De esta manera tenemos que el último caso confirmado en este escenario sería el día 315 desde el primer caso. A partir de fijar estos valores de cantidad de casos confirmados diarios, luego planteamos algunos escenarios. Todos estos escenarios tienen la característica de que la letalidad diaria se mantiene constante $L(d) = p_1$ desde del día 0 hasta un cierto día d^* y; a partir del día $d^* + 1$ hasta el día 315 se mantiene constante $L(d) = p_2$. Para cada simulación elegimos dos valores distintos de p_1 y p_2 en el conjunto $\{0{,}02; 0{,}05; 0{,}10\}$ y un valor d^* en el conjunto $\{70, 120\}$. Para las distribuciones de los

tiempos entre confirmación y muerte entre aquellos que fallecen, planteamos en todos los casos $F_d \equiv F$ para todo d, donde F es una distribución binomial negativa con media $\mu = 10{,}8$ y tamaño $r = 0{,}88$. Usamos esta distribución con estos parámetros porque su función de distribución acumulada aproxima bien a la distribución de los tiempos de fallecimiento observados en la Argentina entre el 3 marzo y el 6 de agosto. A partir de fijar dichos parámetros uno puede armar una matriz de fin de epidemia, donde la i-ésima fila representaría lo que ocurrió con el í-ésimo caso confirmado (en este escenario ficticio). La matriz tiene tres columnas: una primera columna que indica la fecha de confirmación, una segunda columna si la persona falleció o no por Covid y finalmente una tercera columna que indica la fecha en la que falleció. Para las personas que según el sorteo de la simulación no mueren por Covid se les puede asignar una fecha lejana. Luego a partir de dicha matriz que se genera, se pueden calcular los estimadores presentados en el capítulo 2, además de lo que llamaremos Tasa de Letalidad acumulada final que definimos como:

$$\mathrm{T}LC_{t,f} = \frac{M_{t,f}}{\sum_{d=0}^{t} c(d)}. \tag{4}$$

Notar que la cantidad $TLC_{t,f}$ definida en (4) es un valor observable al finalizar la epidemia que no coincide necesariamente con $TLC_{t,e}$ definido en (1), pero es esperable que $TLC_{t,f}$ sea un valor cercano a $TLC_{t,e}$ cuando la cantidad de confirmados hasta el día t es grande. O sea, podemos interpretar que la curva que corresponde a $TLC_{t,f}$ es a la que deseamos acercarnos, ya que en la realidad la curva de $TLC_{t,e}$ nunca la observaremos. Por lo tanto, en el ejemplo de la simulación graficamos $TLC_{t,f}$ por más que en este caso sí tengamos disponible $TLC_{t,e}$ porque los parámetros los elegimos nosotros al armar la simulación.

En todos los escenarios analizados, aquí solo presentamos dos de ellos, observamos (ver Figuras 2 y 3) que tanto el estimador de Garske et al. (2009), como el que proponemos nosotros, dan muy buenas estimaciones en comparación con la tasa que se informa comúnmente. Esta tasa, durante un gran período de la epidemia subestima de una manera grosera los verdaderos valores de la Tasa de Letalidad, comenzando a estimar bien en la etapa final de la misma, cuando la cantidad de nuevos casos es mínima proporcionalmente respecto al total de casos acumulados. Al comparar nuestra propuesta con la de Garske et al. (2009), se puede observar que la nuestra aproxima bastante bien a $TLC_{t,f}$ durante toda la epidemia y tiene una excelente reacción en el día d^*, percibiendo el cambio de la letalidad, mientras que el estimador de Garske et al. (2009) demora en reaccionar en esos días posteriores al día d^*. En

ese sentido, estos ejemplos parecen mostrar un mejor desempeño del estimador que proponemos. Notar además que en la etapa final de la epidemia todos los estimadores se aproximan a la tasa la $TLC_{t,f}$ inclusive a la tasa que se informa diariamente $TLC(t)$. Para visualizar este fenómeno es que decidimos plantear escenarios en los cuales la epidemia finaliza. Para distinguir mejor las curvas, debido a que los primeros días tienen mucha variabilidad, realizamos los gráficos desde el día 15 de la epidemia.

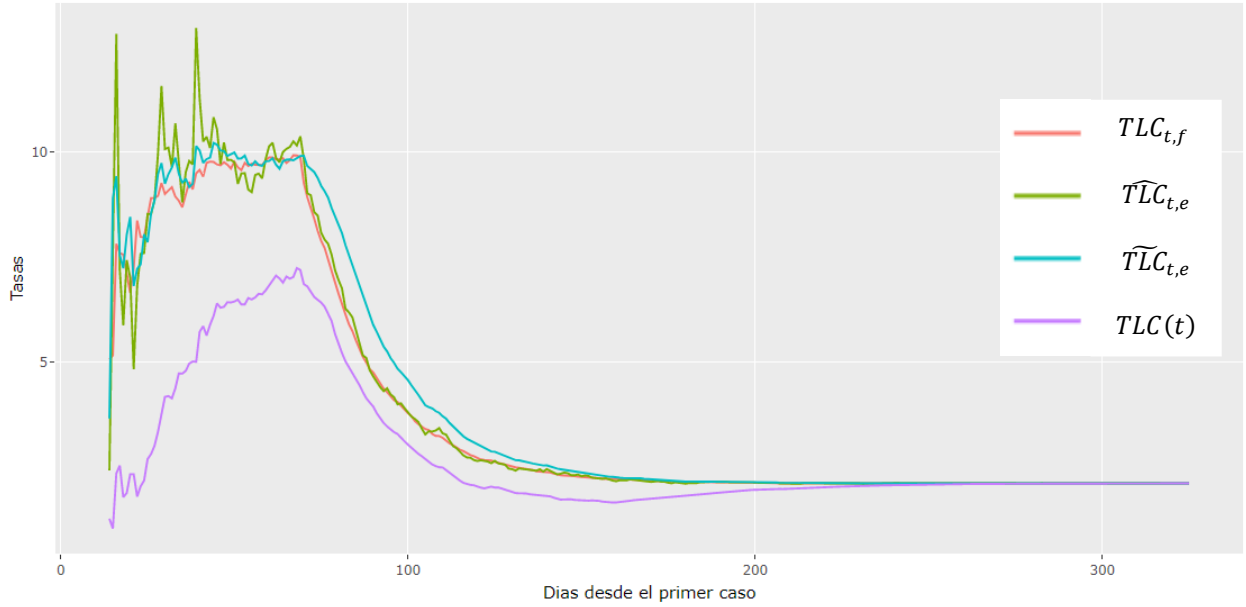


Figura 2: Escenario con letalidad 10% hasta el día 70 y con letalidad 2% luego, con F conocida.

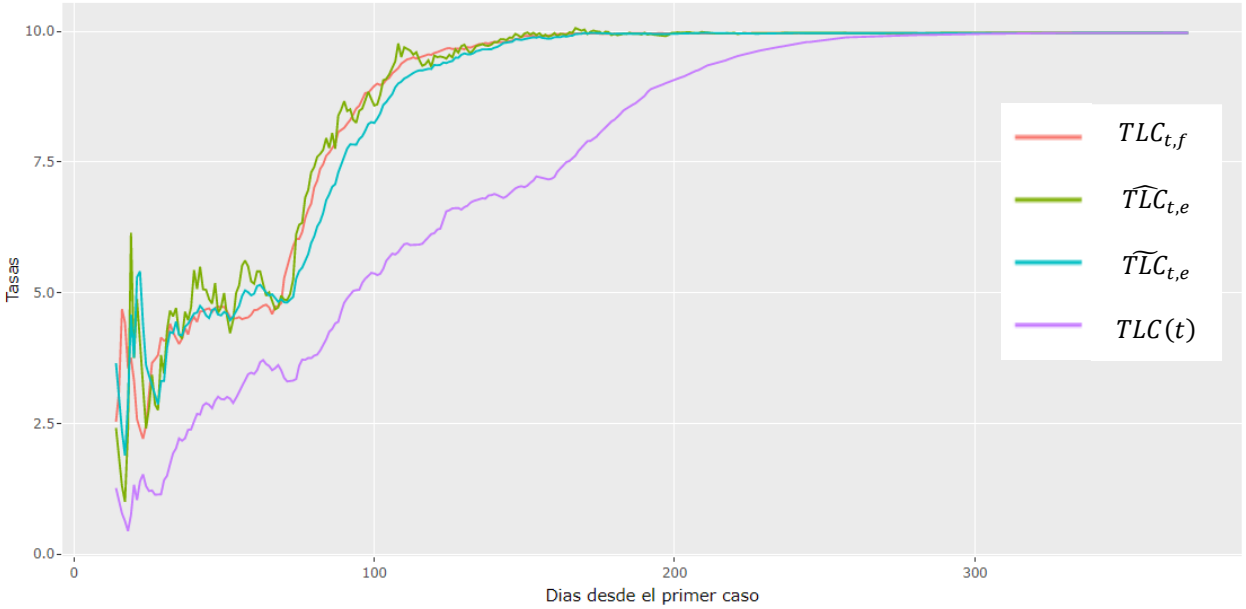


Figura 3: Escenario con letalidad 5% hasta el día 70 y con letalidad 10% luego, con F conocida.

5. Datos reales

En esta sección calculamos, para cada día t de los meses de junio, julio y agosto, diferentes estimaciones de la tasa de letalidad esperada $TLC_{t,e}$. Suponiendo que estamos en el día t, lo que proponemos es estimar todas las F_d, de los días anteriores a t de manera constante igual a F. Vamos a estimar $F(k)$, como la proporción de personas fallecidos que mueren en k días o menos, entre las personas que se les confirmó Covid hasta el día t - 45. Una idea más detallada de por qué para estimar la F nos corremos 45 días (para atrás) puede verse en la sección Recursos para la enseñanza.

Hay otras maneras más sofisticadas de estimar $F_d(k)$, por ejemplo, usando modelos paramétricos que, para simplificar el texto no presentamos en este capítulo.

La Figura 4 muestra los resultados obtenidos con diferentes métodos. Se observa que todas las propuestas se acercan a la curva roja, que es la letalidad final calculada el 31 de diciembre de 2020, mucho más que la rosa, que es la letalidad que se anuncia día a día.

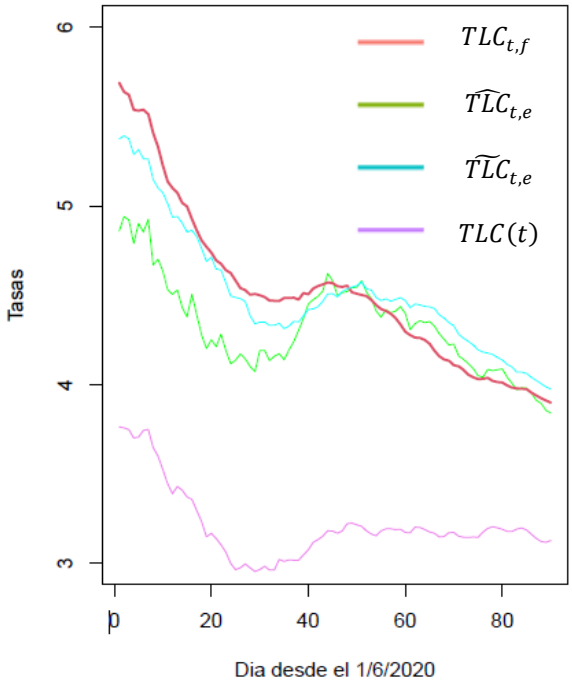
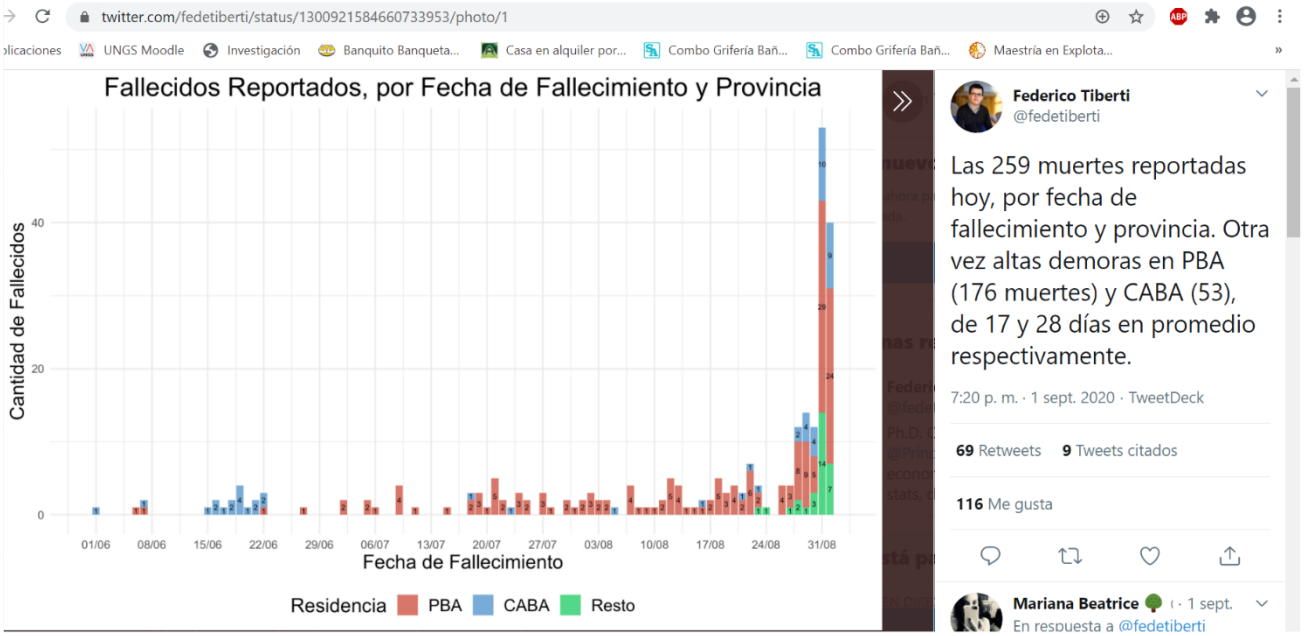


Figura 4: Diferentes estimaciones para la Tasa de letalidad esperada $TLC_{t,e}$ de Covid 19, para los meses de junio, julio y agosto de 2020 en Argentina.

5.1. Demoras en la carga de datos.

Otra de las dificultades que tiene poder estimar la tasa de letalidad sobre confirmados durante la epidemia es la demora en la carga de datos. En Argentina, en el período que analizamos, hubo sobre todo demoras importantes en la carga de fallecidos a la base de datos, como puede verse en el siguiente Tuit de Federico Tiberti.



Esto hacía que todas las propuestas subestimen la $TLC_{t,f}$. Debemos comentar que la base de datos que usamos para el análisis de los datos reales de Argentina entre junio y agosto de 2020 fue una base de datos del 31 de diciembre de 2020. Utilizando dicha base, pudimos reconstruir cual hubiera sido la base de datos completa que se debería haber tenido si no hubiera habido demora en la carga de datos para cada día entre el 1 de junio y el 31 de agosto, y con esas bases completas calculamos los estimadores.

El problema de cómo corregir en el momento la demora en la carga de datos es abordado por ejemplo por Bastos *et al.* (2019).

6. Recursos para la enseñanza

6.1. Ideas principales, cómo transmitirlas.

Hay ideas principales que creemos que se pueden transmitir en un curso de secundaria, incluso sin pasar por las definiciones formales de las variables aleatorias y por la definición de la Tasa de Letalidad acumulada esperada $TLC_{t,e}$. Por ejemplo, podríamos comenzar por preguntar en el aula quiénes escucharon algo sobre la letalidad de la Covid, qué piensan que significa, y cómo creen que se calcula.

Consensuar que la letalidad de un virus es la proporción de personas que finalmente fallecen por el virus entre las que se contagian. Hacer notar que en Covid no se sabe cuántas personas se contagian porque por ejemplo, muchas son asintomáticas. Entonces en principio sólo se pretende saber la letalidad sobre casos confirmados, o sea la proporción de personas que fallecen entre las que se les confirma el virus. Distinguir en este punto que hay una TLC (Tasa de Letalidad sobre confirmados) y una TLI (Tasa de Letalidad sobre infectados) y que nosotros vamos a tener en cuenta la TLC. Para quien quiera investigar, estas tasas se conocen, en inglés, como *case fatality rate* (CFR), la TLC e *infection fatality rate* (IFR), la TLI. Luego contar cómo se calcula con un ejemplo concreto. Por ejemplo, ¿qué se informa como la letalidad al finalizar el día 30 de junio de 2020 en Argentina? Por un lado, se cuenta x: "la cantidad de personas fallecidas hasta el 30 de junio" e y "la cantidad de personas confirmadas hasta el 30 de junio", y se informa como letalidad acumulada hasta el 30 de junio a la división x/y. Llegado a este punto consultar en el aula si les parece que es correcto hacer este cálculo para obtener la proporción de personas que finalmente mueren por Covid entre los casos confirmados hasta el 30 de junio. Tendría que poder surgir que seguramente hay personas que aún no murieron por Covid (sobre todo las que se contagiaron hace poco) y no las estamos considerando en el numerador x. Entonces sería lógico que la TLC que se informa habitualmente esté dando valores más bajos que los verdaderos valores de la TLC. Preguntar en el aula qué se podría hacer para intentar subsanar dicho problema. Llegar a la conclusión que lo que habría que hacer es poder predecir cuántas personas faltan morir y entonces corregir el numerador x con un numerador x^* que no cuente cuántos se murieron, sino que prediga cuántos morirán en total entre los confirmados hasta el 30 de junio. Preguntar al curso cómo se podría hacer para llegar a dicho valor x^*. Comentar que lo que serviría es saber cuánto demora en morir la gente que se le confirma el virus. Si, por ejemplo, supiéramos que el 40% de las personas que mueren por el virus, mueren en una semana o menos, entonces, rastreando a las personas que se les confirmó el virus el 23 de junio (justo 7 días antes del 30 de junio) y viendo cuántas de ellas murieron, podríamos predecir cuántas van a morir en total. Si vemos que murieron por ejemplo 10 de estas personas al 30 de junio, entonces podemos predecir que en total van a morir z_1, donde $z_1 \cdot 0{,}4 = 10$. Podemos despejar $z_1 = 10/0{,}4 = 25$. Luego dar un ejemplo similar. Si por ejemplo sabemos que el 70% se mueren en 10 días o menos, entonces podemos contar las personas que se murieron entre las que se contagiaron el 20 (10 días antes del 30 de junio). Preguntarles cómo predecirían cuántos van a morir en total entre las personas que se les confirmó el virus el 20 de junio, si por ejemplo se murieron 15 de estas personas al 30 de junio. Luego resolver: si llamamos z_2 a la cantidad que predecimos que finalmente va a morir entre las que se les confirmó el 20 de junio, debería valer que $z_2 \cdot 0{,}7 = 15$. Por lo tanto, podemos despejar $z_2 = 15/0{,}7 = 21{,}43$. De esta manera si

supiéramos cuántas personas se murieron hasta el 30 de junio entre las que se los confirmó el virus justo k días antes del 30 de junio, y además supiéramos qué proporción de personas que mueren por el virus, mueren en k días o menos (número que llamamos $F(k)$), entonces podríamos predecir cuántos morirán en total entre los confirmados hasta el 30 de junio del siguiente modo:

Si llamamos $M^*(k)$ a la cantidad de personas que murieron hasta el 30 de junio entre los que se les confirmó justo k días antes del 30 de junio, entonces podríamos calcular:

$$x^* = \frac{M^*(0)}{F(0)} + \frac{M^*(1)}{F(1)} + \frac{M^*(2)}{F(2)} + \cdots + \frac{M^*(119)}{F(119)}$$

donde 119 es la cantidad de días que pasaron desde el inicio de la epidemia en Argentina (3 de marzo de 2020). En esta instancia se puede preguntar cómo se podrían calcular o estimar los valores $F(k)$. Proponer para entenderlo un ejemplo ficticio. Seguimos pensando que estamos en el 30 de junio de 2020. Podemos suponer que las personas a las que se les confirmó el virus durante marzo y abril, o ya fallecieron o ya se recuperaron para el 30 de junio. Supongamos que de los casos confirmados en marzo y abril hay 100 casos fallecidos y para cada caso fallecido tenemos cuántos días pasaron entre la confirmación del virus y el fallecimiento. Luego, podríamos tener una tabla como la siguiente:

Días entre confirmación y muerte	0	1	2	3	4	5	6	7	8	9	10
Cantidad de casos	20	10	10	10	10	10	10	5	5	5	5

O sea, todos los casos en este ejemplo murieron en 10 días o menos. En este caso, calcularíamos:

k	0	1	2	3	4	5	6	7	8	9	10
$F(k)$	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,85	0,90	0,95	1

¿Por qué si estamos en el 30 de junio solo contemplamos los casos confirmados de marzo y abril? ¿Por qué no contemplamos todos los casos confirmados hasta el 30 de junio? Bueno, porque si consideráramos todos los casos confirmados hasta el 30 de junio, aquellos casos que se confirmaron pocos días antes del 30 de junio, si figuran entre los fallecidos es porque murieron en pocos días. Sin embargo, los que mueren luego del 30 de junio no los estaríamos teniendo en cuenta para la estimación de $F(k)$, quedando esta distorsionada. Así estimaríamos valores $F(k)$ más grandes que los reales para los k pequeños y valores $F(k)$ más chicos que los reales para valores de k grandes. Para finalizar se podría mostrar el gráfico de la Sección 5 del análisis de los meses entre junio y agosto hechos con la base de diciembre, comparar las tasas y ver cuánto mejor aproximan a la verdadera letalidad si se calcula de esta manera.

6.2. Función de distribución acumulada

Las variables aleatorias de conteo tienen como imagen o rango (conjunto de valores posibles), subconjuntos de los números enteros no negativos 0, 1, 2, …, etc. Por ejemplo, si definimos T como "cantidad de días de una semana (lunes a viernes), que un determinado individuo podrá dormir más de 8 horas"; los valores posibles de esta variable aleatoria son 0, 1, 2, 3, 4 y 5. En este caso se dice que T tiene rango finito. Ahora supongamos que observamos a una persona elegida al azar que se enferma y muere a causa de una enfermedad infecciosa y definimos T como "días que transcurren desde que se le confirma la enfermedad hasta que muere". Estamos asumiendo que la confirmación de la infección fue previa a su muerte, eventualmente el mismo día. Esta variable puede tomar valores enteros no negativos 0, 1, 2, …, etc., donde el valor 0 simplemente se interpreta como que la confirmación y muerte ocurrieron el mismo día. En este caso, decimos que el rango de la variable aleatoria T es infinito numerable.

Identificar la distribución de estas variables no es una tarea sencilla en general, especialmente con datos reales de conteo como los que trabajamos en este capítulo. La necesidad de conocer la distribución de una variable aleatoria se debe a que a partir de ella podemos responder preguntas como:

- ¿Qué tan probable es que una de estas personas muera a los 5 días de la confirmación de la enfermedad? o ¿cuál es la probabilidad de que tarde más de 10 días en morir?
- ¿Qué proporción de enfermos mueren antes del primer mes?
- ¿Qué es más probable: morir antes del día 5 o después?
- En promedio, aquellos que mueren por una enfermedad, ¿cuánto tardan en morir?

Toda la información que necesitamos para responder preguntas como estas, está contenida en una función que se llama la Función de distribución acumulada. Esta función determina completamente "el comportamiento" de la variable aleatoria. Si se la conoce de manera exacta, suele ser una tarea sencilla responder esas preguntas, en caso contrario se la aproxima de la mejor forma posible, siempre que se pueda.

A partir de la función de distribución acumulada podemos calcular las probabilidades puntuales para la variable aleatoria T, por ejemplo: la probabilidad de que un enfermo muera a los 9 días de habérsele

confirmado la enfermedad, es decir P(T = 9) o P(T=23), etc., y viceversa: si tenemos las probabilidades puntuales podemos calcular la función de distribución acumulada.

La función de distribución acumulada F de una variable aleatoria T se define, para cada número real t, como $F(t) = P(T \leq t)$, donde el evento $T \leq t$ significa que T tome valores menores o iguales a t. Veamos en un caso sencillo cómo obtener la función de distribución acumulada a partir de la función de distribución puntual y al revés. Sea T una variable que toma los valores {0, 2, 6} con probabilidades (conocidas):

$$P(T = 0) = 0.1,\ P(T = 2) = 0.5,\ P(T = 6) = 0.4$$

Luego

$F(t) = 0$ si $t < 0$

$F(t) = P(T = 0) = 0.1$ si $0 \leq t < 2$

$F(t) = P(T = 0) + P(T = 2) = 0.6$ si $2 \leq t < 6$

$F(t) = P(T = 0) + P(T = 2) + P(T = 6) = 1$ si $6 \leq t$.

Recíprocamente si sabemos que F está definida como arriba entonces podemos recuperar las probabilidades puntuales haciendo:

$P(T = 0) = F(0) = 0.1$

$P(T = 2) = F(2) - F(0) = 0.5$

$P(T = 6) = F(6) - F(2) = 0.4$.

7. Conclusiones y tareas pendientes

7.1. Conclusiones

En este trabajo hemos analizado el problema que tiene la tasa de letalidad que se informa habitualmente. Hemos notado que debido al tiempo que transcurre entre confirmación y fallecimiento de las personas que fallecen por el virus y la alta proporción de casos recientes que se tiene durante el brote de la epidemia, dicha tasa subestima la verdadera tasa de letalidad. Hemos propuesto un modelo matemático que nos permitió definir claramente la tasa de letalidad. Propusimos una nueva manera para estimar la tasa de letalidad y la comparamos con otra propuesta existente y con la tasa que se informa habitualmente. Hicimos la comparación en escenarios simulados observando ventajas de nuestra propuesta sobre las demás. Luego analizamos, en los meses de junio a agosto de 2020 en Argentina, cómo hubieran dado dichas tasas y las comparamos con la tasa final obtenida a fin de año, donde las personas confirmadas en dicho período asumimos que ya se recuperaron o ya fallecieron por Covid.

En dicho ejemplo también observamos una gran mejoría de las tasas propuestas sobre la tasa informada habitualmente. También hemos incorporado la sección "Recursos para la enseñanza" dónde intentamos resumir cuáles son las ideas que se les podría transmitir a un grupo de estudiantes sobre este capítulo guiando al docente con ejemplos qué podría utilizar.

7.2. Tareas pendientes

Cuando uno propone un estimador de cierto parámetro es deseable tener no solo el estimador sino también unas bandas de confianza que nos den una noción de la precisión del estimador. Al haber escrito nuestra propuesta como suma de variables independientes creemos que esto lo podremos lograr utilizando alguna generalización del Teorema Central del Límite.

Nuestra propuesta de estimación será buena siempre que logremos estimar bien las distribuciones de los tiempos entre confirmación y fallecimiento F_d. Hemos observado, aunque no lo presentamos en este capítulo, que dichas distribuciones pueden modelarse bien usando binomiales negativas con ceros inflados, ver Hilbe (2011) y Zeileis *et al.* (2008). De cualquier manera, estimar bien las F_d durante la epidemia es un problema pendiente que tendríamos que abordar. Es un problema que se conoce como estimación de distribución con datos censurados, ver por ejemplo Saffari y Adnan (2011). Otro tema que se podría abordar es el de las demoras en la carga de datos. Habría que incorporar alguna de las propuestas existentes para solucionar esas demoras y para que los métodos de estimación propuestos sean efectivos durante la epidemia.

8. Bibliografía

- Garske, T.; Legrand, J.; Donnelly, C. A.; Ward, H.; Cauchemez, S.; Fraser, C., ... y Ghani, A. C. (2009). "Assessing the severity of the novel influenza A/H1N1 pandemic". Bmj, Vol. nº. 339.
- Bastos, Leonardo S., et al. "A modelling approach for correcting reporting delays in disease surveillance data". Statistics in medicine, 2019, vol. 38, no 22, p. 4363-4377.
- Hilbe, Joseph M. "Negative binomial regression". Cambridge University Press, 2011.
- Zeileis, A., Kleiber, C., y Jackman, S. (2008). "Regression models for count data in R". Journal of statistical software, 27(8), 1-25.
- Saffari, S. E.; Adnan, R. (2011). "Zero-inflated negative binomial regression model with right censoring count data". Journal of Materials Science and Engineering. B, 1(4B), 551.