Estimación bayesiana de la tasa de mortalidad por COVID-19 dada la infección por SARS-Cov-2: el caso español

Anabel Blasco-Moreno[1,2], Sergio Baena-Mirabete[2], Daniel Baena-Mirabete[3] y Pere Puig[2]

[1] Servei d'Estadística Aplicada, Univ. Autònoma de Barcelona, Cerdanyola del Vallès, Spain. anabel.blasco@uab.cat

[2] Departament de Matemàtiques, Univ. Autònoma de Barcelona, Cerdanyola del Vallès, Spain

[3] Departament d'Estadística i Investigació Operativa, Univ. Politècnica de Catalunya, Barcelona, Spain

3 de julio de 2020

1. Introducción

La enfermedad por coronavirus 2019 (COVID-19) es una enfermedad infecciosa causada por el coronavirus 2 del síndrome respiratorio agudo severo (SARS-CoV-2). Se identificó por primera vez en diciembre de 2019 en Wuhan, China, y desde entonces se ha extendido a nivel mundial, lo que resulta en la actual pandemia. Hasta el 28 de junio de 2020, se han reportado más de 10 millones de casos en 213 países y territorios, lo que ha resultado en más de 500.000 muertes. Más de 5,5 millones de personas se han recuperado (Worldometer).

Uno de los aspectos de la nueva enfermedad es que muchas de las personas infectadas con SAR-Cov-2 son asintomáticas y solo pueden diagnosticarse si se someten a pruebas de PCR. Desafortunadamente, dichas pruebas no siempre están disponibles para la población general. Por lo tanto, es evidente que el recuento oficial de infectados por SARS-CoV-2 subestima severamente el número real de individuos infectados.

Los datos reportados hasta el momento revelan una gran dispersión en la tasa de letalidad (Case Fatality Rate (CFR), en inglés), del 0,5 % al 15 % (Worldometer, WHO). Esta tasa se calcula como el número de muertes por infección confirmada y las grandes diferencias observadas se deben principalmente a la disponibilidad de pruebas PCR de los diferentes países. Por otro lado, la tasa de mortalidad por infección (Infection Fatality Rate (IFR), en inglés) se calcula como el número de muertes por infección. Estos datos son más fiables al depender, el CFR, notoriamente de la disponibilidad de pruebas PCR. Por ello, la toma de decisiones políticas se sustenta a menudo en el IFR.

En las primeras etapas del brote y en países donde los kits de diagnóstico eran limitados, solo se evaluaron pacientes hospitalizados con síntomas avanzados de COVID-19. Bajo estas circunstancias, el CFR es una estimación inflada del IFR porque no se identifican muchas infecciones en la población. Esta inflación es intrínseca a los datos reportados en todos los países, aunque es mucho menor para los países que tienen una mayor capacidad de test diagnósticos, como es el caso de Corea del Sur. Por contra, en España, ni las personas con síntomas leves ni asintomáticos se someten a pruebas en base a los protocolos españoles establecidos debido a la escasez de pruebas de PCR. No obstante, en España se ha llevado a cabo la encuesta de seroprevalencia más grande implementada hasta el momento para confirmar la presencia de infecciones no reportadas. Sin embargo, la especificidad de las pruebas utilizadas se cuestiona constantemente por lo que los resultados obtenidos también son discutibles.

Por otro lado, el caso del crucero Diamond Princess reporta el IFR de una muestra de población completa, donde todos los clientes estuvieron expuestos al SARS-CoV-2. No obstante, no es necesariamente representativo de la demografía global, los ingresos o la disponibilidad de atención médica de los países en general. En este caso, el IFR fue del 1,3 % (Russell et al., 2020).

En este artículo proporcionamos un nuevo enfoque estadístico para estimar el IFR basado en un modelo jerárquico bayesiano considerando Sexo y Edad como factores explicativos. Se han utilizado los datos de

Corea del Sur para estimar los valores a priori asociados a la mortalidad por COVID-19 si se está infectado y, los datos del estudio de seroprevalencia español para estimar los valores a priori correspondientes a la probabilidad de infección por SARS-Cov-2.

2. Metodología

2.1. Definición del problema

Para el análisis de los datos observados de la enfermedad COVID-19 se parte de la base que existe un proceso no observado (latente) asociado a la probabilidad de infectarse por SARS-CoV-2. Sea X_i la variable aleatoria asociada a este proceso. Se tiene:

$$X_i = \begin{cases} 1 & \text{con probabilidad } p \\ 0 & \text{con probabilidad } 1-p \end{cases}$$

donde p corresponde a la probabilidad de estar infectado por SARS-CoV-2, es decir, la prevalencia de infección por SARS-Cov-2.

Sea Z_i la variable aleatoria que toma el valor 1 si el individuo $i-ésimo$ fallece de COVID-19 y 0 en caso contrario. Este es el proceso que realmente se observa.

Nótese que una persona fallece por COVID-19 si previamente se ha infectado por SARS-Cov-2. Luego, dicha información se interpreta en términos de las siguientes probabilidades condicionadas:

$$P(Z_i = 1 \mid X_i = 1) = w$$
$$P(Z_i = 1 \mid X_i = 0) = 0$$

donde w corresponde a la probabilidad de fallecer por COVID-19 si se está infectado por SARS-Cov-2, es decir, el IFR.

Por el Teorema de la Probabilidad Total, podemos obtener la probabilidad de fallecer por COVID-19 como sigue:

$$P(Z_i = 1) = P(Z_i = 1 \mid X_i = 1)P(X_i = 1) + P(Z_i = 1 \mid X_i = 0)P(X_i = 0) = pw$$

Es decir, la probabilidad de fallecer por COVID-19 es el producto de las dos probabilidades anteriores.

Para una población de N individuos, los datos disponibles corresponden a los n fallecimientos por COVID-19 observados, esto es:

$$\left\{ \sum_{i=1}^{N} Z_i = n; \quad N \right\}$$

Nótese que la probabilidad de fallecer por COVID-19 en esta población corresponde a la proporción muestral n/N. No obstante, el objetivo del estudio es estimar la probabilidad de fallecer por COVID-19 si se está infectado por SARS-Cov-2. Sin embargo, dado que el número total de infectados no es observable en la población, dicha probabilidad no puede ser estimada de forma directa.

El número de muertes observadas (n) en la población se distribuye según una distribución Binomial, $n \sim Bin(N, pw)$, donde pw corresponde a la probabilidad de morir por COVID-19, mencionada anteriormente. En este caso, el problema de la estimación de los parámetros no es identificable, sin embargo, admite una solución desde el enfoque Bayesiano.

La función de verosimilitud de los datos corresponde a la función de probabilidad de la $Bin(N, pw)$ donde, a cada uno de los parámetros p y w, se les asocia una distribución a priori. A partir del teorema de Bayes, la distribución a posteriori de los parámetros se obtiene de la siguiente manera:

$$P(p, w \mid n, N) = \frac{P(n, N \mid p, w)\pi(p, w)}{\int\int P(n, N \mid p, w)\pi(p, w)dpdw}$$

donde $P(n, N \mid p, w)$ es la función de verosimilitud y $\pi(p, w)$ es la distribución a priori conjunta de los parámetros. Asumiendo independencia entre p y w, se obtiene $\pi(p, w) = \pi_1(p) \cdot \pi_2(w)$.

2.2. Modelo estadístico

Hasta la fecha, la literatura recoge que existen diferencias en el IFR entre sexos y grupos de edad. Para la prevalencia de la infección por SARS-Cov-2, en algunos casos como el estudio español, se han detectado diferencias a nivel de edad. Dadas estas evidencias, se ha considerado oportuno incorporar las covariables Edad y Sexo en la estimación de los parámetros de interés. Para ello, se propone utilizar el siguiente modelo jerárquico bayesiano:

$$\begin{aligned}
n_{ij} &\sim Bin(N_{ij}, p_j w_{ij}), \quad i = 0, 1 \quad j = 0, 1, .., 9 \\
log\left(\tfrac{p_j}{1-p_j}\right) &= \alpha_0 + \alpha_1 (Edad_j - \xi)_-, \quad j = 0, 1, .., 9 \\
log\left(\tfrac{w_{ij}}{1-w_{ij}}\right) &= \beta_0 + \beta_1 Sexo_i + \beta_2 Edad_j + \beta_3 Sexo_i Edad_j, \quad i = 0, 1 \quad j = 0, 1, .., 9 \\
\alpha_0 &\sim N(\mu_{\alpha_0}, \sigma_{\alpha_0}), \\
\alpha_1 &\sim N(\mu_{\alpha_1}, \sigma_{\alpha_1}), \\
\xi &\sim Unif[\xi_1, \xi_2], \quad 0 \leq \xi_1, \xi_2 \leq 9 \\
\beta_0 &\sim N(\mu_{\beta_0}, \sigma_{\beta_0}), \\
\beta_1 &\sim N(\mu_{\beta_1}, \sigma_{\beta_1}), \\
\beta_2 &\sim N(\mu_{\beta_2}, \sigma_{\beta_2}), \\
\beta_3 &\sim N(\mu_{\beta_3}, \sigma_{\beta_3}).
\end{aligned}$$

donde, $\alpha_0, \alpha_1, \beta_0, ..., \beta_3$ y ξ son los parámetros del modelo. La variable $Sexo$ presenta las categorías $0 = Mujer$ y $1 = Hombre$. La variable $Edad$ se toma agrupada en 10 categorías: $0 =$ "$0-9$", $1 =$ "$10-19$", $2 =$ "$20-29$", $3 =$ "$30-39$", $4 =$ "$40-49$", $5 =$ "$50-59$", $6 =$ "$60-69$", $7 =$ "$70-79$", $8 =$ "$80-89$" y $9 =$ "$90\,o\,+$" años.

Por otro lado, $\mu_{\alpha_0}, \mu_{\alpha_1}, \sigma_{\alpha_0}$ y σ_{α_1} son los hiperparámetros (valores a priori) para las distribuciones de los parámetros α_0 y α_1. De la misma forma, $\mu_{\beta_0}, .., \mu_{\beta_3}$ y $\sigma_{\beta_0}, .., \sigma_{\beta_3}$ son los hiperparámetros para las distribuciones de los parámetros β. Finalmente, ξ_1 y ξ_2 son los hiperparámetros de la distribución uniforme de ξ. Este último parámetro corresponde a un punto de corte en la variable Edad a partir del cual se observa un cambio de tendencia en la probabilidad de contagio para los datos del estudio de seroprevalencia español, como se verá más adelante. En este caso, $(Edad_j - \xi)_-$ toma el valor 0 si $Edad_j > \xi$ y $Edad_j - \xi$ si $Edad_j \leq \xi$.

2.2.1. Método de estimación de la distribución a posteriori: Aproximación de Laplace

El método de Laplace consiste en maximizar la función log posterior conjunta para encontrar los máximos asociados a los parámetros de la misma. A continuación, se aproxima la densidad conjunta a posteriori a partir de una distribución Normal multivariante con vector de medias los máximos anteriores y matriz de variancias y covariancias la inversa de la matriz Hesiana evaluada en dichos máximos.

Concretamente, la función de densidad conjunta a posteriori tiene el siguiente aspecto:

$$f(\alpha_0,\alpha_1,\beta_0,\beta_1,\beta_3,\xi;N_{ij},n_{ij}) \propto \prod_{i=0}^{1}\prod_{j=0}^{9}(p_j w_{ij})^{n_{ij}}(1-p_j w_{ij})^{N_{ij}-n_{ij}}\cdot$$
$$\frac{1}{\sigma_{\alpha_0}\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{\alpha_0-\mu_{\alpha_0}}{\sigma_{\alpha_0}}\right)^2}\cdot ... \cdot \frac{1}{\sigma_{\beta_3}\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{\beta_3-\mu_{\beta_3}}{\sigma_{\beta_3}}\right)^2}$$

Tomando logaritmo, se obtiene la función log posterior:

$$logf(\alpha_0,\alpha_1,\beta_0,\beta_1,\beta_3,\xi;N_{ij},n_{ij}) \propto \sum_{i=0}^{1}\sum_{j=0}^{9}(n_{ij}(log(p_j)+log(w_{ij}))+(N_{ij}-n_{ij})log(1-p_j w_{ij}))+$$
$$log\left(\frac{1}{\sigma_{\alpha_0}\sqrt{2\pi}}\right)-\frac{1}{2}\left(\frac{\alpha_0-\mu_{\alpha_0}}{\sigma_{\alpha_0}}\right)^2+...+$$
$$log\left(\frac{1}{\sigma_{\beta_3}\sqrt{2\pi}}\right)-\frac{1}{2}\left(\frac{\beta_3-\mu_{\beta_3}}{\sigma_{\beta_3}}\right)^2$$

donde

$$p_j=\frac{1}{1+exp(-(\alpha_0+\alpha_1(Edad_j-\xi)\_))},\quad j=0,1,..,9$$

$$w_{ij}=\frac{1}{1+exp(-(\beta_0+\beta_1 Sexo_i+\beta_2 Edad_j+\beta_3 Sexo_i Edad_j))},\quad i=0,1\quad j=0,1,..,9$$

Finalmente, maximizando la función log posterior se obtienen los estimadores *maximum a posteriori (MAP)* de los parámetros.

3. Resultados

En primer lugar se presentan los resultados de los modelos iniciales realizados con la información del estudio de seroprevalencia español para obtener los hiperparámetros $\mu_{\alpha_0},\mu_{\alpha_1},\sigma_{\alpha_0}\sigma_{\alpha_1},\xi_1$ y ξ_2 y con los datos de Corea del Sur para obtener los hiperparámetros $\mu_{\beta_0},..,\mu_{\beta_3}$ y $\sigma_{\beta_0},..,\sigma_{\beta_3}$.

3.1. Estimación de los hiperparámetros para la probabilidad de infección por SARS-Cov-2

La información disponible corresponde al estudio ene-COVID-19 de sero-epidemiologia de la infección por SARS-Cov-2 en España. Los resultados de dicho estudio fueron reportados en dos rondas separadas. Los resultados de la primera ronda (Ministerio de Sanidad y ISCIII, 2020) se presentaron el 13 de mayo de 2020. Hasta la fecha se había recopilado información para un total de 60.897 participantes de los cuales 3.106 dieron positivo a SARS-Cov-2. El estudio se completó en la segunda ronda (Ministerio de Sanidad y ISCIII, 2020) donde el 95 % de los participantes de la primera ronda aceptaron participar de la segunda. Los resultados fueron presentados el 3 de junio de 2020 con un total de 63.564 participantes de los cuales 3.350 dieron positivo a SARS-Cov-2.

Los hiperparámetros $\mu_{\alpha_0},\mu_{\alpha_1},\sigma_{\alpha_0}\sigma_{\alpha_1},\xi_1$ y ξ_2 de las distribuciones a priori de los parámetros α_0, α_1 y ξ se han obtenido a partir de los datos recogidos en la segunda ronda, más actual y completa. El análisis de los datos de la primera ronda se presenta en el Apéndice A.

El término ξ corresponde a un salto observado entre los grupos de edad en la probabilidad de infectarse (Muggeo, 2003). Estudios anteriores indican que esta probabilidad no depende de la edad del sujeto y/o el sexo. No obstante, sí se observó en el caso de España, un salto cualitativo en dicha probabilidad a partir de una cierta edad en la primera ronda del estudio de seroprevalencia (ver Apéndice A). En la segunda ronda, pese a no ser tan evidente, los datos muestran cierta variabilidad en los grupos de edad más avanzados (Figura 1).

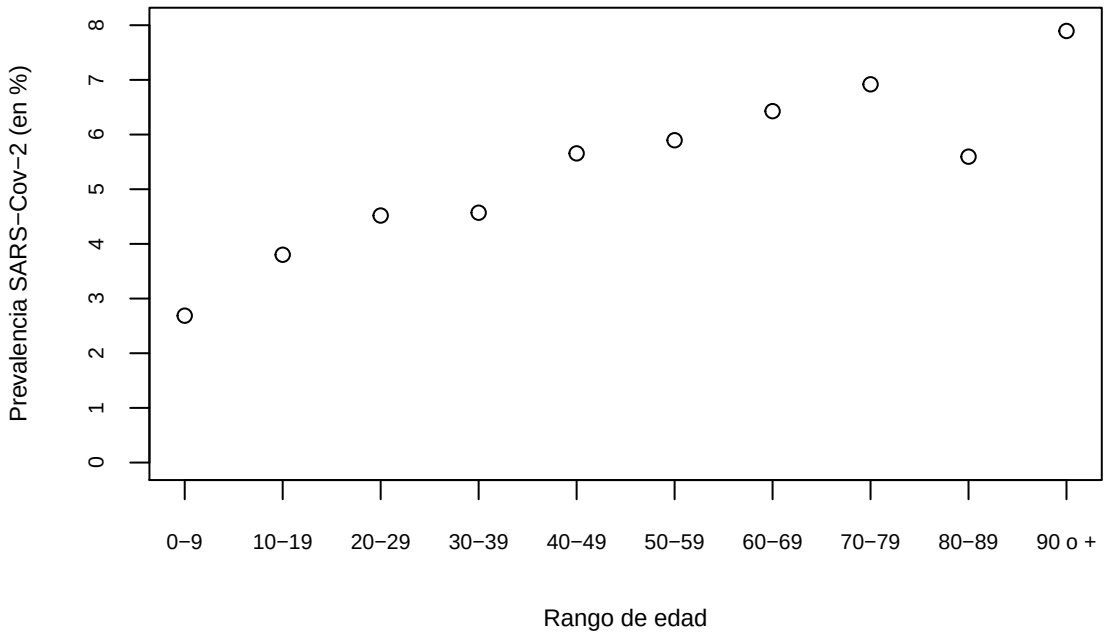

Figura 1: Prevalencia de infección por SARS-Cov-2 en España por categoría de edad

En particular llama la atención la prevalencia de infectados en el grupo de 80 a 89 años. En este grupo se evaluó a 2.484 personas, representan el 3,9 % del total, de las cuales 139 dieron positivo por SARS-Cov-2. Luego, se consideró más oportuno estimar una prevalencia común para los grupos de mayor edad como se había observado en la primera ronda del estudio (ver Apéndice A).

El número de personas infectadas por SARS-Cov-2 sigue una distribución Binomial. Luego, la probabilidad de estar infectado por SARS-Cov-2 en los datos del estudio de seroprevalencia español ($\tilde{p}$) se modela asumiendo una función logística como sigue:

$$log\left(\frac{\tilde{p}_j}{1-\tilde{p}_j}\right) = \tilde{\alpha_0} + \tilde{\alpha_1}(Edad_j - \tilde{\xi})_{\_}, \quad j = 0, .,9.$$

donde

$$(Edad_j - \tilde{\xi})_{\_} = \begin{cases} 0 & si \quad Edad_j > \tilde{\xi} \\ Edad_j - \tilde{\xi} & si \quad Edad_j \leq \tilde{\xi} \end{cases}$$

El modelo se estima por máxima verosimilitud. Los resultados se muestran en la Tabla 1 que recoge la estimación, la desviación estándar y el intervalo de confianza (IC) al 95 %.

Tabla 1: Estimadores máximo verosímiles para la probabilidad de infección por SARS-Cov-2

			IC (95 %)	
	Estimación	Desv. estándar	Inferior	Superior
$\tilde{\alpha_0}$	-2.665	0.031	-2.727	-2.602
$\tilde{\alpha_1}$	0.138	0.014	0.111	0.166
$\tilde{\xi}$	5.446	0.345	4.756	6.137

En la Figura 2 se muestra la estimación de la prevalencia de infección por SARS-Cov-2 para cada grupo de edad en comparación con el valor observado en los datos.

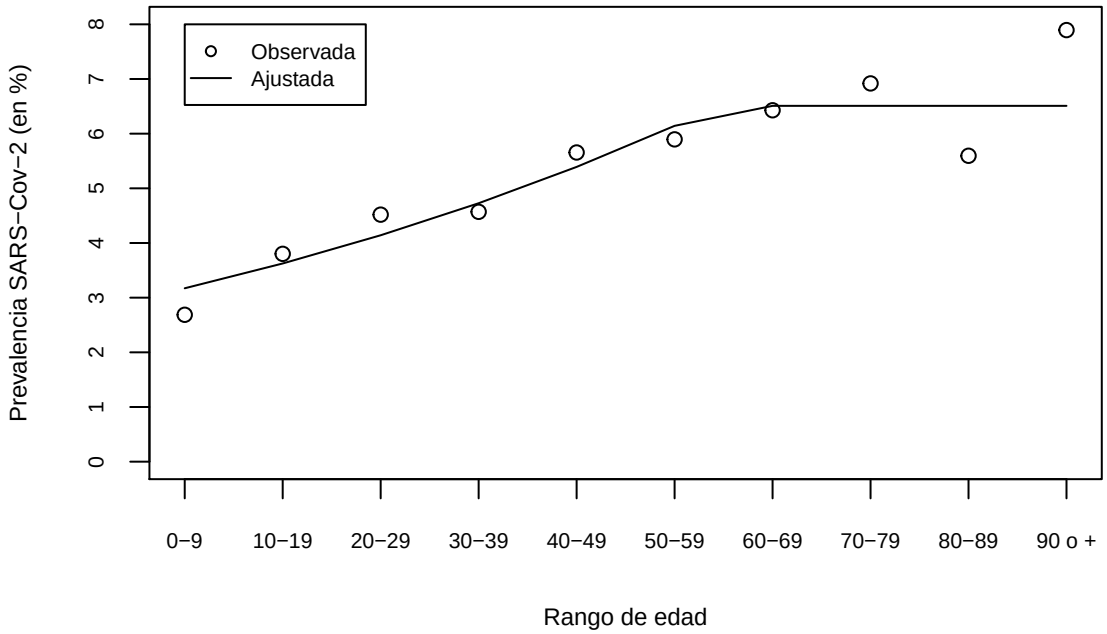

Figura 2: Prevalencia de infección por SARS-Cov-2 en España por grupo de edad: Observada vs Ajustada

A partir del rango de edad de 60 a 69 años, se estima una probabilidad de estar infectado por SARS-Cov-2 constante y cercana al 7 %.

3.2. Estimación de los hiperparámetros para la probabilidad de muerte por COVID-19 sobre casos confirmados

Los hiperparámetros $\mu_{\beta_0}, .., \mu_{\beta_3}$ y $\sigma_{\beta_0}, .., \sigma_{\beta_3}$ se obtienen a partir de los datos disponibles para Corea del Sur a fecha 13 de junio de 2020 extraídos de la web del KCDC (Korea Centers for Disease Control & Prevention, www.cdc.go.kr).

Corea del Sur es uno de los países del mundo que mayor número de pruebas PCR ha llevado a cabo, considerado como uno de los países con mayor fiabilidad en cuanto a los casos confirmados. En consecuencia, el cociente entre el número de fallecidos y el número de diagnosticados arrojaría un valor más próximo al IFR real comparado con otros países en los cuales se han realizado menos pruebas diagnósticas. La Figura 3 muestra la ratio de fallecidos por casos confirmados por grupo de edad y sexo.

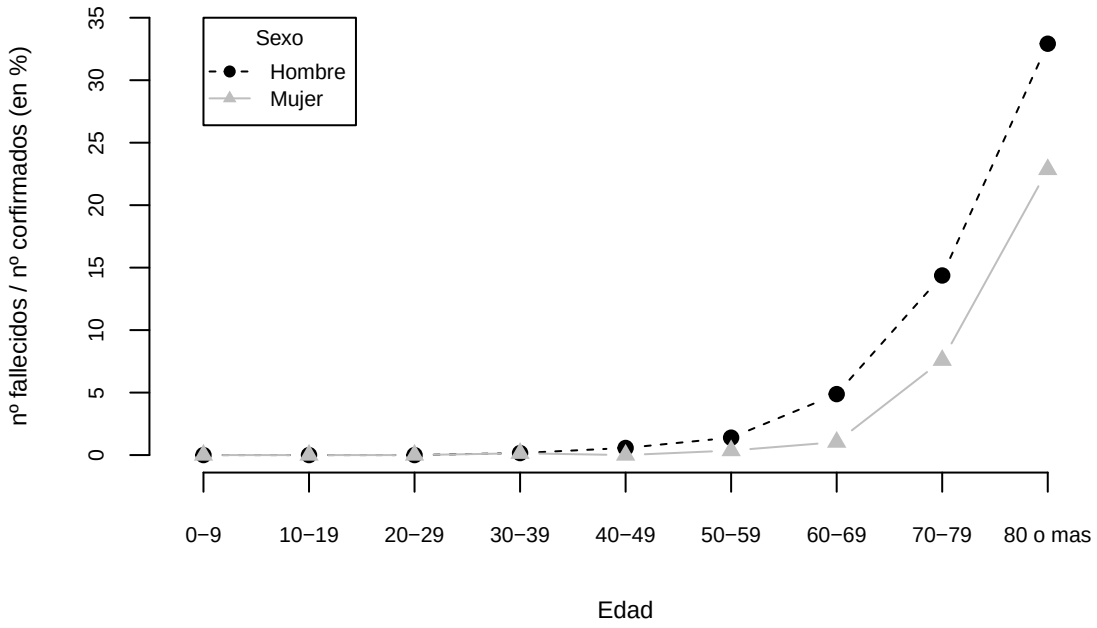

Figura 3: Número de fallecidos sobre casos confirmados en Corea del Sur

Se aprecia como el porcentaje de fallecidos (sobre el número de confirmados) del país es cercano al 35 % en hombres de más de 80 años y al 25 % en el caso de las mujeres (Figura 3). Estas diferencias entre sexo y edades son consistentes con las encontradas por Shim et al (2020).

El número de personas fallecidas por COVID-19 en Corea del Sur sigue una distribución binomial. La probabilidad de morir por COVID-19 sobre casos confirmados ($\tilde{w}$) se estima asumiendo una función logística como sigue:

$$log\left(\frac{\tilde{w}_{ij}}{1-\tilde{w}_{ij}}\right) = \tilde{\beta}_0 + \tilde{\beta}_1 Sexo_i + \tilde{\beta}_2 Edad_j + \tilde{\beta}_3 Sexo_i Edad_j. \quad i = 0,1 \quad j = 0,1,..,9$$

Como antes, el modelo se estima por máxima verosimilitud. Los resultados del modelo se encuentran en la Tabla 2.

Tabla 2: Estimadores máximo verosímiles para la probabilidad de muerte por COVID-19 sobre casos confirmados

			IC (95 %)	
	Estimación	Desv. estándar	Inferior	Superior
$\tilde{\beta}_0$	-12.655	0.755	-14.228	-11.262
$\tilde{\beta}_1$	1.404	0.102	1.214	1.613
$\tilde{\beta}_2$	3.176	0.915	1.420	5.022
$\tilde{\beta}_3$	-0.342	0.126	-0.595	-0.099

3.3. Ajuste del modelo jerárquico

Los datos de fallecidos en España por COVID-19 (n) se han extraído de la web del Ministerio de Sanidad a fecha 22 de mayo de 2020. Los datos referidos a la población (N) por grupo de edad y sexo, se han obtenido del Instituto Nacional de Estadítica (INE, www.ine.es)

Una vez obtenidos los hiperparámetros para las distribuciones a priori, el modelo jerárquico bayesiano que se ajusta es el siguiente:

$$
\begin{aligned}
n_{ij} &\sim Bin(N_{ij}, p_j w_{ij}) \qquad i = 0,1 \quad j = 0,1,..,9 \\
log\left(\tfrac{p_j}{1-p_j}\right) &= \alpha_0 + \alpha_1 (Edad_j - \xi)\_ \qquad j = 0,1,..,9 \\
log\left(\tfrac{w_{ij}}{1-w_{ij}}\right) &= \beta_0 + \beta_1 Sexo_i + \beta_2 Edad_j + \beta_3 Sexo_i Edad_j \qquad i = 0,1 \quad j = 0,1,..,9 \\
\alpha_0 &\sim N(-2,665; 0,031) \\
\alpha_1 &\sim N(0,138; 0,014) \\
\xi &\sim Unif[4,756; 6,137] \\
\beta_0 &\sim N(-12,655; 0,755) \\
\beta_1 &\sim N(1,404; 0,102) \\
\beta_2 &\sim N(3,176; 0,915) \\
\beta_3 &\sim N(-0,342; 0,126).
\end{aligned}
$$

Como se ha comentado en la Sección 2.2.1, se utiliza la aproximación de Laplace para encontrar los estimadores MAP de los parámetros (Tabla 3).

Tabla 3: Estimaciones maximum a posteriori del modelo

			IC (95 %)	
	Estimación MAP	Desv. estándar	Inferior	Superior
α_0	-2.767	0.032	-2.832	-2.703
α_1	0.145	0.014	0.118	0.173
ξ	5.191	0.289	4.628	5.766
β_0	-12.725	0.079	-12.883	-12.569
β_1	1.243	0.010	1.223	1.263
β_2	1.380	0.091	1.202	1.558
β_3	-0.074	0.012	-0.098	-0.051

La Tabla 4 recoge los estadísticos de resumen media y desviación estándar, así como el intervalo de credibilidad (IC) al 95 % obtenidos a partir de la distribución predictiva a posteriori de los parámetros w y p, respectivamente.

Tabla 4: Estimaciones para el IFR (w) y la prevalencia de infección por SARS-Cov-2 (p)

		Estimaciones para w (en %)				Estimaciones para p (en %)			
				IC (95 %)				IC(95 %)	
Edad	Sexo	Media	Desv. estándar	Inferior	Superior	Media	Desv. estándar	Inferior	Superior
0-9	Mujer	0.00030	0.00002	0.00025	0.00035	2.87957	0.27293	2.35720	3.42731
10-19	Mujer	0.00103	0.00007	0.00090	0.00118	3.31106	0.27327	2.77925	3.85112
20-29	Mujer	0.00358	0.00022	0.00316	0.00403	3.80540	0.27023	3.27569	4.33209
30-39	Mujer	0.01241	0.00065	0.01116	0.01373	4.37105	0.26483	3.84782	4.88289
40-49	Mujer	0.04298	0.00194	0.03923	0.04687	5.01730	0.26002	4.50136	5.52132
50-59	Mujer	0.14879	0.00577	0.13746	0.16036	5.72104	0.23208	5.24455	6.15654
60-69	Mujer	0.51379	0.01776	0.47912	0.54920	5.91342	0.18285	5.56316	6.27870
70-79	Mujer	1.75855	0.05784	1.64688	1.87362	5.91366	0.18294	5.56332	6.27963
80-89	Mujer	5.84235	0.19445	5.47024	6.23027	5.91366	0.18294	5.56332	6.27963
90 y +	Mujer	17.70026	0.58138	16.57607	18.86956	5.91366	0.18294	5.56332	6.27963
0-9	Hombre	0.00119	0.00009	0.00102	0.00138	2.87957	0.27293	2.35720	3.42731
10-19	Hombre	0.00382	0.00026	0.00334	0.00435	3.31106	0.27327	2.77925	3.85112
20-29	Hombre	0.01229	0.00071	0.01095	0.01375	3.80540	0.27023	3.27569	4.33209
30-39	Hombre	0.03951	0.00195	0.03584	0.04344	4.37105	0.26483	3.84782	4.88289
40-49	Hombre	0.12703	0.00529	0.11693	0.13748	5.01730	0.26002	4.50136	5.52132
50-59	Hombre	0.40760	0.01466	0.37939	0.43646	5.72104	0.23208	5.24455	6.15654
60-69	Hombre	1.30000	0.04288	1.21658	1.38424	5.91342	0.18285	5.56316	6.27870
70-79	Hombre	4.06681	0.13396	3.80831	4.33330	5.91366	0.18294	5.56332	6.27963
80-89	Hombre	12.00628	0.40780	11.21661	12.82089	5.91366	0.18294	5.56332	6.27963
90 y +	Hombre	30.51125	0.95969	28.61994	32.43187	5.91366	0.18294	5.56332	6.27963

La prevalencia de infección por SARS-Cov-2 es bastante baja en ambos sexos y en todos los grupos de edad, siendo de aproximadamente el 3 % en los grupos más jóvenes y cercano al 6 % en los grupos de mayor edad. Respecto a la mortalidad por COVID-19 si se está infectado, las diferencias son más notables, no solo entre hombres y mujeres sino también, entre grupos de edad (Tabla 4).

La Figura 4 muestra el IFR estimado por grupo de edad, para hombres y mujeres, y la ratio de fallecidos por casos confirmados.

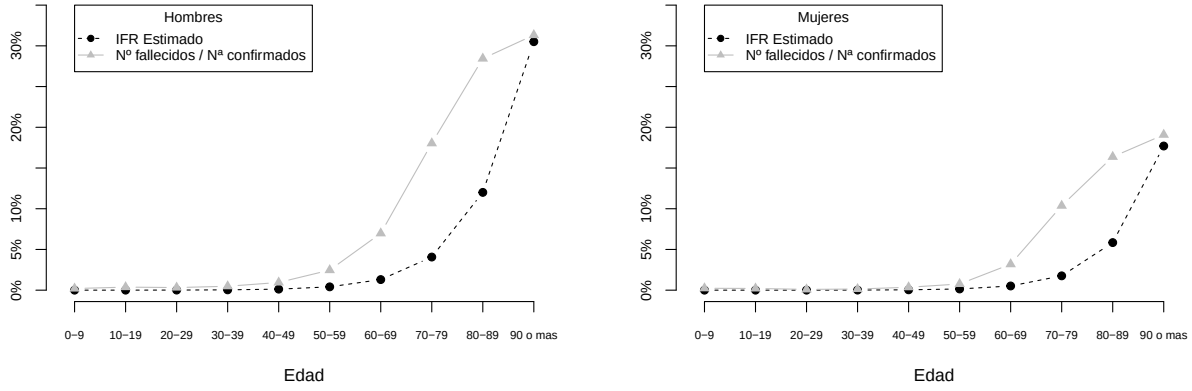

Figura 4: Nº fallecidos sobre casos confirmados e IFR estimado para España

El modelo propuesto estima un IFR inferior a la ratio de fallecidos por casos confirmados (Figura 4) corrigiendo de esta manera el efecto de no observar el total de casos confirmados en la población. Para el grupo de

edad de más de 90 años, el IFR estimado es muy similar a la ratio observada. Esto podría deberse a que el seguimiento de este grupo puede haber sido mayor dado que presentan peor sintomatología.

La Figura 5 muestra la distribución predictiva a posteriori del IFR por sexos en los grupos de edad de 60 a 69 años y de 70 a 79 años. La mortalidad por COVID-19 dada la infección por SARS-Cov-2 en hombres duplica la mortalidad estimada en mujeres.

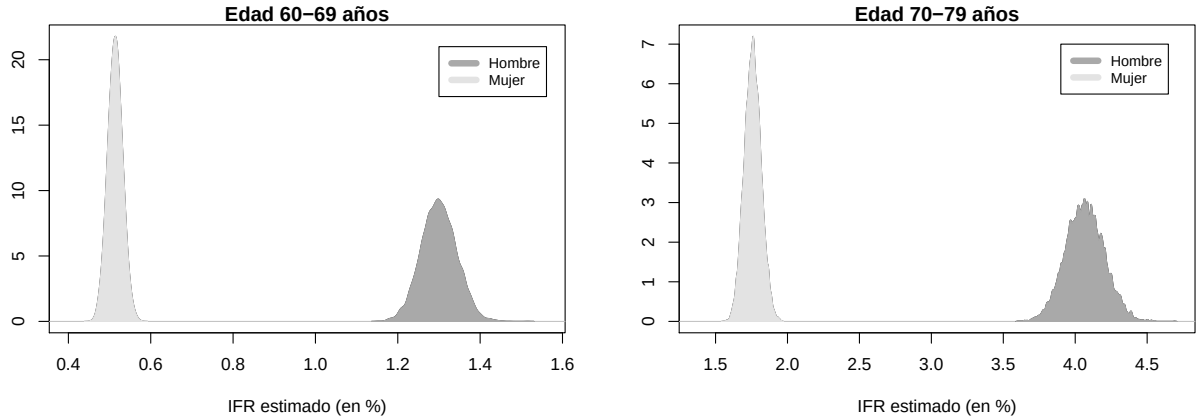

Figura 5: Distribución predictiva a posteriori del IFR estimado para España

4. Discusión

En el presente estudio, hemos presentado una metodología que tiene por objeto estimar el IFR por COVID-19 cuando se desconoce el total de población infectada. Como se ha comentado, el número total de casos infectados es desconocido, entre otras causas, porque existe un número de casos asintomáticos que no presentan ningún tipo de síntomas de la enfermedad y, por tanto, no son contabilizados como población infectada. Por ello, en nuestro estudio, se ha considerado oportuno no usar la información de casos confirmados. Como se observa en la Figura 4, el IFR estimado por rango de edad se sitúa por debajo de la tasa observada de fallecidos sobre casos confirmados.

Una cuestión que no ha sido abordada en nuestro estudio es la fiabilidad de la información del número de personas fallecidas por COVID-19. Por ejemplo, podríamos plantearnos si se han dejado de contabilizar fallecimientos debido a que muchos de los síntomas del COVID-19 son similares a otras enfermedades. Esto especialmente pudo haber sucedido en los primeros meses de la pandemia. Otro aspecto que podría afectar a la contabilidad de fallecidos por COVID-19 es la fecha considerada como inicio de la enfermedad. A día de hoy, parece existir un fuerte consenso en la comunidad científica que afirma que el COVID-19 ya estaba circulando en la sociedad antes de lo que inicialmente se pensó. Por tanto, es muy probable que realmente el número de fallecidos por COVID-19 sea realmente superior al reportado oficialmente. En este sentido, una interesante línea de trabajo futura sería usar la serie temporal de fallecidos en España anterior y posterior al inicio de la pandemia. Desde este enfoque, los fallecidos por COVID-19 podrían inferirse estimando el exceso de muertes respecto al valor histórico esperado. Un estudio interesante al respecto puede consultarse en Rinaldi & Paradisi (2020) en el cual los autores usan la serie histórica del total de fallecidos en Italia, por rango de edad, desde 2015 hasta 2020 con el objetivo de estimar el IFR por COVID-19 sin estar sujetos a estas irregularidades en la contabilización del número de fallecidos e infectados por la enfermedad.

5. Referencias

Flaxman, S., Mishra, S., Gandy, A., Unwin, H., Coupland, H., Mellan, T., ... & Schmit, N. (2020). Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries.

Grewellea, R. E. & De Leoa, G. A. (2020). Estimating the Global Infection Fatality Rate of COVID-19. *Preprint submitted to MedRxiv*. https://doi.org/10.1101/2020.05.11.20098780

Ministerio de Sanidad. Instituto de Salud Carlos III (ISCIII). (2020). Estudio ENE-COVID19: Primera Ronda. Estudio Nacional de Sero-Epidemiología de la Infección por SARS-COV-2 en España. Informe preliminar 13 de mayo de 2020.

Ministerio de Sanidad. Instituto de Salud Carlos III (ISCIII). (2020). Estudio ENE-COVID19: Segunda Ronda. Estudio Nacional de Sero-Epidemiología de la Infección por SARS-COV-2 en España. Informe preliminar 3 de junio de 2020.

Muggeo, V. M. (2003). Estimating regression models with unknown break-points. *Statistics in medicine*, 22(19), 3055-3071.

Rinaldi, G., & Paradisi, M. (2020). An empirical estimate of the infection fatality rate of COVID-19 from the first Italian outbreak. *medRxiv*.

Russell, T. W., Hellewell, J., Jarvis, C. I., Van Zandvoort, K., Abbott, S., Ratnayake, R., ... & CMMID COVID-19 working group. (2020). Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020. *Eurosurveillance*, 25(12), 2000256.

Shim, E., Tariq, A., Choi, W., Lee, Y., & Chowell, G. (2020). Transmission potential and severity of COVID-19 in South Korea. *International Journal of Infectious Diseases*.

Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., ... & Dighe, A. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases*.

World Health Organization, & World health organization. (2020). Coronavirus disease (COVID-2019) situation reports.

Worldometer (2020). www.worldometers.info/coronavirus/. Dover, Delaware, U.S.A.

A. Apéndice

En este Apéndice se presentan los resultados del análisis de los datos de la primera ronda del estudio de seroprevalencia español. Estos resultados se descartaron en favor de los datos de la segunda ronda más completos y actuales.

A.1. Estimación de los hiperparámetros para la probabilidad de infección por SARS-Cov-2: datos estudio de seroprevalencia español primera ronda

La información disponible corresponde a la primera ronda (Ministerio de Sanidad y ISCIII, 2020) del estudio ene-COVID-19 de sero-epidemiologia de la infección por SARS-Cov-2 en España.

En los datos de la primera ronda se observó un cambio de tendencia en la prevalencia de infección por SARS-Cov-2. A partir del grupo de 50 a 59 años, la prevalencia se estabiliza alrededor del 7 % (Figura 6).

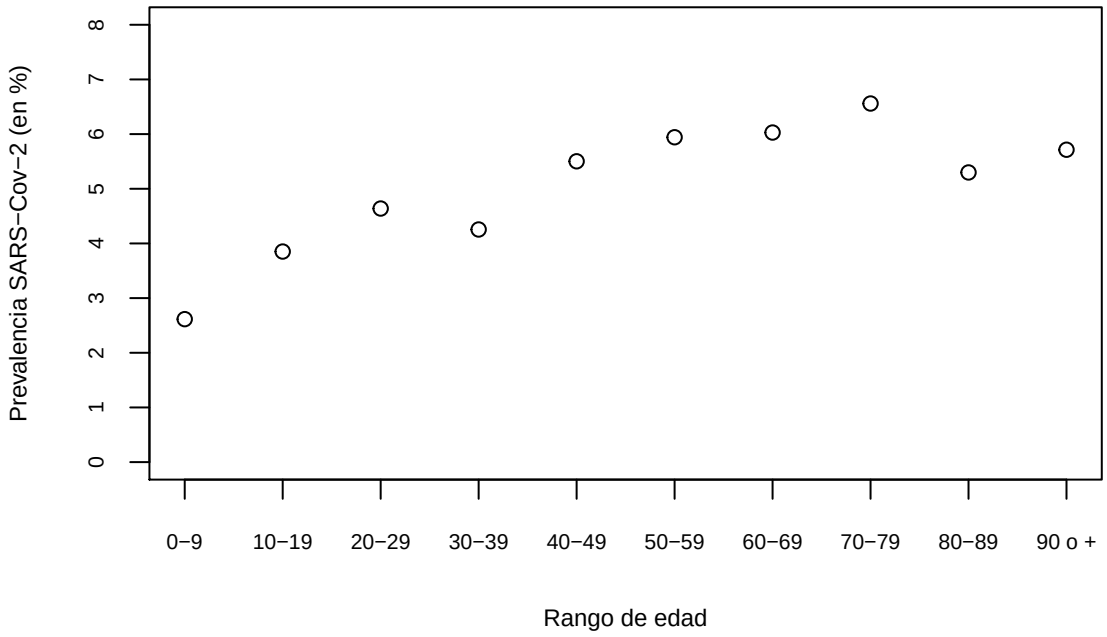

Figura 6: Prevalencia de infección por SARS-Cov-2 en España por categoría de edad

Las estimaciones proporcionadas por el modelo binomial comentado en la sección 3.1 se muestran en la Tabla 5.

Tabla 5: Estimadores máximo verosímiles para la probablidad de contagio

			IC (95 %)	
	Media	Desv. estándar	Inferior	Superior
$\tilde{\alpha_0}$	-2.747	0.026	-2.799	-2.696
$\tilde{\alpha_1}$	0.150	0.019	0.112	0.188
$\tilde{\xi}$	4.703	0.361	3.980	5.426

\

A partir del rango de edad de 50 a 59 años, se estima una probabilidad de estar infectado por SARS-Cov-2 constante y ligeramente superior al 6 % (Figura 7).

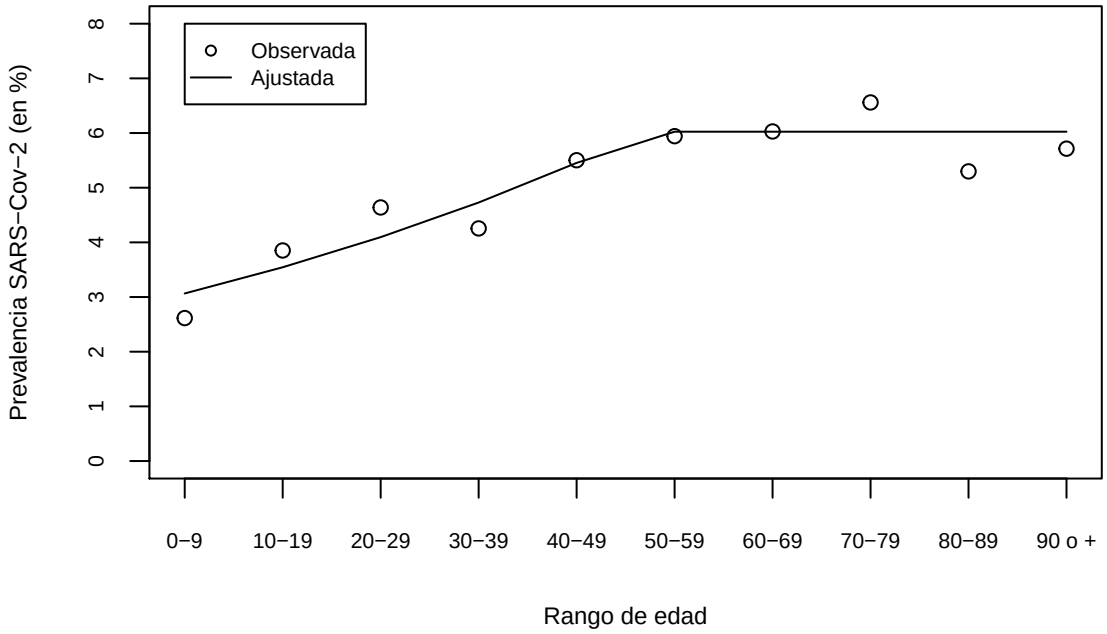

Figura 7: Prevalencia de infección por SARS-Cov-2 en España por grupo de edad: Observada vs Ajustada