Estadística básica para la pandemia del coronavirus COVID-19

Ignacio Cascos, ignacio.cascos@uc3m.es
Departmento de Estadística, Universidad Carlos III de Madrid

Mayo de 2020

Contents

1 Introducción

El 13 de Mayo de 2020, el *Instituto de Salud Carlos III* publicó un informe sobre la primera ronda del Estudio Nacional de Sero-Epidemiología de la Infeccion por SARS-COV-2 en España (ENE-Covid19). El informe contiene resultados interesantes y resulta de lectura sencilla. Este documento ha sido confeccionado para facilitar la lectura y comprensión de dicho informe, así como de otros estudios epidemiológicos y estadísticos sobre la pandemia del COVID-19. Su lectura no requiere ningún conociento estadístico previo, pero haber seguido un curso a nivel de grado sobre Introducción a la Estadística la facilitaría.

2 Test biomédicos y jerga asociada

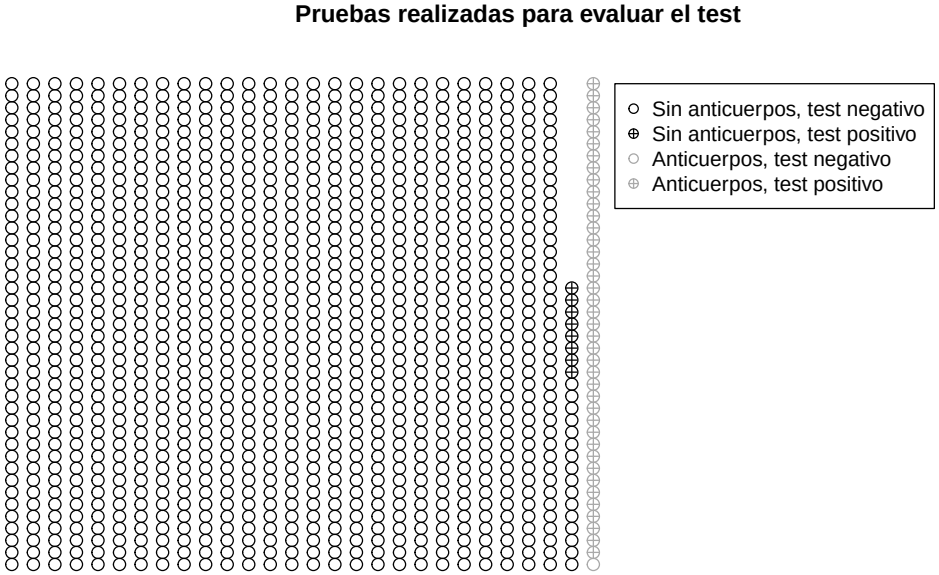
Los términos *prevalencia*, *sensibilidad* y *especificidad* son habituales en la jerga de las Ciencias Biomédicas.

- La *prevalencia* es el porcentaje (o porporción) de individuos de una población que tienen cierta condición médica.

  - Ejemplo 1: La prevalencia actual de anticuerpos IgG contra el SARS-Cov2 en España es aproximadamente del 5% (ENE-Covid19).

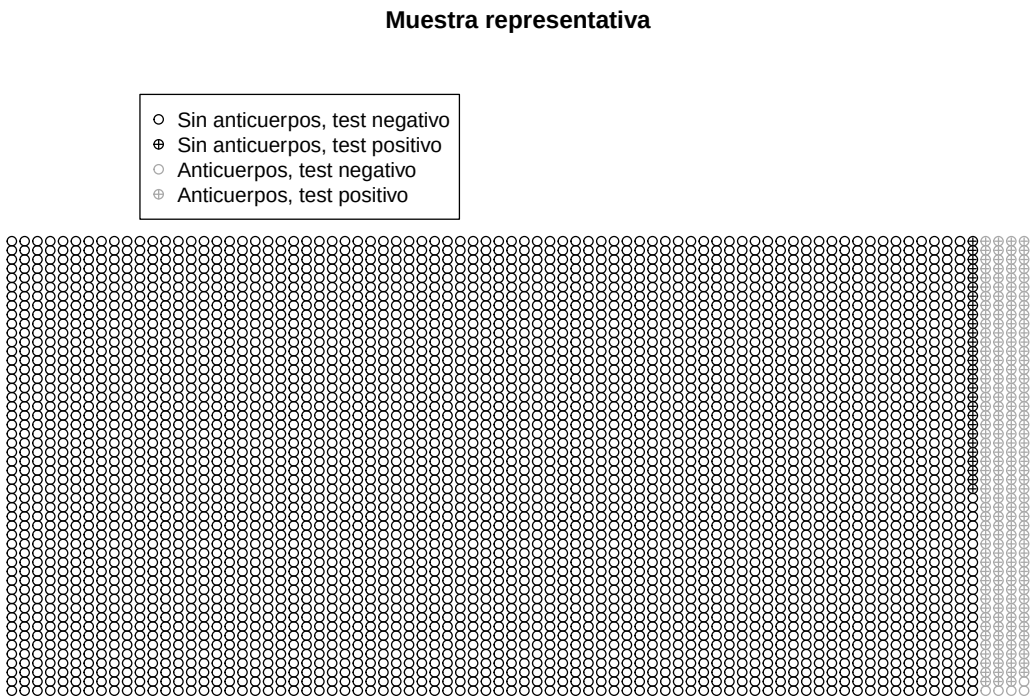
- La *sensibilidad* (Tasa de Verdaderos Positivos, TVP) es la proporción de indiviuos infectados que son identificados como tales.
  - Ejemplo 2: La Sensibilidad declarada del *DiaSorin LIAISON SARS-CoV-2 S1/S2 IgG* para la detección de anticuerpos IgG es del 96.67% dado que fue probado en 41 individuos portadores de anticuerpos dando resultado positivo en 40 de ellos. Ver el informe sobre los Test Serológicos Autorizados de la FDA.
- La *especificidad* (Tasa de Verdaderos Negativos, TVN) es la proporción de individuos sanos que son identificados como tales.
  - Ejemplo 3: La Especificidad declarada del *DiaSorin LIAISON SARS-CoV-2 S1/S2 IgG* para la detección de anticuerpos IgG es del 99.3% dado que fue probado en 1090 individuos sin anticuerpos dando resultado negativo en 1082 de ellos. Ver el informe sobre los Test Serológicos Autorizados de la FDA.

El gráfico presentado a continuación refleja una circunferencia negra por cada test realizado sobre un individuo sin anticuerpos en el Ejemplo 3, y una circunferencia gris por cada test realizado en cada individuo con anticuerpos en el Ejemplo 2. Si el resultado del test fue positivo, hay un signo más '+' dibujado dentro de la circunferencia.



El gráfico anterior no es representativo de la población española. En él hay un 3.63% de circunferencias grises (representando individuos con anticuerpos), mientras que el porcentaje de la población española con anticuerpos es aproximadamente el 5%, ver Ejemplo 1.

Justo a continuación, aparece un nuevo gráfico en el que el porcentaje de circunferencias grises es el 5%, mientras que la razón de signos más dentro de las circunferencias negras es aproximadamente el mismo que en el gráfico superior, que se corresponde con la especificidad del Ejemplo 3. Del mismo modo, la razón de signos más entre las circunferecias grises se corresponde con la sensibilidad del Ejemplo 2.



La *sensibilidad* y *especificidad* cuantifican la precisión de una prueba diagnóstica, pero para un individuo sobre el que ya ha sido realizada la prueba habiendo indicado un resultado positivo, lo que le preocupa es la probabilidad de que realmente tenga la condición médica (que recibe el nombre de Valor Predictivo Positivo, VPP). Dicho de otra manera, sabe que tiene asociada una circunfernecia con un '+' dentro de ella, pero ¿se trata de una circunferencia gris?

Por otro lado, si el resultado de su prueba diagnóstica fue negativo, lo que le preocupará es la probabilidad de no tener la condición médica (Valor Predictivo Negativo, VPN). Observa que la probabilidad de tener anticuerpos pese a que el resultado del test haya sido negativo es su complementario, $1 - \text{VPN}$.

3 Notación probabilística

A continuación introducimos notación probabilística básica:

- Denotamos como D el suceso que indica que un individuo tiene anticuerpos (en realidad D hace referencia a enfermedad, *disease*). Su probabilidad $P(D)$ representa la verosimilitud de que un invididuo seleccionado al azar sea portador anticuerpos, y es igual a la prevalencia.
  - Ejemplo 1: Si D representa tener anticuerpos IgG contra el SARS-Cov2, su probabilidad es $P(D) = 0.05$.
- Denotamos como '+' el suceso que indica que el resultado de la prueba diagnóstica es positivo. La sensibilidad del test es la probabilidad condicionada de un resultado positivo dado que el individuo es portador de anticuerpos, lo que denotamos como Sensibilidad $= P(+|D)$, donde la línea vertical

'|' representa una *probabilidad condicionada* (el suceso que aparece a su derecha refleja la información disponible, mientras que el que está a su izquierda es el suceso cuya probabilidad queremos calcular). Representa la verosimilitud de que el resultado de la prueba diagnósitica de un portador de anticuerpos sea positivo.

  - Ejemplo 2: $P(+|D) = 0.9667$.

- Denotamos como '$-$' el suceso que indica que el resultado de la prueba diagnóstica es negativo. La especificidad del test es la probabilidad condicionada de un resultado negativo dado que el individuo no tiene anticuerpos, lo que denotamos como Especificidad $= P(-|\overline{D})$. El suceso $\overline{D}$ es el *suceso complementario* a D, es decir, el individuo no tiene anticuerpos IgG.

  - Ejemplo 3: $P(-|\overline{D}) = 0.993$.

- La probabilidad de que un individuo cuyo resultado en el test es positivo sea portador de anticuerpos es VPP $= P(D|+)$.
- La probabilidad de que un individuo cuyo resultado en el test es negativo sea portador de anticuerpos es VPN $= P(\overline{D}|-)$.

4 Fórmula de Bayes (y de la probabilidad total)

Vamos ahora a calcular la probabilidad de que un individuo tenga anticuerpos dado que ha obtenido un resultado positivo en la prueba diagnóstica, $P(D|+)$. En el gráfico con la muestra representativa, esto es equivalente a restringirse a las circunferencias con un signo más dentro de ellas (test positivos) y calcular la razón de circunferencias grises (portadores de anticuerpos) de entre ellas, es decir,

$$P(D|+) = \frac{P(D \cap +)}{P(+)}\,,$$

donde la intersección $\cap$ impone que ambos sucesos deben ocurrir simultáneamente (test positivo y portador de anticuerpos).

Paso a paso, el numerador $P(D \cap +)$ es la proporción de circunferencias grises con un signo más dentro de ellas en la muestra representativa. Se calcula multiplicando la proporción de individuos con anticuerpos, $P(D)$ (proporción de circunferencias grises, prevalencia), por la razón de test positivos de entre los que tiene anticuerpos, $P(+|D)$ (razón de circunferencias con un signo más de entre las grises, sensibilidad).

Para calcular el denominador tenemos que recurrir a un procedimiento un poco más largo (la *regla de la probabilidad total*). Comenzamos partiendo los resultados positivos en dos grupos, por un lado aquellos que están asociados a portadores de anticuerpos y por otro, aquellos que se obtuvieron en individuos sin anticuerpos. La probabilidad de un resultado positivo es la suma de las probabilidad de los dos grupos anteriores, que se calcula multiplicando la proporción de individuos con (o sin) anticuerpos por la razón de test positvos en cada uno de los dos grupos,

$$\begin{aligned} P(D|+) = \frac{P(D \cap +)}{P(+)} &= \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|\overline{D})P(\overline{D})} \\ &= \frac{P(+|D)P(D)}{P(+|D)P(D) + (1 - P(-|\overline{D}))(1 - P(D))} \\ &= \frac{\text{Sensibilidad} \times \text{Prevalencia}}{\text{Sensibilidad} \times \text{Prevalencia} + (1 - \text{Especificidad}) \times (1 - \text{Prevalencia})}\,. \end{aligned}$$

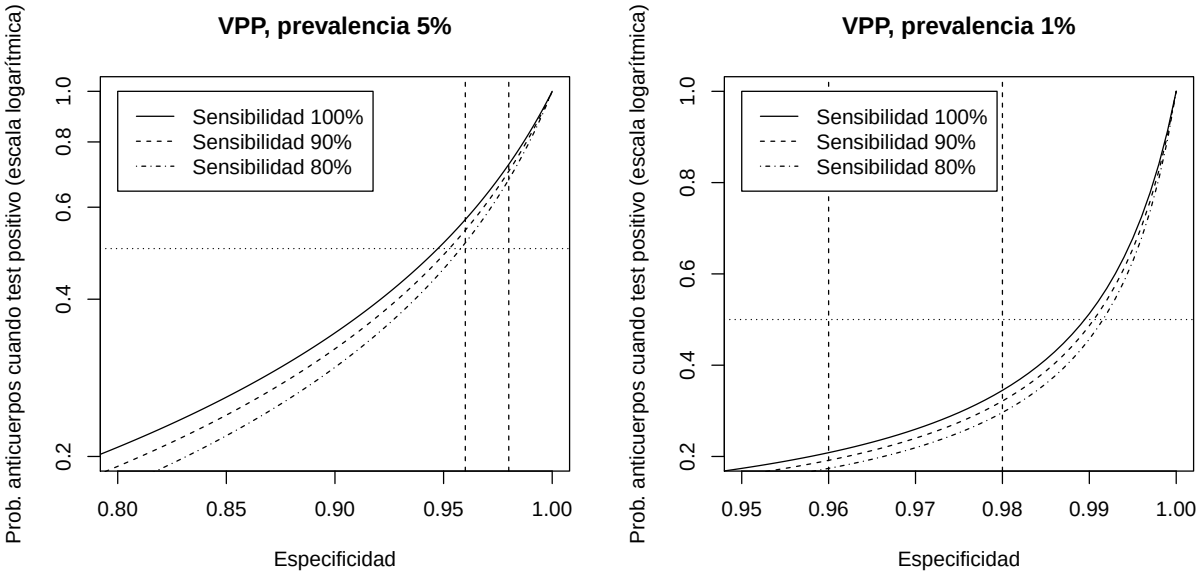
De modo análogo (y saltándonos los detalles) obtenemos que la TPN es

$$P(\overline{D}|-) = \frac{P(\overline{D} \cap -)}{P(-)} = \frac{P(-|\overline{D})P(\overline{D})}{P(-|D)P(D) + P(-|\overline{D})P(\overline{D})}$$
$$= \frac{\text{Especificidad} \times (1 - \text{Prevalencia})}{(1 - \text{Sensibilidad}) \times \text{Prevalencia} + \text{Especificidad} \times (1 - \text{Prevalencia})}\,.$$

- Ejemplo: Si la prevalencia de anticuerpos IgG contra el SARS-Cov2 es 5%, entonces
  - Para un individuo cuyo resultado en el test es positivo, la probabilidad de que realmente porte anticuerpos es
$$P(D|+) = \frac{0.9667 \times 0.05}{0.9667 \times 0.05 + 0.007 \times 0.95} = 0.879\,.$$
  - Para un individuo cuyo resultado en el test es negativo, la probabilidad de que no porte anticuerpos es
$$P(\overline{D}|-) = \frac{0.993 \times 0.95}{0.993 \times 0.95 + 0.0333 \times 0.05} = 0.998\,.$$

5 Valor Predictivo Positivo (VPP)

Suponiendo que la prevalencia es bien el 5% o el 1%, en los gráficos a continuación, se observa que la proporción de individuos con anticuerpos dentro de aquellos que obtuvieron resultado positivo en la prueba diagnóstica (VPP) depende de la sensibilidad y especificidad.



En un informe de la Sociedad Americana de Enfermedades Infecciosas puede leerse "Algunos test de anticuerpos para el COVID-19 autorizados por la FDA tienen una especificidad estimada del 96-98%, por lo que un resultado positivo puede ser con mayor probabilidad un falso positivo que un verdadero positivo si la prevalencia de los anticuerpos en la población es del 5% o menos". La línea punteada horizonal de los dos gráficos superiores está fijada en 0.5. Cualquier VPP por debajo proviene de un test para el que la probabilidad de tener anticuerpos dado un resultado positivo (verdadero positivo) es inferior a 0.5. Por

tanto, la probabilidad de no tener anticuerpos dado un resultado positivo (falso positivo) es superior a 0.5. En definitiva, por debajo de la línea punteada horizontal, los falsos positivos son más probables que los verdaderos positivos. Las líneas discontinuas verticales indican valores de especificidad 0.96 y 0.98, tal como aparecía reflejado en el informe. Observa que los falsos positivos son bastante frecuentes con una prevalencia del 5% si el test no es muy preciso, mientras que son frecuentes con una prevalencia del 1% incluso con contrastes muy precisos.

6 Valor Predictivo Negativo (VPN)

La prueba diagnóstica utilizada para obtener los datos del (ENE-Covid19) es el *Zhejiang Orient Gene Biotech IgG rapid test*. El fabricante declaró una sensibilidad del 97% y una especificidad del 100%, mientras que estudios de fiabilidad posteriores revelaron una sensibilidad del 79%, mientras que la especificidad es 100%.

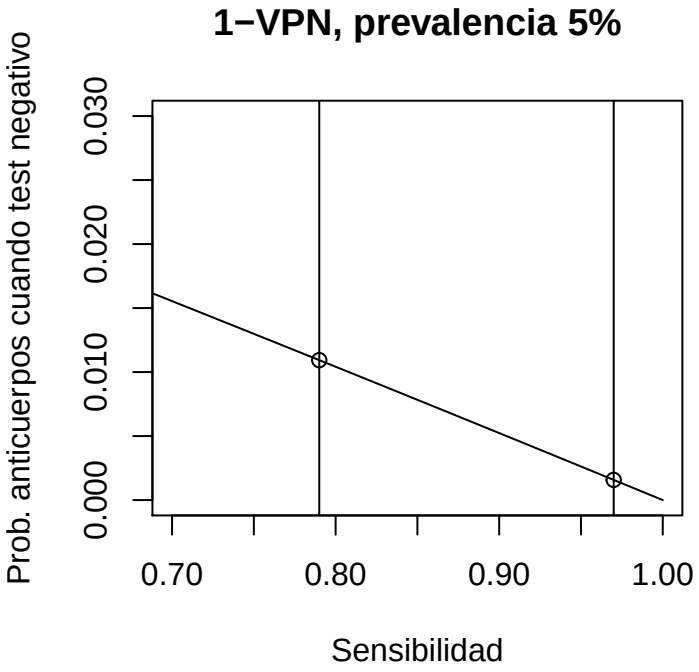
Si una test tiene una especificidad del 100%, $P(-|\overline{D}) = 1$, todos los individuos que no tienen anticuerpos IgG obtendrán un resultado negativo en el test, así que la única posibilidad para que un individuo obtenga un resultado positivo es que sea portador de anticuerpos IgG. En consecuencia, VPP es 1, pero existe la posibilidad de que un individuo obtenga un resultado negativo pese a portar anticuerpos. En concreto, si la especificidad es 79%, tal como indicaban los estudios de fiabilidad posteriores, aproximadamente 1.1% de los individuos con un resultado negativo en la prueba, portarán anticuerpos.

- Si Sensibilidad $= 0.79$, entonces

$$P(D|-) = 1 - \frac{\text{Especificidad} \times (1 - \text{Prevalencia})}{(1 - \text{Sensibilidad}) \times \text{Prevalencia} + \text{Especificidad} \times (1 - \text{Prevalencia})} = 0.0109\,.$$

- Si Sensibilidad $= 0.97$, entonces

$$P(D|-) = 0.00158\,.$$



7 Informes estadísticos

Según el (ENE-Covid19), la prevalencia actual de anticuerpos IgG contra el SARS-Cov2 en España es aproximadamente 5%. En el estudio participaron más de 60000 individuos y se estimó la prevalencia sobre áreas geográficas y grupos de edad. Junto con cada proporción estimada, el informe refleja un Intervalo de Confianza (IC) al 95%. La fórmula genérica del IC aproximado para una proporción p es

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

donde $\hat{p}$ es la *proporción muestral* (cociente de dividir el número de individuos en la muestra con anticuerpos entre el número total de individuos en la muestra), n es el *tamaño muestral* (número de individuos en la muestra), y 1.96 es el cuantil de la distribución normal estándar que tiene asociada una probabilidad igual a $0.025 = 0.05/2$ para su cola superior, de tal modo que la probabilidad de que una variable aleatoria normal estándar tome un valor entre -1.96 y 1.96 es 0.95 (nivel de confianza). El valor $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ es el *error estándar* del estimador, que cuantifica su precisión, mientras que el *error estándar relativo* cuantifica la precisión relativa del estimador (por tanto puede reflejarse como un porcentaje) y se obtiene dividiendo el error estándar entre la proporción muestral. Observa que el número de habitantes que tiene España (tamaño poblacional) no aparece ni en la fórmula del IC, ni en la del error estándar, ni en la del error estándar relativo. Es decir, la precisión del estimador depende del número de observaciones disponible, pero no del tamaño de la población que queremos estudiar.

Si la proporción de individuos con anticuerpos es $p = 0.05$, el error estándar de $\hat{p}$ para $n = 60000$ será aproximadamente 1.78%, mientras que la amplitud del IC para p sería aproximadamente $2 \times 1.96 \times 1.78 = 7\%$ de $\hat{p}$. El IC global que refleja el informe es $[0.047, 0.054]$, cuya amplitud es el 14% de $\hat{p}$. ¿Por qué nos dan un IC más amplio de lo esperado? A continuación reflejamos algunas consideraciones relevantes sobre la precisión de la estimación de la proporción de individuos con anticuerpos.

- El estudio fue realizado sobre *hogares*. Para cada hogar seleccionado, se realizaron pruebas diagnósiticas a todos los individuos del mismo. Cuando uno de ellos sufre de COVID-19, debido a su alto nivel de contagio, es muy probable que los demás miembros del hogar también se contagien. Esto es equivalente a reducir el tamaño muestral, con lo que el efecto final es que el error estándar del estimador aumenta. Los hogares españoles, tienen, en promedio, 2.5 mienbros. Al dividir el tamaño muestral entre 2.5 se incrementa el error estándar (y la amplitud del CI) de $\hat{p}$ por un factor de $\sqrt{2.5}$, aproximadamente el 58%.
- Para cada individuo del estudio, pese a que el resultado de su test haya sido negativo, existe la posibilidad de que sea portador y que su resultado haya sido un falso negativo. De este modo, la proporción de individuos con anticuerpos no es la proporción de pruebas positivas, sino que es $1/0.79 = 1.266$ multiplicado por la proporción de pruebas positivas. El efecto final es un incremento en el error estándar (y amplitud del IC) de $\hat{p}$ por un factor de 1.266, aproximadamente el 27%.

Estas dos observaciones explican, de un modo bastante preciso, la amplitud de los ICs del estudio. No obstante, queda por realizar un comentario relevante sobre el tipo de muestreo realizado (que también afecta a la precisión del estimador).

- Para seleccionar los individuos que tomaron parte en el estudio, se realizó un *muestreo estratificado*. Esto quiere decir que se formaron subgrupos de la población (*estratos*) y el muestreo se realizó en cada uno de los mismos. Como la proporción de portadores de anticuerpos no es la misma en todas las provincias españolas y el Instituto Nacional de Estadísitica (INE) tiene registros precisos de la población de cada provincia, se decidió estimar en primer lugar la proporción de portadores de anticuerpos en cada provincia tomando un número de observaciones proporcional a su población y obtener luego la estimación global de la proporción de portadores ponderando cada estimación provincial proporcionalmente al número de habitantes de la provincia. Este procedimiento sirve para reducir el error estándar de la estimación global.

8 Exceso de mortalidad

Se ha suscitado una fuerte polémica en torno al número oficial de fallecidos por COVID-19 en España. En una pandemia como ésta no es posible recoger de un modo fidedigno el número de muestres diarias. No obstante, a largo plazo, será posible cuantificar el número aproximado de fallecimientos con una causa relacionada con la COVID-19. Esto se hace modelando el número de muertes en España con una serie temporal y obteniendo el exceso de mortalidad. El número diario de muertes en España se recoge (con una ligera demora) en la web MoMo del Instituto de Salud Carlos III que permite realizar la comparación directa entre el número de muertes durante la pandemia de la COVID-19 con el número esperado de muertes en el mismo período.

Observa que la mortalidad se modela utilizando una serie temporal y, como tal, tiene *tendencia* (a crecer o decrecer a lo largo del tiempo, en este caso debido al crecimiento vegetativo, envejecimiento de la población, cambios en la esperanza de vida,...) y *estacionalidad* (patrones en intervalos regulares, la mortalidad depende de la temperatura, y por tanto de la época del año). En los días de mayor mortalidad, fue muy difícil evaluar la causa de muchos fallecimientos, pero a largo plazo puede calcularse el exceso de mortalidad. Observa que la COVID-19 aumenta el número de muertes, pero viene acompañada de otros factores que la reducen (reducción de la actividad debido al confinamiento) o la aumentan (efecto cosecha, que provoca un aumento en el número de muertes, a corto plazo, entre los que ya están enfermos).