Análisis de la evolución de la incidencia de la COVID-19 en España

Juan Matorras Díaz-Caneja

30/11/2020

Introducción

Este es un ejercicio básico de análisis de los datos de la incidencia de la COVID-19 en España a lo largo de 2020. Habiendo la cantidad de informes y herramientas para el análisis de los datos sobre la incidencia de la COVID-19 que ya existen, este documento no pretende aportar nada singularmente nuevo y su razón de ser no es otra que poner en práctica y profundizar por mi parte en el aprendizaje de las técnicas de análisis de datos y el lenguaje R que he iniciado en la segunda mitad de septiembre de 2020.

Los datos de partida son los publicados por el Gobierno de España en la web **datos.gob.es** a través del siguiente enlace: https://datos.gob.es/es/catalogo/e05070101-evolucion-de-enfermedad-por-el-coronavirus-covid-19.

El grueso del informe se centra sobre los totales en España agregando los datos disponibles por Comunidades Autónomas, aunque también se muestran información de las CCAA de Madrid y Cantabria. La razón de la selección de estas dos comunidades y no otras es simple y llanamente que nací y crecí en la última, mantenniendo allí vínculos familiares y de amistad, mientras que en la primera he pasado prácticamente la mitad de mi vida, sigo viviendo en ella y previsiblemente así seguirá siendo en los próximos años.

Proceso metodológico y software utilizado

El archivo de datos no ha sido sometido a ningún tipo de modificación o alteración previa y su manipulación en este análisis es el mínimo imprescindible para permitir el tratamiento de los datos y obtención de resultados.

La fecha y hora de descarga de los datos que han sido utilizados para las tablas y gráficos incluídos en este informe ha sido (aaaa-mm-dd hh:mm:ss): **2020-11-30 15:15:05**

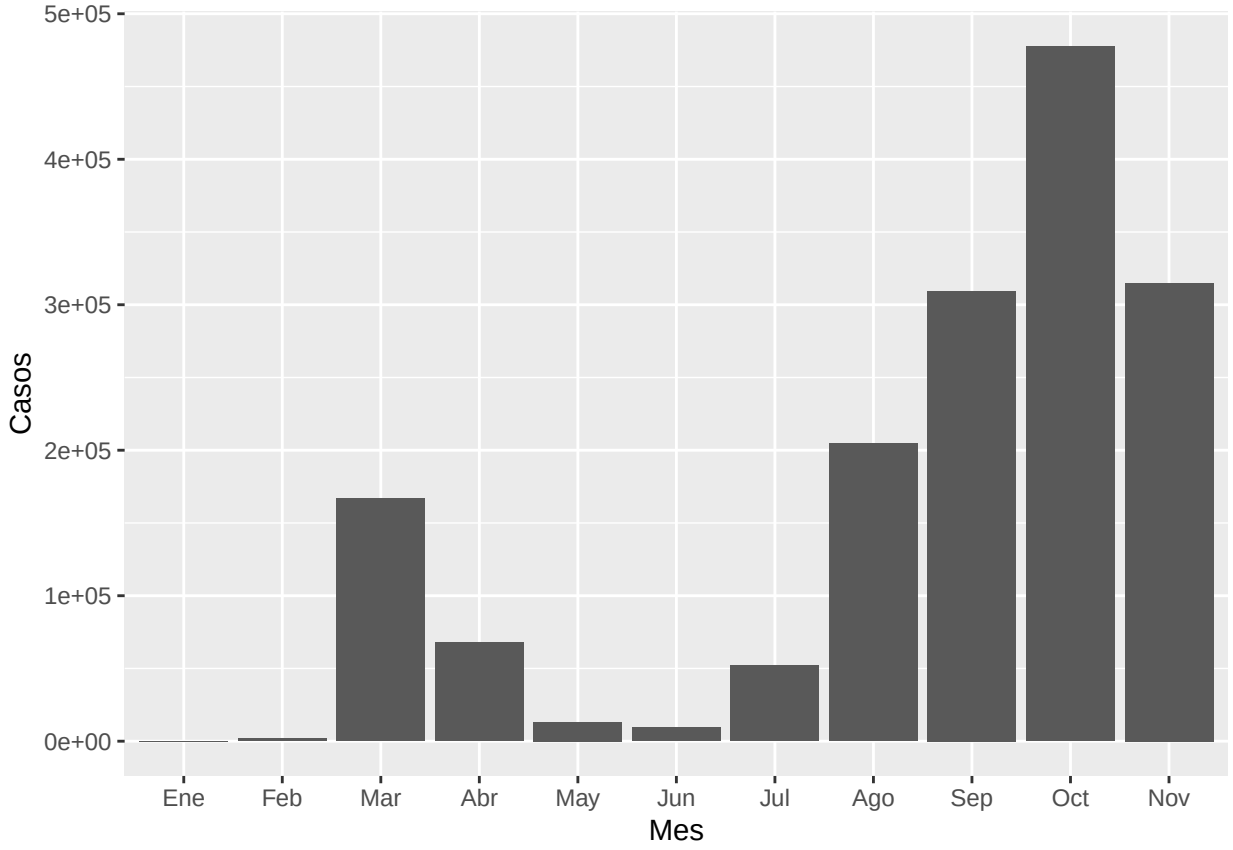
El análisis se ha llevado a cabo utilizando el software libre para análisis estadístico **R**, versión 4.0.2. (1)

Se ha hecho uso también de los paquetes complementarios:

- **lubridate** para facilitar el manejo de fechas. (2)
- **knitr** para mejorar la apariencia de tablas. (3)
- **tydiverse** por los paquetes que incluye para ayudar en la extracción de la información y los gráficos mejorados de **ggplot2**. (4)
- **data.table** con el objeto de manipular las tablas más eficientemente. (5)

Incidencia mensual y número total de casos detectados desde el inicio de 2020

La evolución de número de casos notificados por meses se refleja en el gráfico que se muestra a continuación:



Correspondiente a los valores que se incluyen en la tabla siguiente:

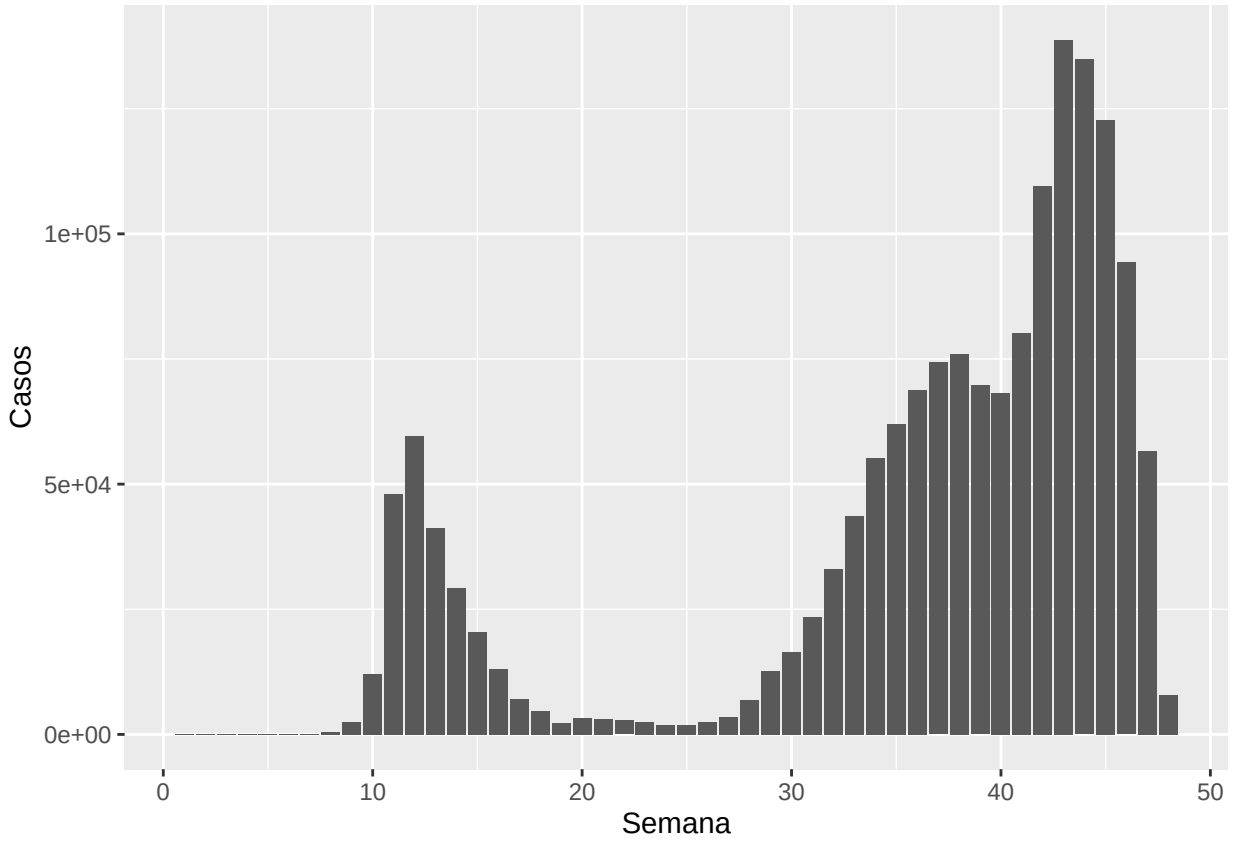
Mes	Nº casos mensuales
Ene	45
Feb	1.745
Mar	166.759
Abr	67.800
May	13.306
Jun	9.423
Jul	52.186
Ago	204.433
Sep	309.355
Oct	477.610
Nov	314.842

El número total de casos acumulados desde el 1 de enero de 2020 hasta la fecha indicada en el punto anterior según los datos oficiales disponibles en ese momento ascienden a un total de **1.617.504**.

Considerando una población en España de **47,33** millones de personas según los datos publicados por el INE (Instituto Nacional de Estadística) correspondientes al inicio del año 2020, el porcentaje de contagio de la población es del **3,418 %** hasta la fecha.

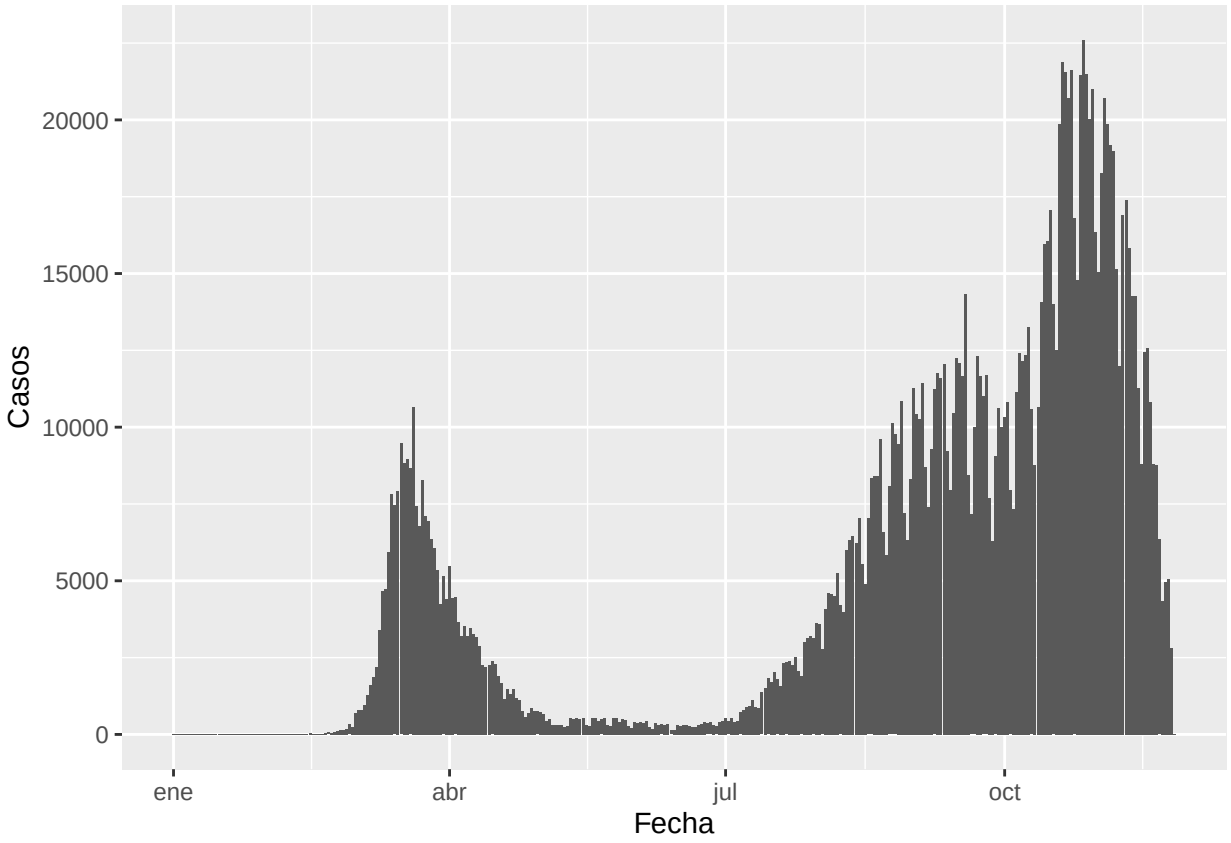
Incidencia semanal

- Evolución de número de casos identificados por semanas:

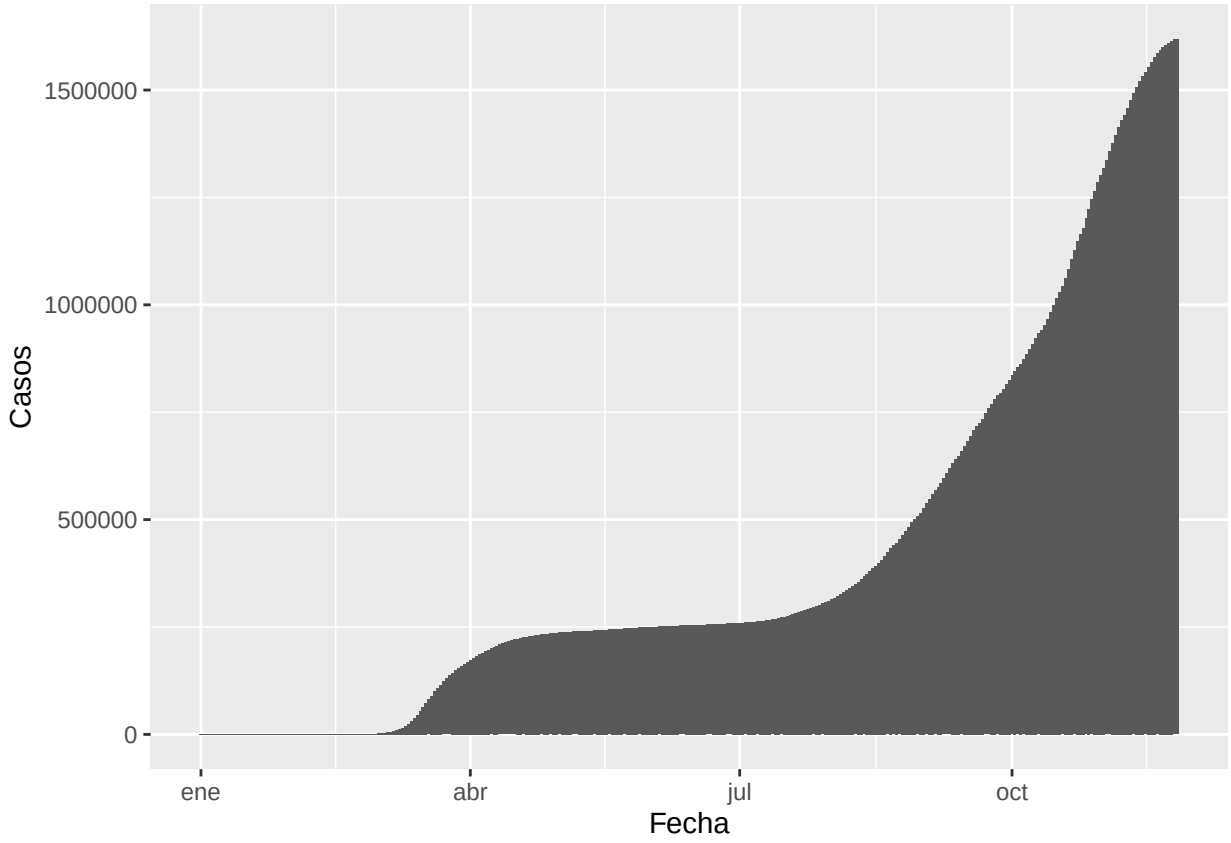


Incidencia diaria

- Curva epidémica de los casos notificados por días:

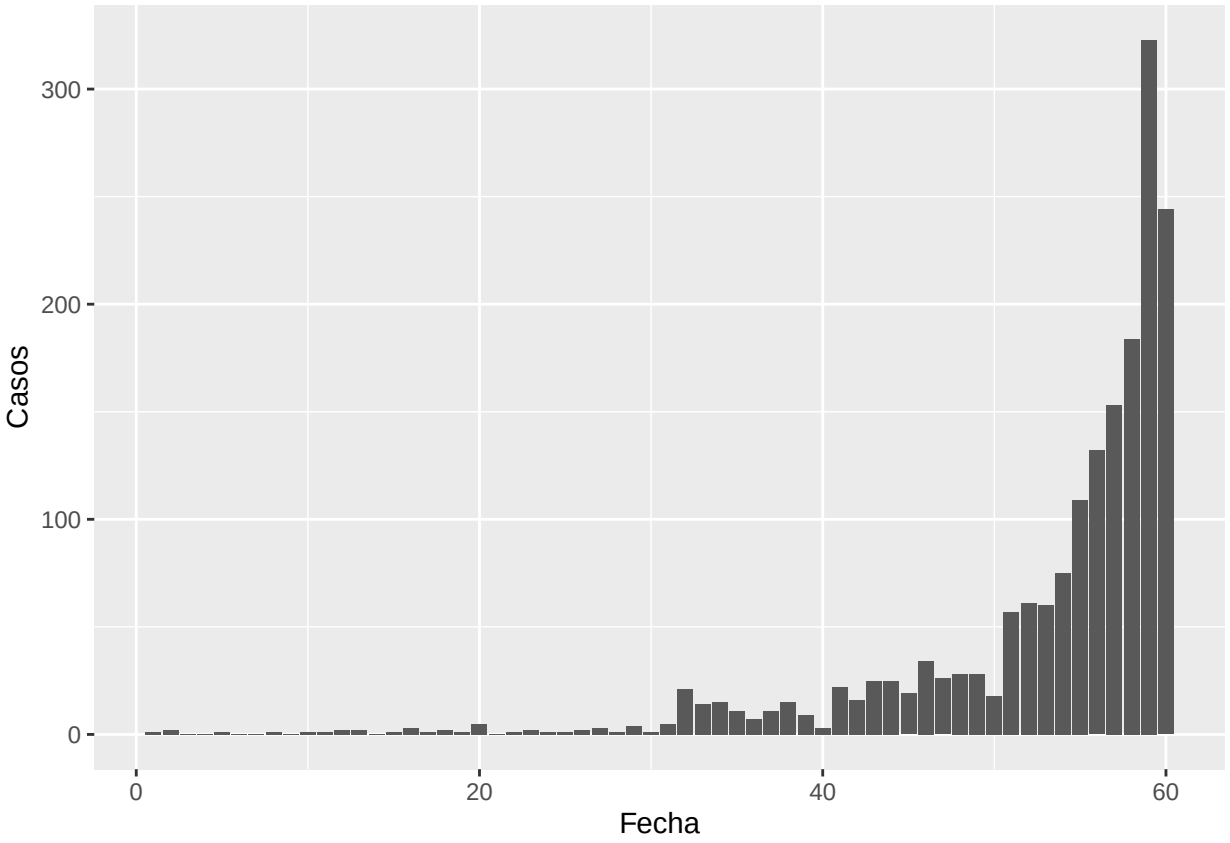


- Gráfico de casos acumulados a origen por días:



Detalle del número de casos en los dos primeros meses de 2020

- Evolución diaria del número de casos durante los dos primeros meses del año:



Los casos reportados totales a lo largo de esos dos meses son **1.790**, si bien es claro que se produce una acusada inflexión en la pendiente de crecimiento a partir del día 50.

Siendo así que el desglose de número agregado de casos identificados en dichos primeros 50 días y los siguientes 10 días queda de la siguiente manera:

- Periodo 1-50: 392
- Periodo 51-60: 1.398

En 10 días se detectan **3,6** veces los casos que se habían producido en los 50 días anteriores.

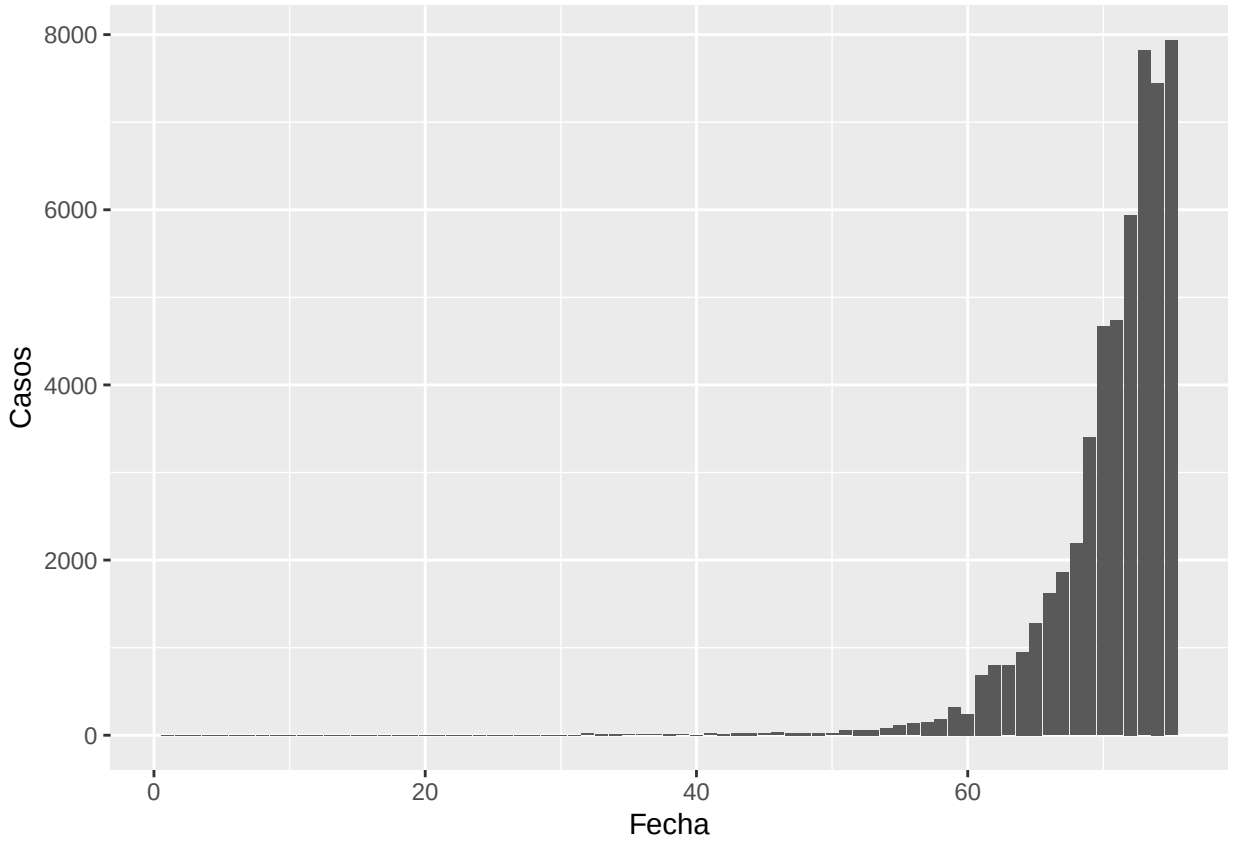
Subsiguiente evolución durante la primera quincena de marzo

En este apartado analizamos cómo continúa desarrollándose la propagación de la pandemia a principos del mes de marzo, estableciendo por su relevancia en lo ocurrido en España durante esos días dos periodos de tiempo diferenciados, del 1 al 8 y del 9 al 15.

En los primeros ocho días de marzo la progresión diaria de nuevos casos siguió disparándose, resultando un total de **10.186** casos a añadir al total anterior, siendo éstos **5,7** veces los registrados a lo largo de todo enero y febrero.

Durante los siguientes siete días, del 9 al 15 de marzo, los casos a sumar fueron **41.939**, lo que supone **4,1** veces los notificados en los 8 primeros días del mes.

- Gráfico del número de casos diarios desde el 1 de enero hasta el 15 de marzo de 2020:



Incidencia acumulada por 100.000 habitantes en los 14 días previos a la declaración del estado de alarma del 14 de marzo

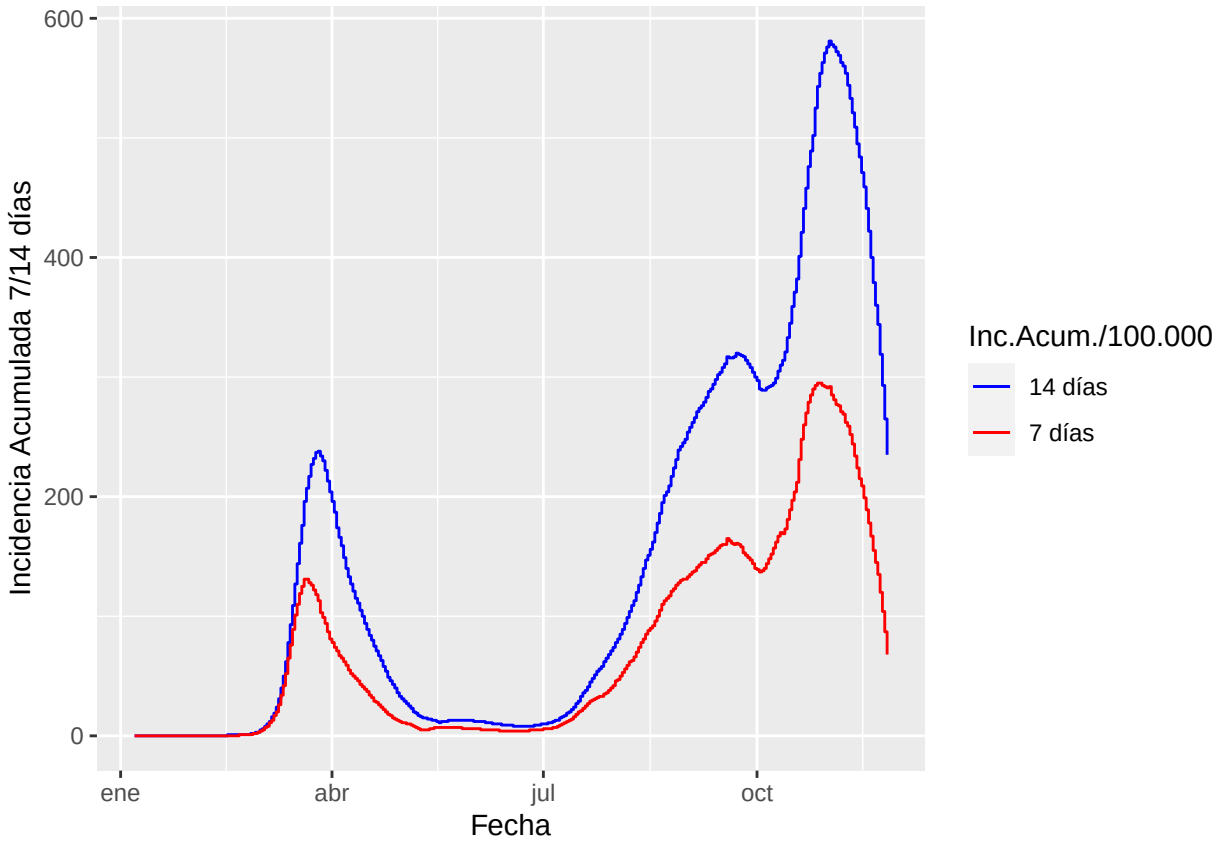
Pasemos ahora a calcular la incidencia acumulada por cada 100.000 habitantes en los 14 días previos a la declaración del estado de alarma que tuvo efecto

Tomando esos 14 días previos, es decir, entre el 29 de febrero y el 13 de marzo, la incidencia acumulada por cada 100.000 habitantes, con el mismo dato de población presentado más arriba fue de **78** casos/100.000 hab.

Contrasta este valor de forma muy llamativa con los límites que se han estado manejando en la segunda ola de infecciones, donde se ha hablado de 200, 500 e incluso 1.000 casos/100.000 hab.

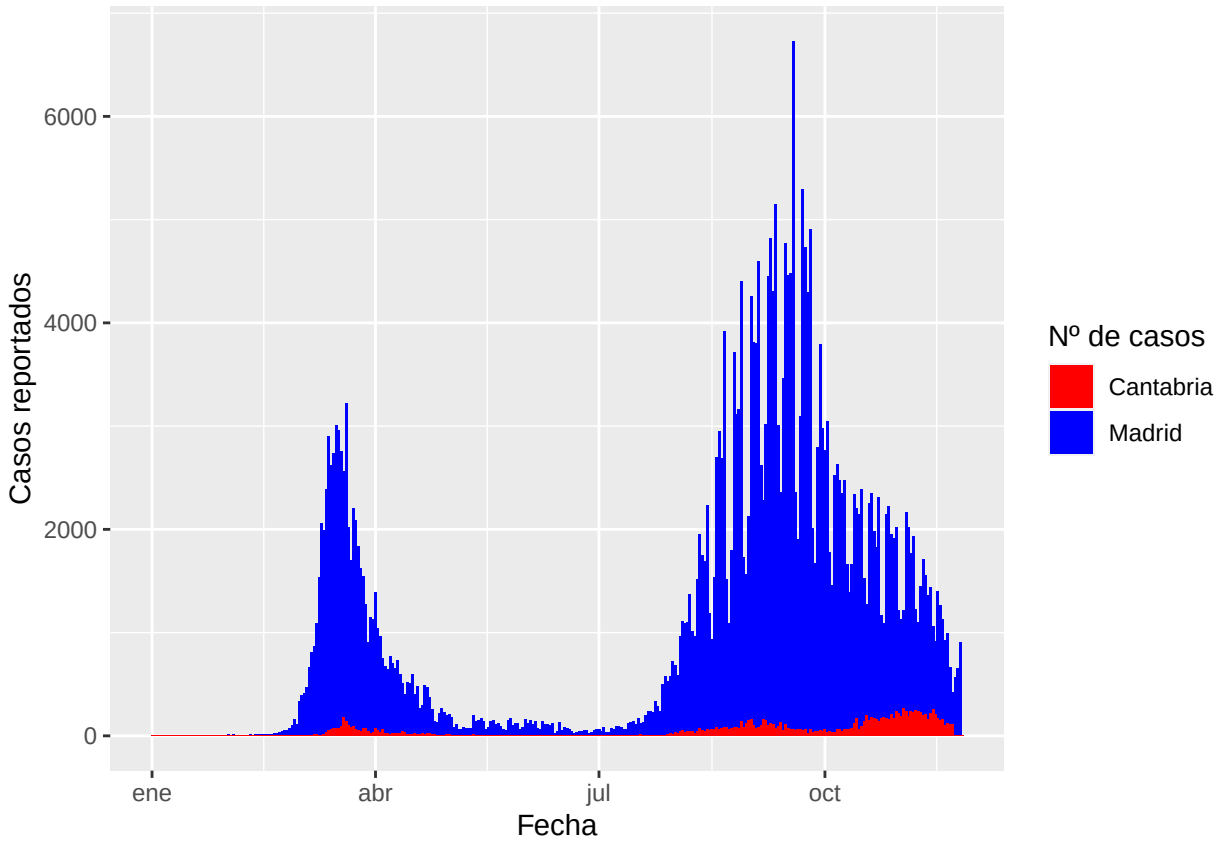
Evolución de la incidencia acumulada a lo largo de todo el año

En el siguiente gráfico se representan las incidencias acumuladas por cada 100.000 habitantes correspondientes a periodos de 14 y 7 días:



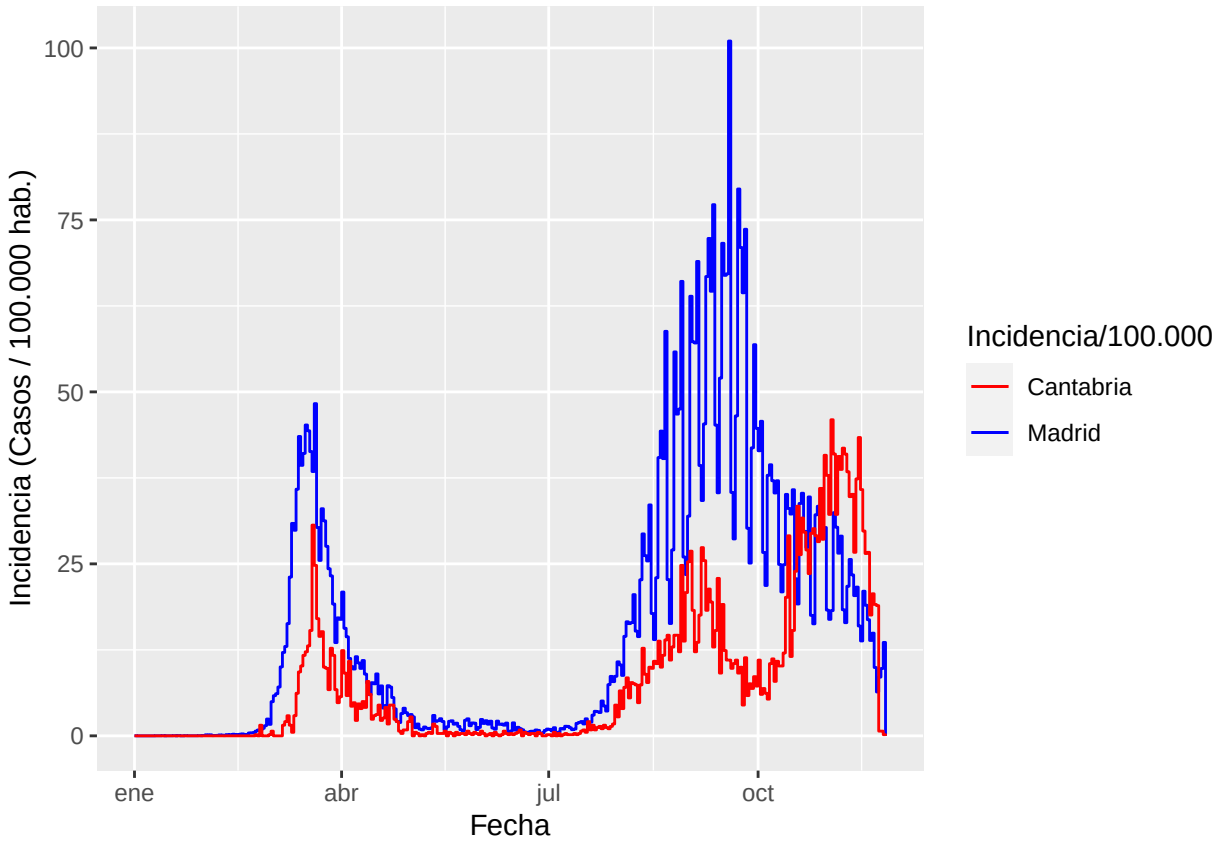
Comparación de la evolución del número de casos entre Cantabria y Madrid

En el siguiente gráfico se compara la evolución de la enfermedad entre dos comunidades muy diferentes, Cantabria y la Comunidad Autónoma de Madrid:

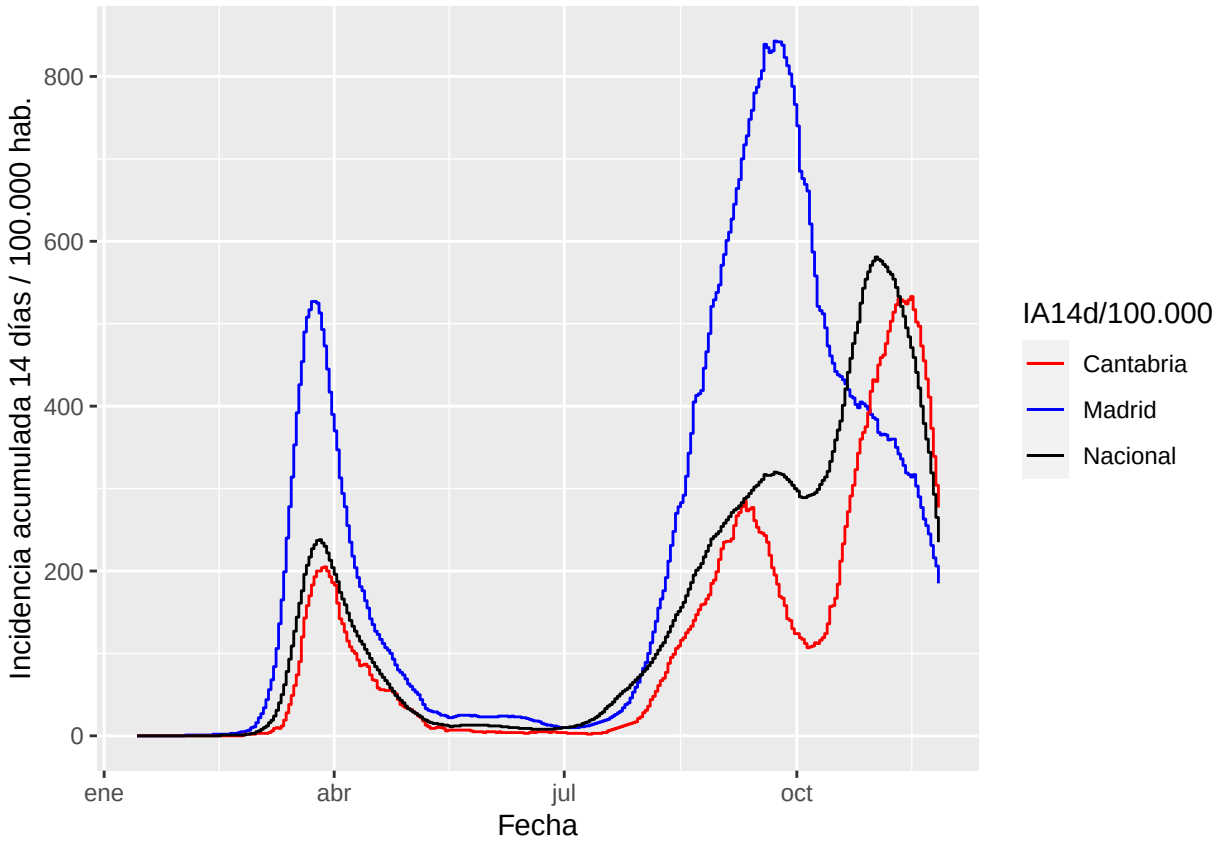


Como es lógico, los datos no son comparables en términos absolutos por la gran diferencia de población entre ambas comunidades autónomas, por no entrar en la forma de vida en una y otra y en cómo eso impacta en la dispersión de la enfermedad.

Para solventar este problmea reprentamos ahora número de casos por cada 100.000 habitantes, con los datos de población en cada comunidad disponibles en el momento en el INE, que corresponden a 2019, reflejando los datos de Cantabria en color rojo y los de Madrid en azul:



Por completar la información comparativa entre ambas comunidades se adjunta también la incidencia acumulada en 14 días para ambas áreas geográficas, junto con la correspondiente al conjunto del territorio nacional:

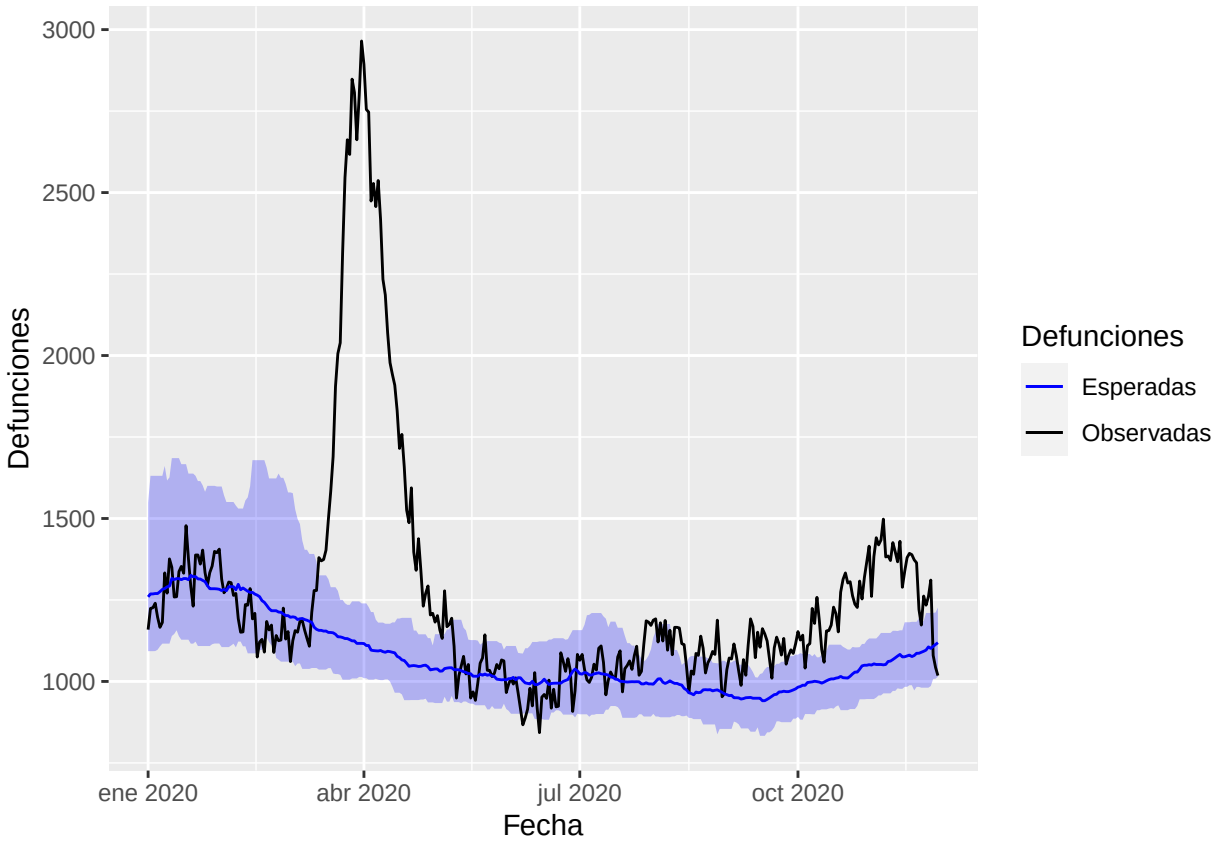


Exceso de mortalidad

Como último paso del análisis obtendremos cifras del exceso de mortalidad registrado en este año, presumiblemente achacable a la incidencia de la pandemia de la COVID-19. Los datos se han obtenido del enlace del **Instituto de Salud Carlos III**: https://momo.isciii.es/public/momo/data.

La fecha y hora de descarga de los datos de mortalidad utilizados para la elaboración de los siguientes gráficos y tablas fue (aaaa-mm-dd hh:mm:ss): **2020-11-30 15:13:41**

Representemos en primer lugar la evolución del número de defunciones en comparación con las esperadas y su rango para los percentiles 1 y 99:

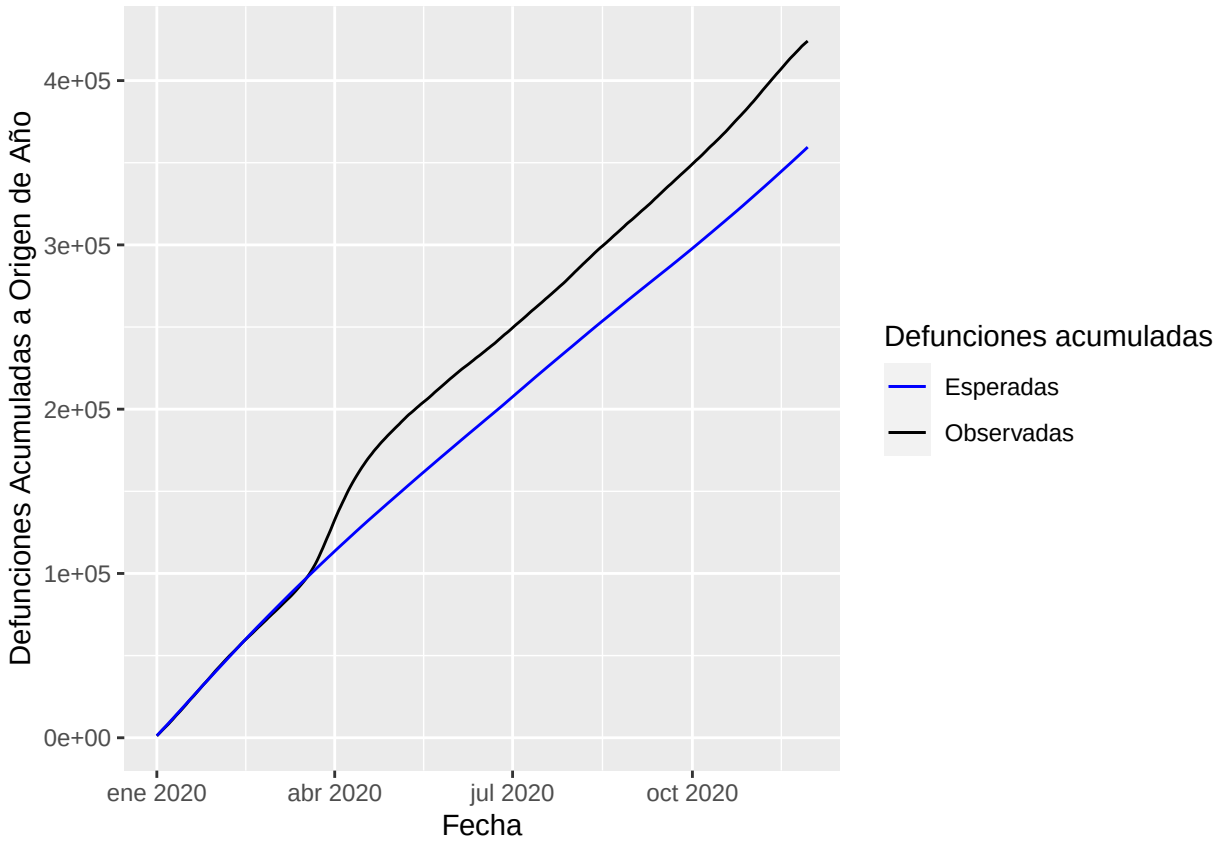


Como se puede ver, existe un periodo de mortalidad totalmente disparada a lo largo de los meses de marzo a mayo, mientras que luego se aprecia otro periodo de desviación al alza, no tan acusado pero más prolongado en el tiempo y con una tendencia creciente, que cubríría desde agosto hasta bien avanzado octubre.

Técnicamente se define "periodo de exceso de mortalidad" cuando se cumplen las siguientes condiciones:

- Se observa al menos dos días consecutivos con defunciones observadas por encima del percentil 99 de las estimadas.
- La fecha de inicio del periodo es el primer día con las defunciones observadas por encima de las estimadas.
- La fecha de fin del periodo es el último día con las defunciones observadas por encima de las estimadas.
- Si entre la fecha de fin de un periodo y la fecha de inicio del siguiente hay dos días, se unifican ambos periodos, tomando la fecha de inicio del primer periodo y fecha de fin del último.

Con estas premisas podemos aislar los periodos en los que se han producido dichas circunstancias y calcular el exceso de defunciones durante esos lapsos de tiempo concretos. Ahora bien, antes de pasar a realizar dichos cálculos, realicemos uno más básico, comparando directamente las cifras de defunciones esperadas acumuladas a lo largo del año con las que realmente se han registrado:



Como podíamnos esperar, esta gráfica no aporta gran valor a la hora de la interpretación de la información, más allá del hecho de que las defunciones observadas se despegan de las esperadas de forma muy visible a lo largo de los meses de marzo a mayo, y que dicho distanciamento se vuelve a incrementar, ya a menor ritmo, a partir del mes de agosto, aunque vuelve a repuntar en noviembre.

Más interesentante resulta la comparación directa de las cifras acumuladas hasta la fecha. En este caso tenemos, con los datos disponibles, una total de **424.094** defunciones observadas y **359.518** defunciones esperadas, resultando un exceso de **64.576** defunciones. Expresando dicho exceso en términos porcentuales, nos encontramos con un **18 %** más fallecimientos de los esperados.

Por afán de completar la visión de la evolución de estas variables, presentamos a continuación esos mismos valores en la fecha en la que se levantó el estado de alarma, 21 de junio de 2020, buscando una fecha en lo que podríamos denominar **"final de la primera ola"**, que no ***"la derrota de la pandemia"*** (sic):

- Defunciones acumuladas observadas: **239.309** personas
- Defunciones acumuladas esperadas: **197.398** personas
- Esceso de defunciones: **41.911** personas
- En tanto por ciento: **21,2 %**

Retomando la senda de la ortodoxia y aplicando ahora sí los criterios técnicos "oficiales" que presentábamos más arriba que definen los periodos de exceso de mortalidad, las fechas que delimitan el principio y final de los periodos de exceso padecidos a lo largo de 2020 son:

- Antes de unificar periodos de exceso próximos:

Inicio	Fin
2020-03-10	2020-05-09
2020-07-20	2020-08-29
2020-08-31	2020-11-26

- Después de unificar los periodos de exceso cercanos (<= 2 días):

Inicio	Fin
2020-03-10	2020-05-09
2020-07-20	2020-11-26

Los excesos de defunciones en estos 2 periodos son:

Inicio	Fin	Exceso de defunciones
2020-03-10	2020-05-09	44.599
2020-07-20	2020-11-26	22.443

Siendo el total agregado de exceso de defunciones de **67.042** personas.

Expresándolo en términos porcentuales, el exceso de defunciones es un **18,6 %** superior al total de las esperadas hasta la fecha. Como es lógico, este valor porcentual se irá reduciendo a medida que transcurra el tiempo desde el final del último episodio de exceso de defunciones.

Aunque en el exceso de defunciones haya casos de fallecimiento no directamente imputables a la COVID-19, hay que asignar dichas muertes a la crisis del COVID-19. Si determinadas patologías no son debidamente atendidas en tiempo y forma por la sobrecarga del sistema sanitario provocado por la pandemia, los fallecimientos asociados a las mismas son por tanto atribuibles a la COVID-19 aunque el virus no haya tenido presencia en el paciente correspondiente.

No podemos dejar de llamar la atención sobre el hecho de que en la determinación de las cifras de exceso de defunciones se ha utilizado como nivel de referencia el número de defunciones esperadas. Es perfectamente argumentable que durante el periodo de estado de alarma este nivel de comparación debería ser inferior ya que el propio estado de alarma tuvo por necesidad incidencia negativa en el número de fallecimientos por accidente laboral y por accidente de tráfico, teniendo que rebajarse por tanto el patrón de referencia de defunciones durante el estado de confinamiento y arrojando un exceso de defunciones por causa de la COVID-19 superiores a los mostrados más arriba. Aunque es posible realizar estimaciones de estas desviaciones con datos disponibles públicamente, dejamos esa posibilidad de perfeccionamiento del estudio para mejor oportunidad.

. . . . . .

Referencias

(1) R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

(2) Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL: http://www.jstatsoft.org/v40/i03/

(3) Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

(4) Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

(5) Matt Dowle and Arun Srinivasan (2020). data.table: Extension of `data.frame`. R package version 1.13.2. https://CRAN.R-project.org/package=data.table