Análisis de la evolución de la incidencia de la COVID-19 en España desde el 1 de enero de 2020 hasta la actualidad

Una exploración de los datos de casos detectados de la COVID-19 y la mortalidad en España durante la pandemia mediante R y R Markdown (RStudio ®)

Juan Matorras Díaz-Caneja

04/02/2021

1.- Introducción

Este es un ejercicio básico de análisis de los datos de la incidencia de la COVID-19 en España desde el comienzo del año 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 inicié en la segunda mitad de septiembre de 2020.

Hay que matizar que, si bien este estudio empezó cubriendo únicamente fechas dentro de 2020, por la propia duración de la pandemia, el alcance del mismo se ha extendido para acomodar los datos disponibles correspondientes ya al año 2021.

Por su propia esencia, éste es un documento vivo que además der ser puesto al día periódicamente con los nuevos datos disponibles, va sufriendo adiciones, modificaciones y correcciones de erratas. La última versión disponible de este documento y de los datos empleados en su elaboración se pueden encontrar en el repositorio de GitHub: https://github.com/JuanMatorras/Covid-19.

Los datos de partida son los publicados por el Gobierno de España. Estos datos estaban durante 2020 disponibles directamente a través la web **datos.gob.es** en el enlace: https://datos.gob.es/es/catalogo/e05070101-evolucion-de-enfermedad-por-el-coronavirus-covid-19. Con el cambio de año este enlace ha quedado anulado y es necesario acudir a: https://cnecovid.isciii.es/covid19/#documentación-y-datos donde se encuentran los nuevos enlaces de descarga. En este caso el archivo utilizado es el de datos por CCAA: https://cnecovid.isciii.es/covid19/resources/casos_tecnica_ccaa.csv

El grueso del informe se centra sobre los totales en España agregando los datos disponibles por Comunidades Autónomas, aunque también se muestran información de la incidencia en las CCAA de Madrid, Cantabria, Asturias, Galicia y Castilla y León. La razón de la selección de estas comunidades y no otras responde a consideraciones personales y no obedece a ningún criterio técnico.

2.- 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.

Al ejecutar el código se descargan los archivos de datos directamente de la web si no se encuentran ya disponibles en el directorio /data.

La fecha y hora de descarga de los datos que han sido utilizados para la elaboración de las tablas y gráficos incluidos en este informe ha sido (aaaa-mm-dd hh:mm:ss): **2021-02-04 17:59:27**

El análisis se ha llevado a cabo utilizando el entorno de desarrollo integrado de **RStudio** © versión 1.4.1103 (1) para el software libre de análisis estadístico **R**, versión 4.0.3 (2).

Se ha hecho uso también de los siguiente paquetes complementarios para R:

- **lubridate** ver.1.7.9.2 para facilitar el manejo de fechas. (3)
- **knitr** ver.1.31 para mejorar la apariencia de tablas. (4)
- **tidyverse** ver.1.3.0 por los paquetes que incluye para ayudar en la extracción y manipulación de la información en tablas y los gráficos mejorados de **ggplot2** ver.3.3.3. (5) y (6)
- **data.table** ver.1.13.6 con el objeto de manipular las tablas más eficientemente. (7)

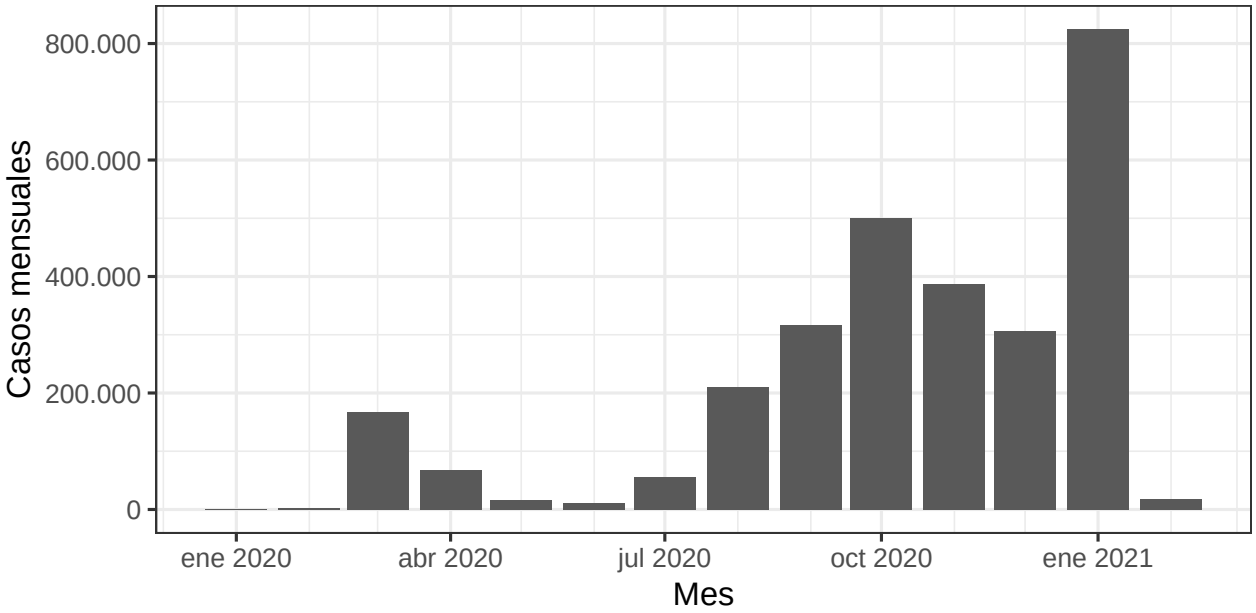
Nota 1: La codificación empleada en los datos fuente para la identificación de comunidades autónomas y provincias se corresponde con la ISO 3166-2:ES. La correspondencia entre dicha codificación y la comunidad autónoma o provincia referida puede ser consultada, entre otras fuentes, en: https://es.wikipedia.org/wiki/ISO_3166-2:ES

Nota 2: En algunos puntos la programación puede ser lejos de óptima, en primera instancia por mi falta de fluidez en la programación con R por estar en proceso de aprendizaje, pero también afectado significativamente por la manera en la que se ha ido construyendo este documento, que empezó siendo algo sencillo, muy limitado en su alcance, pero que ha sufrido múltiples ampliaciones en su alcance en sucesivas etapas y nunca siguiendo un guión definido. Un ejemplo claro son los apartados donde se compara la información entre CCAA. En primera instancia empezaron siendo sólo dos comunidades y por eso se programó de una manera, que se debería haber modificado cuando la lista empezó a hacerse larga a medida que incorporaba nuevas CCAA a los análisis comparativos.

Nota 3: Hay que llamar la atención sobre la naturaleza de los datos, los cuales proceden de la agregación de las aportaciones desde las CCAA al sistema RENAVE. Esto resulta en que los datos de casos de fechas recientes no están completos y esto hace que todas las gráficas con desagregación diaria o variables derivadas, como la incidencia acumulada, presenten siempre un tramo descendente en su extremo final, con independencia de si realmente el progreso de la enfermedad es de expansión o de contracción.

3.- Casos por meses 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	Casos
ene 2020	613

Mes	Casos
feb 2020	1.802
mar 2020	167.228
abr 2020	68.253
may 2020	15.464
jun 2020	10.727
jul 2020	56.067
ago 2020	210.081
sep 2020	316.884
oct 2020	500.060
nov 2020	387.271
dic 2020	306.253
ene 2021	824.447
feb 2021	18.366

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 **2.883.516** personas.

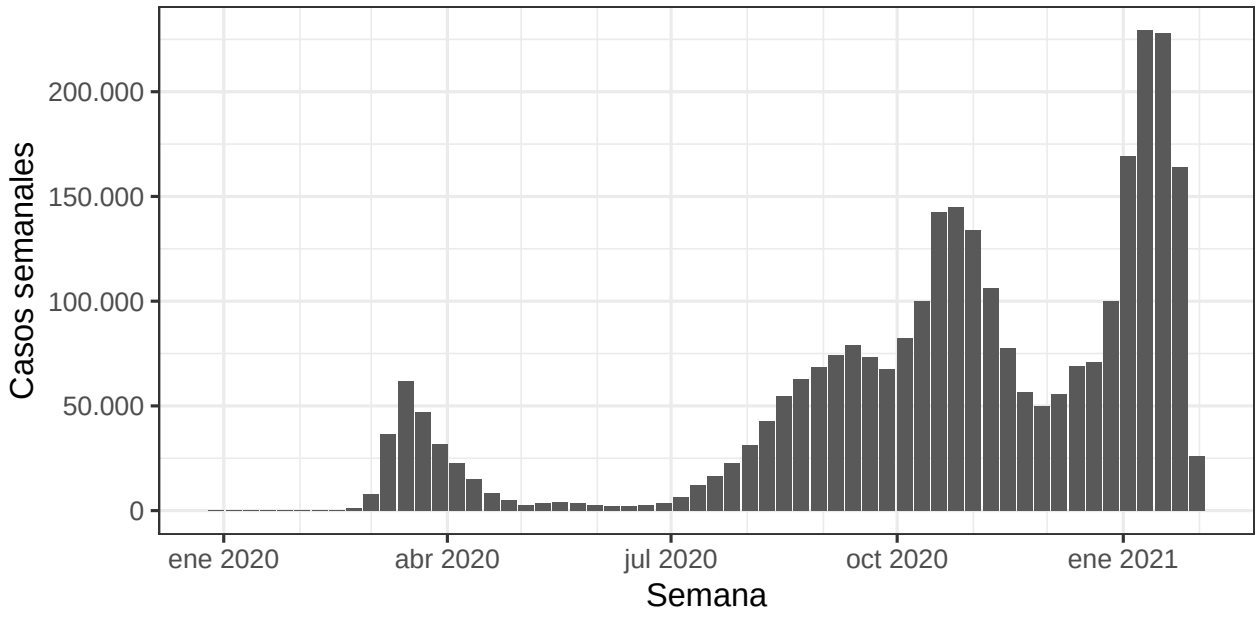
Nota: Véase el apartado 6 donde se explican los errores que están apareciendo en los datos correspondientes a enero de 2020 ya que se han estado cometiendo errores en el registro de nuevos casos en enero de 2021 que se han asignado incorrectamente al año 2020. De hecho, hasta final de 2020, el número total de casos reportados a nivel nacional para el mes de enero de ese año se movía por debajo de los 50 casos mientras que con los datos disponibles en los primeros días de enero de 2021 esta cifra fue aumentando llegando a alcanzar más de 550 casos.

Lo que en cualquier caso es de conocimiento general es la carestía de pruebas diagnósticas durante la primera fase de la pandemia, lo que apunta a una clara subestimación de los casos reales en esos primeros meses de 2020, particularmente en marzo y abril. Cabría intentar estimar el número real de casos en estos primeros meses aplicando el ratio de pruebas con resultado positivo respecto a pruebas realizadas en la segunda mitad de año al número de pruebas realizadas en dichos meses iniciales pero no deja de ser una aproximación un tanto burda y harían falta datos fiables del número de pruebas que se han realizado, tanto en el ámbito público como privado, además de necesitarse datos complementarios ya que los casos detectados luego tienen múltiples pruebas de seguimiento.

Considerando una población en España de **47,332** 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 **6,092 %** hasta la fecha. Insistimos en que, puesto que la incidencia de la enfermedad en los primeros meses de 2020 está por fuerza minusvalorada por la escasez de pruebas de diagnóstico, y no pudiendo olvidar que además tenemos el fenómeno de los casos de infección asintomáticos, el porcentaje de población afectada realmente es necesariamente más alto que el reflejado en este cálculo y es más que probable que sea del orden del doble. De hecho, para poder cubrir esta laguna y tener una idea más fiable del verdadero impacto de la enfermedad sobre el total de la población están los estudios de sero-epidemiología que se han venido realizando desde junio de 2020 (https://portalcne.isciii.es/enecovid19/).

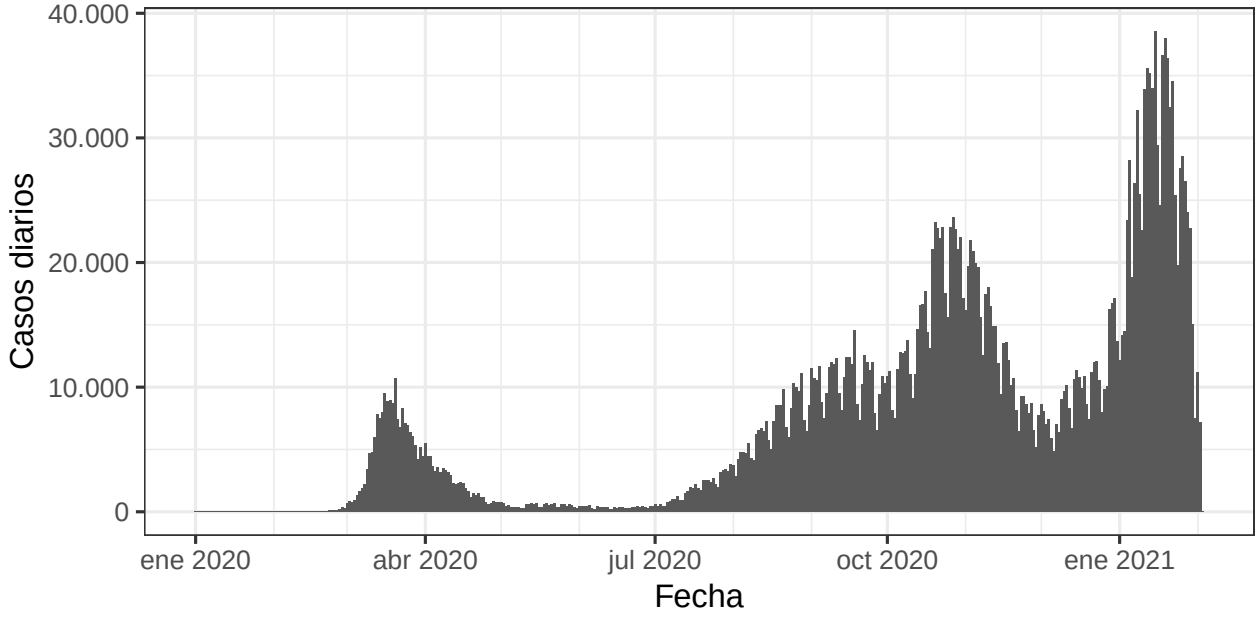
4.- Casos semanales

Evolución de número de casos identificados por semanas:

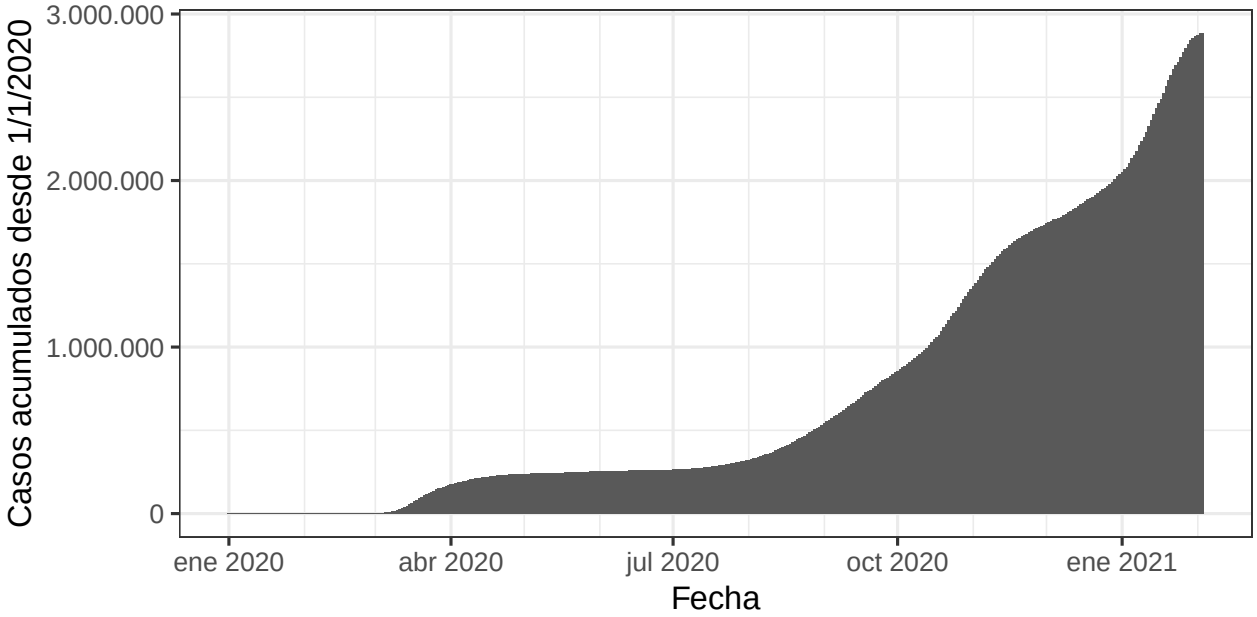


5.- Casos diarios

- Curva epidémica de los casos notificados por días:



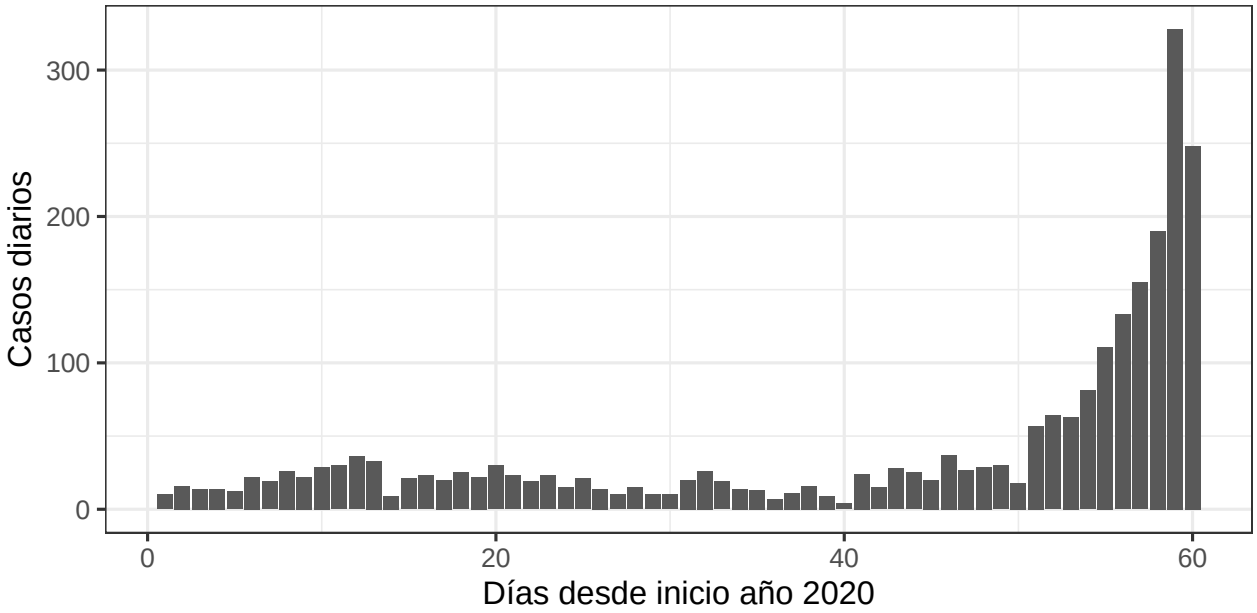
- Gráfico de casos acumulados a origen por días:



El resultado de esta desagregación de datos, o más estrictamente de la ausencia de agregación de los mismos, puesto que los datos de partida son diarios, es la aparición de dientes de sierra que son simplemente el reflejo del descenso en el número de pruebas realizadas en fines de semana y festivos en comparación con los realizados en días laborables.

6.- Detalle del número de casos en los primeros meses de 2020

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



NOTA IMPORTANTE: Este gráfico ha permitido detectar un problema en los datos de origen que se está manifestando en los primeros días de 2021. Están apareciendo casos asignados a los primeros días de 2020 que antes no existían por un problema de deficiente registro de nuevos casos correspondientes realmente a 2021 que se están dando de alta con fecha de 2020. En el repositorio de GitHub al que se hace referencia al principio del documento se adjunta una tabla comparativa donde se puede observar cómo han evolucionado

el número de casos declarados para enero de 2020 en las distintas CCAA entre los días finales de diciembre de 2020 y primeros días de 2021. Las mayores distorsiones se aprecian en los datos de las comunidades autónomas de Madrid y Cataluña en las que para los primeros días de enero de 2020 tenían declarados números de casos con cifras muy bajas, en general 0 o 1, y que durante los primeros días de enero de 2021 empiezan a mostrar cantidades de casos que superan la veintena y la treintena para esas fechas del año 2020. Se puso en conocimiento de este potencial problema al Centro de Estudios Epidemiológicos del Instituto de Salud Carlos III el día 8 de enero de 2021 desde donde confirmaron el problema el día 19 de enero, informando que se esperaba corregir el problema en los días siguientes. Estos datos aparentemente erróneos no se han eliminado ni sufrido ningún tipo de tratamiento especial y se han manipulado como si fuesen correctos en espera de que se complete la revisión y se corrijan los datos fuente.

Los casos reportados totales a lo largo de esos dos meses son **2.415**, 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: 985
- Periodo 51-60: 1.430

En 10 días se detectan **1,5** veces los casos que se habían producido en los 50 días anteriores.

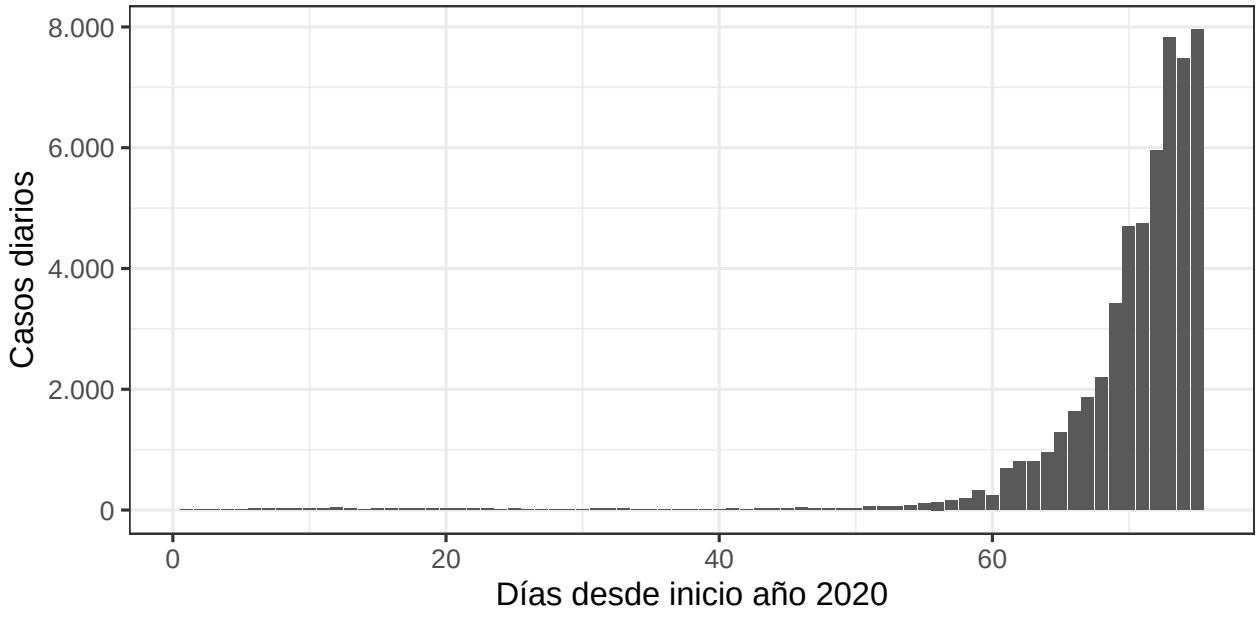
- Subsiguiente evolución durante la primera quincena de marzo:

En este apartado analizamos cómo continúa desarrollándose la propagación de la pandemia a principios 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 **10.239** casos a añadir al total anterior, siendo éstos **4,2** 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 **42.054**, 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

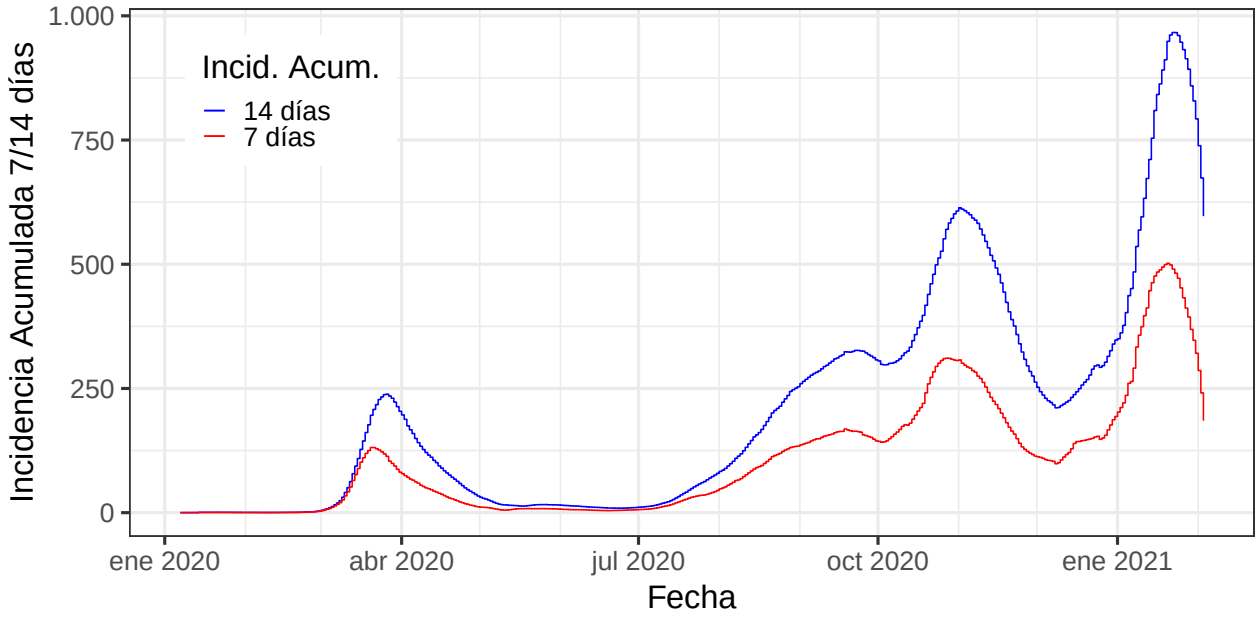
Pasemos ahora a calcular la incidencia acumulada por cada 100.000 habitantes en los 14 días previos a la declaración del estado de alarma que tuvo efecto

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 España durante la segunda ola de infecciones, donde se ha hablado de 200, 500 e incluso 1.000 casos/100.000 hab. para empezar a tomar medidas de limitación de la movilidad y el contacto entre personas, valores totalmente irracionales y desproporcionados en comparación con los manejados por otros países de la Unión Europea.

7.- Evolución de la incidencia acumulada a lo largo de todo el periodo de análisis

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



Detengámonos un momento en este punto para analizar estas curvas, primero comparando la incidencia acumulada por cada 100.000 habitantes en 14 días (IA14) con la acumulada en 7 días (IA7), para luego centrarnos en la propia evolución temporal de estos indicadores.

Dentro de los propios círculos de científicos epidemiológicos existe cierta disparidad de opinión respecto a si hacer el seguimiento de la IA14, la IA7 o incluso ambas en paralelo. En primer lugar no perdamos de vista que el hecho de agregar casos identificados durante un periodo determinado de tiempo, sea éste de 7 días o de 14 días, lo que se persigue con esta operación es laminar los picos que se producen en la detección de casos diarios, principalmente por la influencia de fines de semana y festivos que ya vimos en el punto 5 y así poder determinar con más facilidad la evolución de la propagación de una determinada enfermedad entre la población. La razón para ampliar de 7 a 14 días debería estar en poder absorber mejor las oscilaciones que se provocan por festivos y puentes y no tanto en el mayor o menor periodo de incubación dela enfermedad. De hecho, el efecto laminador de la IA14 frente a la IA7 se puede observar claramente al comparar ambas curvas en el mes de diciembre de 2020. Es más, contrariamente a lo que se pueda presuponer sobre el hecho de que utilizar un periodo más corto de tiempo ayude a tener una mejor idea de la evolución de la enfermedad, esto es solo cierto cuando la curva de incidencia empieza a descender pero no para el momento de aumento de la incidencia. Como se comprueba en la gráfica adjunta, la IA14 empieza a ascender y a mayor ritmo que la

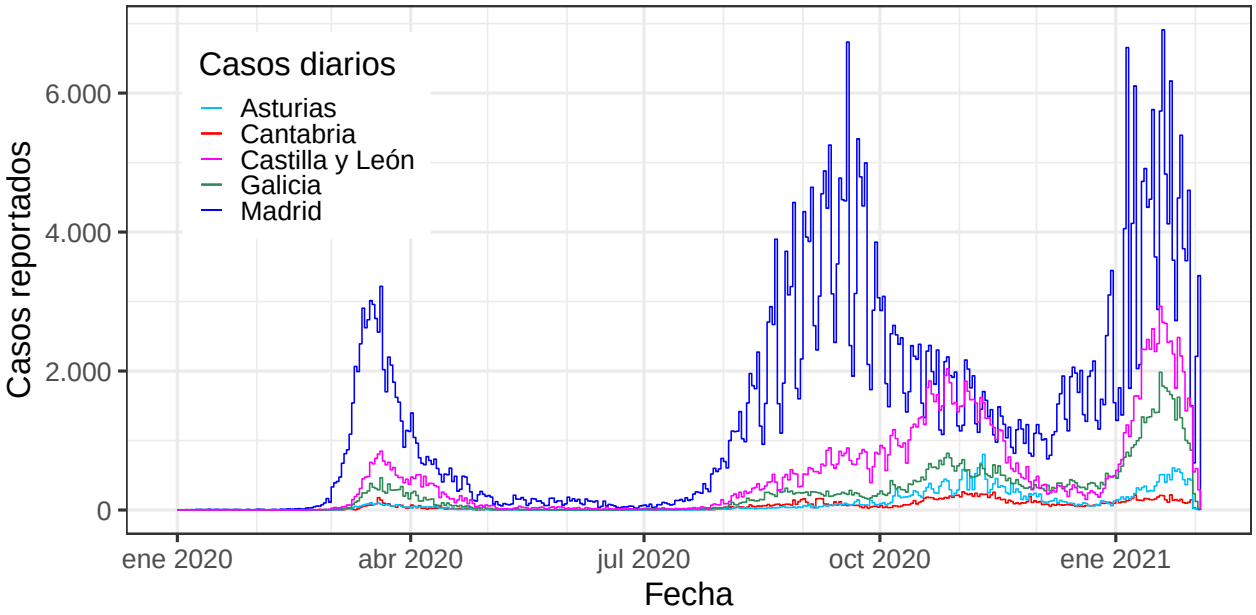
IA7, mientras que el descenso de la primera sí que se demora con respecto a la segunda. Lo que sí es lógico y resulta evidente es que los límites de alarma deben ser distintos para cada una de ellas.

Pasando ahora al análisis de la evolución de ambos indicadores y obviando el hecho ya comentado de la subestimación del primer pico de la ola por la baja realización de pruebas de diagnóstico, queda claro que, a pesar de que los medios de comunicación y, en consecuencia, la propia población general hablen de una tercera ola de la pandemia que arranca en el mes de diciembre, no habiendo bajado la IA14 de 200 sería más apropiado hablar de una segunda fase de la segunda ola puesto que ésta no se puede decir que llegó a estar bajo control. Si analizamos el impacto de la enfermedad desde el punto de vista de la mortalidad (ver apartado 10) sí que podríamos afirmar de algún modo que la segunda ola habría terminado en diciembre, en tanto en cuanto en este mes se cierra el segundo periodo de exceso de mortalidad, pero no así desde el punto de vista de la incidencia acumulada.

8.- Comparación de la evolución del número de casos entre Madrid y otras CCAA

- Casos reportados por CCAA

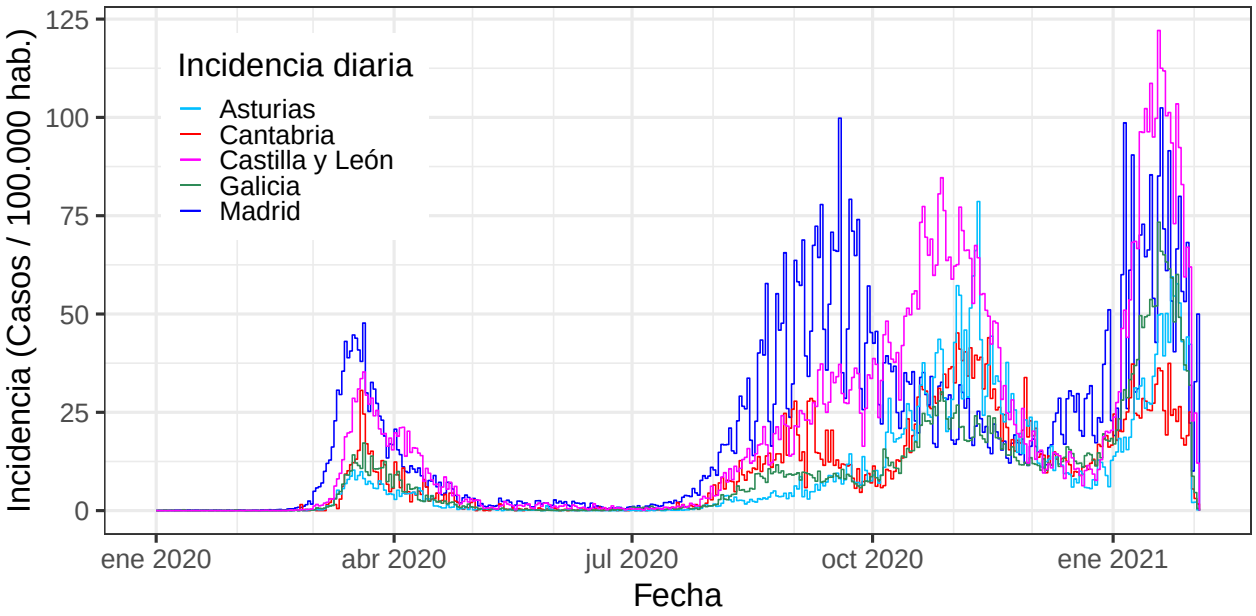
En el siguiente gráfico se compara la evolución de la enfermedad entre comunidades muy diferentes, la Comunidad Autónoma de Madrid, Cantabria, Asturias, Galicia y Castilla y León:



- Incidencia por cada 100.000 habitantes por CCAA

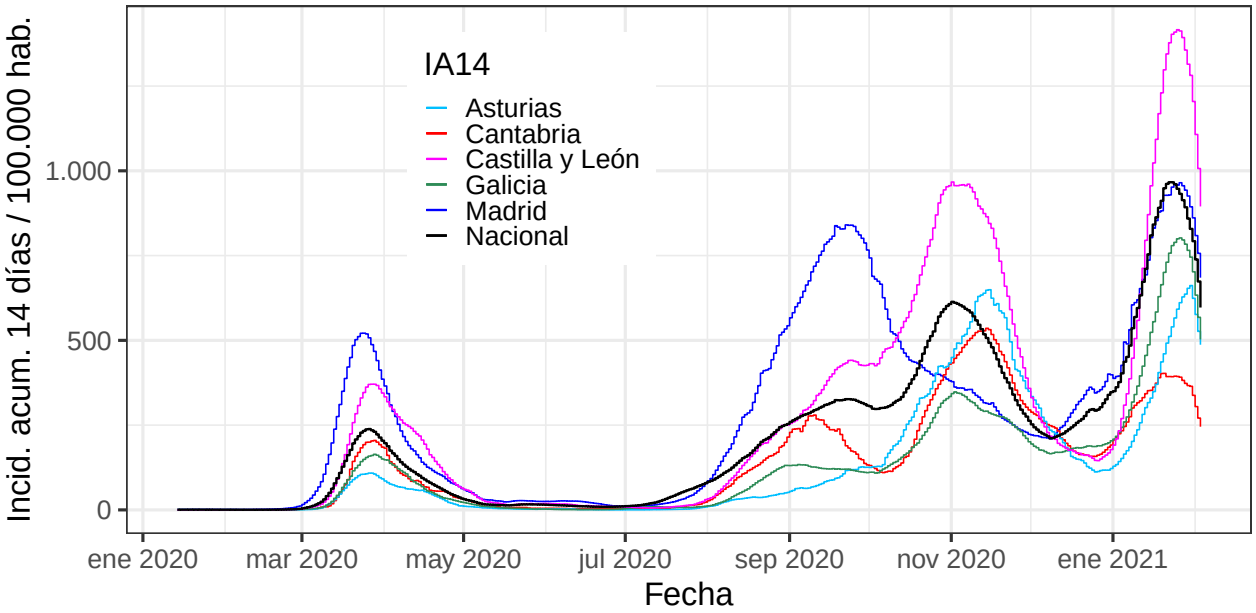
Como es lógico, los datos no son comparables en términos absolutos por la gran diferencia de población entre estas comunidades autónomas, por no entrar en la forma de vida en unas y otras, aunque sí que haya similitudes entre las de la cornisa cantábrica, y en cómo esto impacta en la dispersión de la enfermedad.

Para solventar este problema representamos ahora número de casos por cada 100.000 habitantes, con los datos de población en cada comunidad disponibles en el momento en el INE (https://www.ine.es/dynInfo/Infografia/Territoriales/capitulo.html#!tabla).



- Comparación de la IA14 por CCAA y nacional

Con el objetivo de completar la información comparativa entre estas comunidades se adjunta también la incidencia acumulada en 14 días para estas áreas geográficas, junto con la correspondiente al conjunto del territorio nacional:



9.- Evolución temporal de la tasa de variación de la incidencia acumulada en 14 días por cada 100.000 habitantes (IA14)

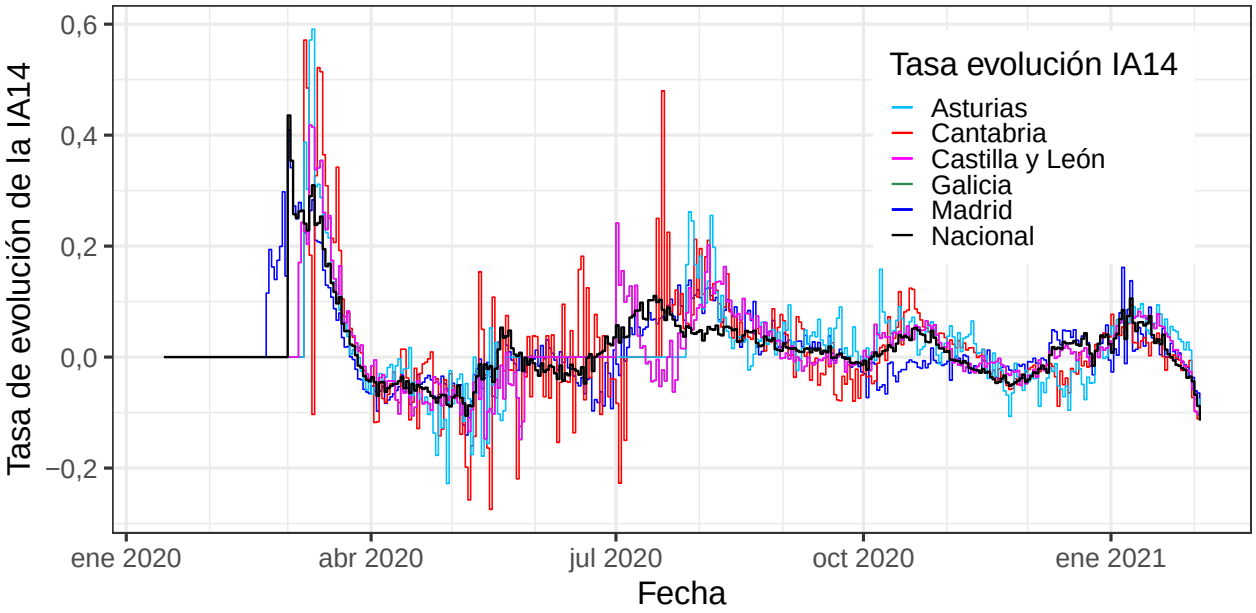
En el siguiente gráfico lo que se representa es cómo varía la incidencia acumulada en 14 días por cada 100.000 habitantes (IA14) expresando la tasa de variación de esta incidencia como:

tasaIA14(i) = (IA14(i)-IA14(i-1))/IA14(i-1)

Se impone la condición para el cálculo de la tasa que IA14(i-1) sea mayor que cero para evitar obtener tasas de crecimiento infinitas por la división con denominador cero y la indeterminación 0/0 en los casos de

secuencias de IA14 con valor 0 en los inicios de las series temporales.

Como para el caso de incidencias acumuladas de valores muy bajos, pequeños cambios de la dicha incidencia representan cambios de tasa de evolución muy elevadas al ser el divisor pequeño, modificaremos la condición indicada en el párrafo anterior, exigiendo que la incidencia acumulada en el día anterior tenga como mínimo un valor de 3 casos por cada 100.000 habitantes.



Hay que hacer notar que, aunque la gráfica resultante tenga una apariencia similar a la del Número reproductivo básico instantáneo - Rt (número promedio de casos secundarios que cada sujeto infectado puede llegar a infectar en una etapa de tiempo (t)), no se trata de este indicador, cuyo cálculo es totalmente diferente al presentado aquí. El número reproductivo básico instantáneo calculado por el Instituto de Salud Carlos III puede ser encontrado en el siguiente enlace: https://cnecovid.isciii.es/covid19/#ccaa. Nótese que el nivel de referencia del número reproductivo es 1 mientras que para la tasa de evolución de la IA14 es 0.

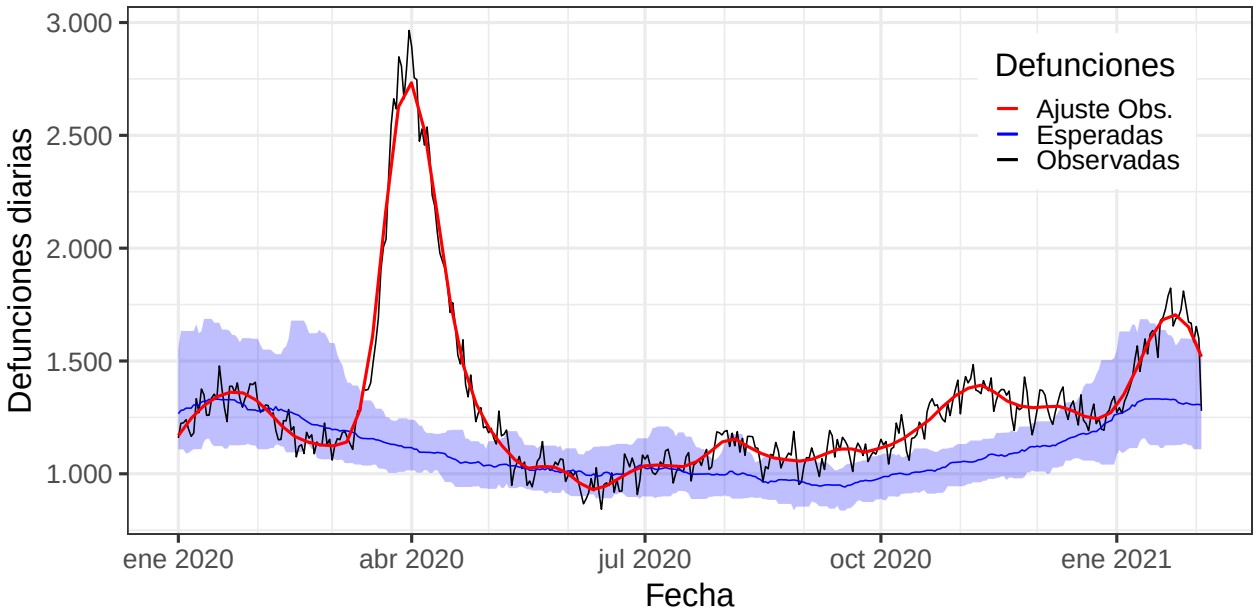
Por otro lado es lógica esta similitud entre las gráficas de la tasa de evolución de la IA14 y la del número reproductivo básico instantáneo, ya que números reproductivos altos se corresponden con evoluciones crecientes en la incidencia de la enfermedad mientras que números reproductivos por debajo de 1 marcan evoluciones decrecientes de la incidencia en el número de casos.

10.- Datos de mortalidad

Como siguiente paso del análisis estudiaremos la mortalidad registrada en el periodo de análisis. 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): **2021-02-04 17:59:31**

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



Se ha agregado una línea de ajuste estadístico de las defunciones observadas para facilitar la visualización de la evolución de las mismas suavizando los dientes de sierra propios de las observaciones por tratarse de datos diarios.

Como se puede ver, existe un periodo de mortalidad totalmente disparada a lo largo de los meses de marzo a mayo, con una punta con valores cercanos al triple de lo esperado, 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 a lo largo de varios meses antes de empezar a descender, que cubriría desde agosto hasta bastante avanzado diciembre.

Antes de avanzar en el estudio cuantitativo del exceso de mortalidad haremos un primer análisis cualitativo comparando las curvas de evolución de la incidencia acumulada (IA14) con la curva de defunciones.

Evidentemente lo que primero salta a la vista es la diferencia de proporción del pico de la primera ola con respecto a la segunda ola que ya hemos comentado anteriormente que en lo que respecta a la incidencia acumulada se debe a una subestimación por falta de pruebas diagnósticas, mientras que en lo que concierne a la mortalidad lo que parecen mostrar los datos que ha sucedido es que la primera ola tuvo un altísimo impacto en la población más sensible. No estábamos preparados, faltaba conocimiento y se reaccionó tarde, con lo que las consecuencias fueron desgraciadamente funestas. El confinamiento generalizado, aunque llegase tarde, era necesario y se mostró efectivo. Ahora bien, tanto a la vista de la incidencia acumulada como de la mortalidad no parece justificado de ningún punto que se prolongase por tanto tiempo como se hizo y el desescalamiento debería haber comenzado antes y haber sido con escalones más prolongados en el tiempo para de verdad haber mantenido un control efectivo de la propagación de la enfermedad.

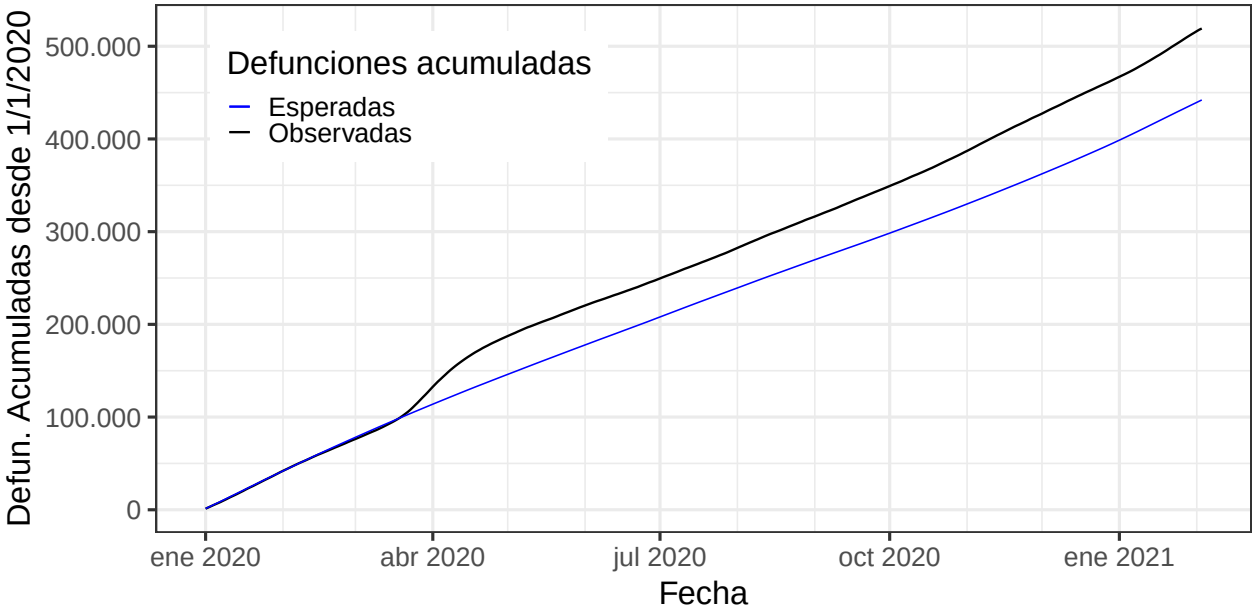
Como era de esperar, tan pronto se levantaron las medidas de restricción, la curva de incidencia empezó a crecer de forma constante. Curiosamente, contrario a lo que se esperaba, el inicio del nuevo curso escolar no tuvo un efecto negativo en la incidencia en el sentido de que ésta se disparase y de hecho en septiembre, primero disminuye el ritmo de crecimiento para luego incluso decrecer la incidencia en la segunda mitad de mes. Por otro lado, y también lógico y esperable, el crecimiento de la incidencia acumulada es acompañado por un crecimiento de la mortalidad que ya en verano supera el percentil 99 de las esperadas. La pregunta inevitable es por qué, con una incidencia creciente una mortalidad superior a la esperada y creciente, no se tomaron medidas antes por parte delas autoridades competentes.

11.- Exceso de mortalidad

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 de 2020 y lo que va de año 2021 con las que realmente se han registrado:



Como podíamos 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 distancian de las esperadas de forma muy visible a lo largo de los meses de marzo a mayo de 2020, y que dicho distanciamiento se vuelve a incrementar, ya a menor ritmo, a partir del mes de agosto, aunque se aprecia que vuelve a repuntar ligeramente en noviembre de 2020, con la punta de la segunda ola, y otro repunte más en enero con la tercera ola de la pandemia.

Más interesante resulta la comparación directa de las cifras acumuladas hasta la fecha. En este caso tenemos, con los datos disponibles, un total de **519.149** defunciones observadas y **442.059** defunciones esperadas, resultando un exceso de **77.090** defunciones. Expresando dicho exceso en términos porcentuales, nos encontramos con un **17,4 %** 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: **197.822** personas
- Esceso de defunciones: **41.487** personas
- En tanto por ciento: **21 %**

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 (fechas en formato aaaa-mm-dd):

- Antes de unificar periodos de exceso próximos:

Inicio	Fin
2020-03-10	2020-05-09
2020-07-20	2020-08-29
2020-09-01	2020-12-25
2021-01-04	2021-02-02

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

Inicio	Fin
2020-03-10	2020-05-09
2020-07-20	2020-12-25
2021-01-04	2021-02-02

Los excesos de defunciones en estos 3 periodos son:

Inicio	Fin	Exceso de defunciones
2020-03-10	2020-05-09	44.573
2020-07-20	2020-12-25	26.256
2021-01-04	2021-02-02	8.733

Siendo el total agregado de exceso de defunciones de **79.562** personas.

Expresándolo en términos porcentuales, el exceso de defunciones es un **18 %** 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.

Para eliminar esta dependencia temporal, veamos estos excesos en términos porcentuales con respecto a las esperadas, pero circunscritos exclusivamente al propio periodo de exceso de defunciones y dejando fuera el resto de la serie temporal:

Inicio	Fin	Exceso de defunciones	Porcentaje de exceso
2020-03-10	2020-05-09	44.573	66,8
2020-07-20	2020-12-25	26.256	16,0
2021-01-04	2021-02-02	8.733	22,1

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 provocada por la pandemia, los fallecimientos asociados a las mismas son por tanto atribuibles a la COVID-19 aunque el virus no haya sido la causa directa del fallecimiento correspondiente.

El índice de mortalidad de la COVID-19 en España en el periodo de estudio, medido como exceso de mortalidad atribuible directa o indirectamente a la COVID por cada mil habitantes, es de **1,68**.

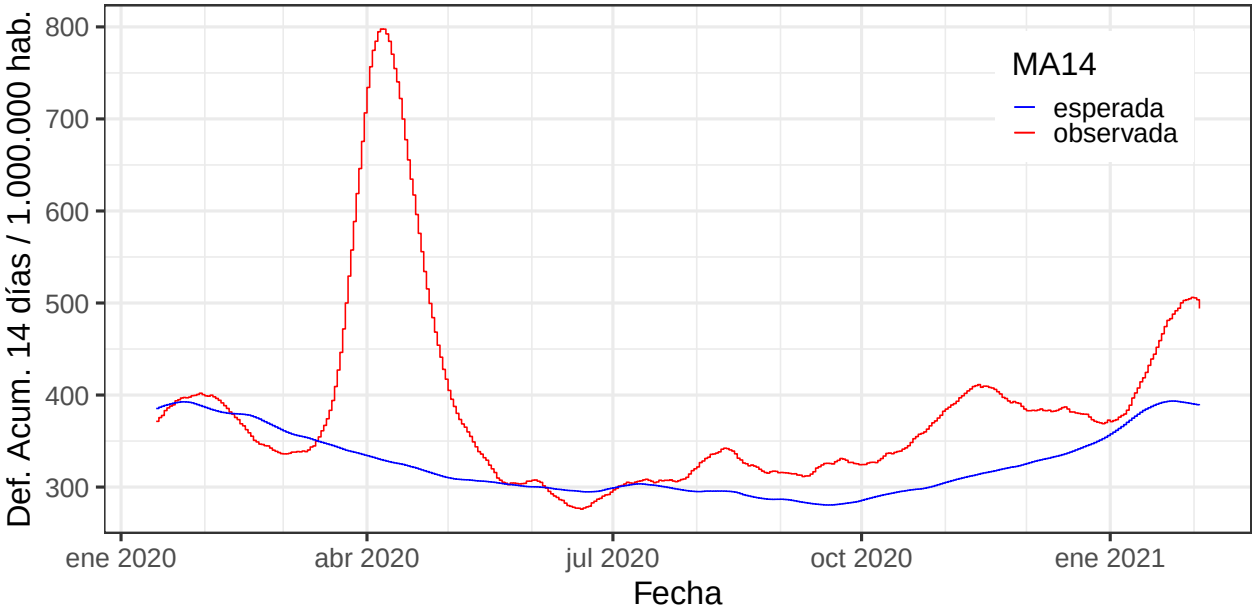
Antes de seguir avanzando no podemos dejar de llamar la atención sobre el hecho de que en la determinación de las cifras de exceso de defunciones se ha utilizado como nivel de referencia el número de defunciones esperadas. Es perfectamente argumentable que durante el periodo de estado de alarma este nivel de comparación debería ser inferior al estadísticamente obtenido con datos de años previos ya que el propio estado de alarma tuvo por necesidad incidencia en el número de fallecimientos por accidente laboral y por accidente de tráfico, sin duda disminuyéndolos. Consecuentemente debería rebajarse el patrón de referencia de defunciones esperadas

durante el estado de confinamiento y el exceso de defunciones por causa de la COVID-19 sería superior al mostrado más arriba. Aunque es posible realizar estimaciones de estas desviaciones con datos disponibles públicamente, dejamos esa posibilidad de perfeccionamiento del estudio para mejor oportunidad.

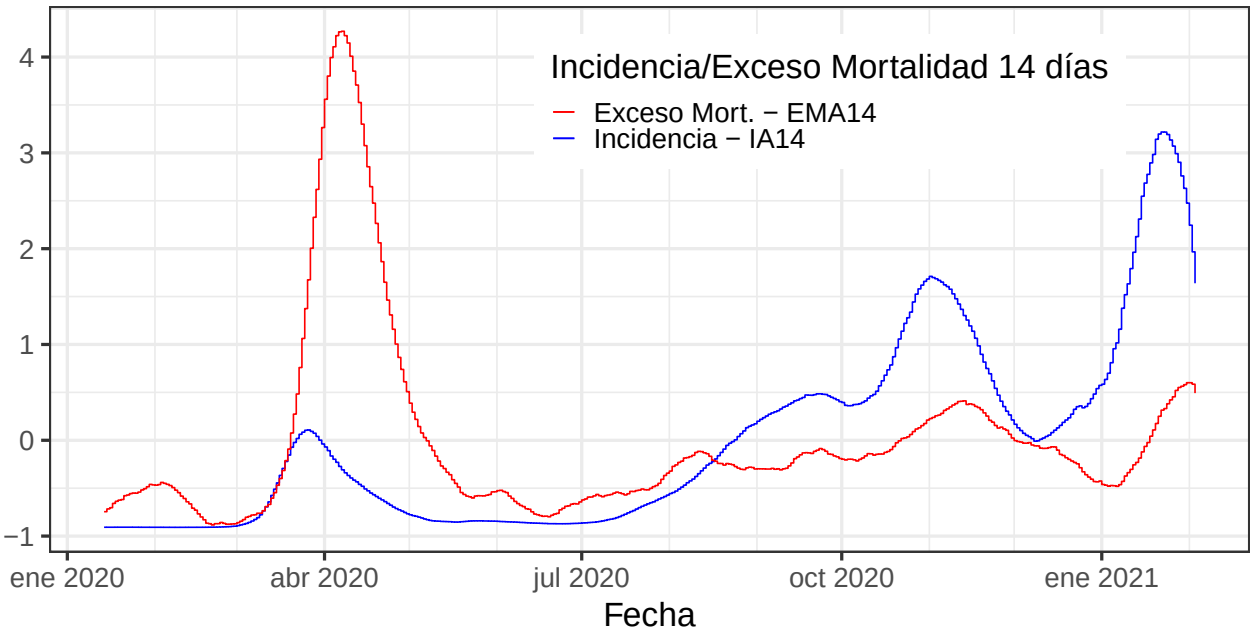
12.- Mortalidad acumulada en 14 días y comparación con la Incidencia acumulada en 14 días

Al hablar de la incidencia diaria de la pandemia veíamos que su valor se veía claramente influenciado por el menor número de pruebas realizadas durante los fines de semana y festivos, generando unos evidentes dientes de sierra a la hora de su representación gráfica. Aunque la razón de la variabilidad en las defunciones observadas no es la misma, recurriremos a una transformación similar, sumando defunciones en periodos móviles de 14 días y dividiendo la cantidad resultante por la población en España, obteniendo una variable que denominaremos mortalidad acumulada en 14 días por cada millón de habitantes (MA14).

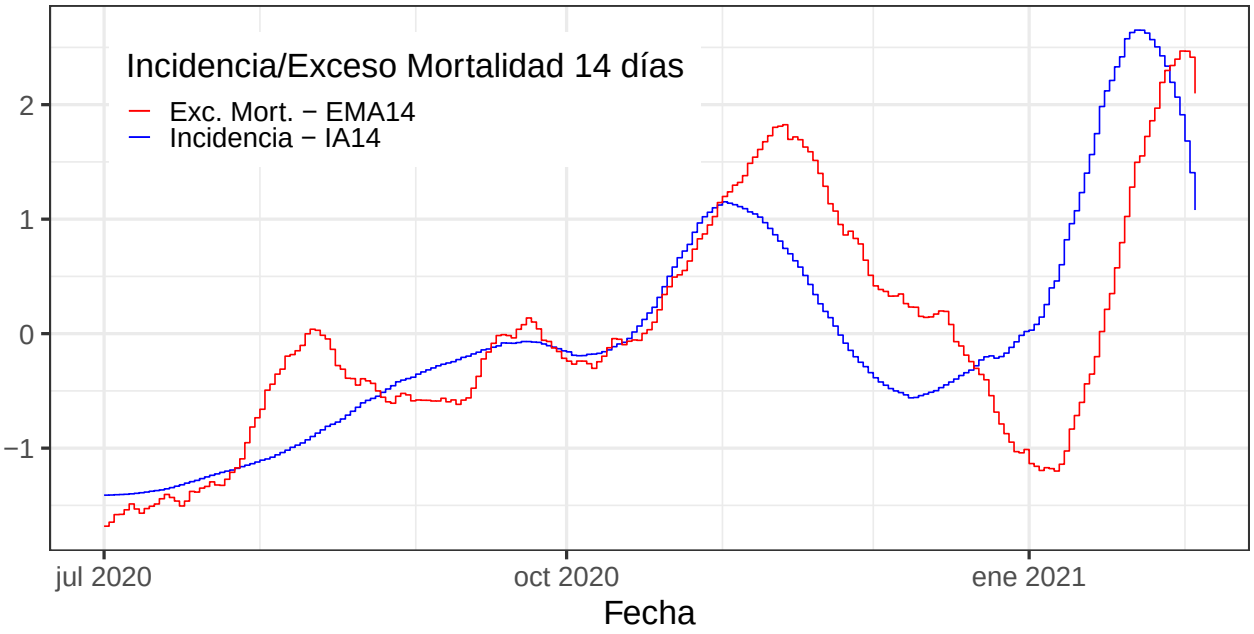
Representando gráficamente la MA14 de las defunciones observadas y la MA14 de las defunciones esperadas, observamos que mantienen la semejanza con los datos de origen, con la lógica distorsión de escala y desplazamiento temporal, y que se ha conseguido el deseado efecto de suavizar los dientes de sierra en lo que respecta a las defunciones observadas.



Ahora bien, lo que realmente nos interesa es comparar la diferencia entre estas variables de mortalidades acumuladas, observadas menos esperadas, que denominaremos exceso de mortalidad acumulada (EMA14), con la de la incidencia acumulada (IA14). Veamos qué pasa cuando escalamos ambas, EMA14 e IA14, y las representamos en el mismo gráfico:



Al comparar ambas gráficas vuelve a resaltar el que se invierta la proporción entre el pico de la primera ola y el de la segunda, resultado de la escasez de ensayos en la primera mitad de 2020, sin poder menoscabar las mejoras en los tratamientos de la enfermedad. En vez de intentar estimar la IA14 durante es primera mitad de año con conjeturas falibles, lo que haremos como siguiente paso es simplemente descartar el periodo desde el 1 de enero de 2020 hasta el 30 de junio de 2020 y analizar los datos correspondientes a fechas posteriores.



La correlación que ya se adivinaba con datos desde el 1 de enero de 2020 se hace mucho más evidente con la serie arrancando el 1 de julio de 2020, además de verificarse el esperado decalaje en el tiempo entre los picos de ambas variables, alcanzándose el máximo de exceso de mortalidad después de que la incidencia acumulada empiece ya a decrecer y el inicio del crecimiento en el exceso de mortalidad se produce con demora respecto al inicio del ascenso en la incidencia, por el tiempo de desarrollo de la enfermedad hasta llegar a consecuencias fatales.

Nota: No perdemos de vista en este apartado y en el siguiente que estamos comparando una variable que sí

que está directamente correlacionada con el virus responsable de la COVID-19, porque son identificaciones del virus mediante pruebas médicas, con defunciones de todo tipo y causa. Esto plantea una debilidad de base en los resultados obtenidos y en conclusiones que se infieran de éstos, pero la magnitud de la incidencia de la pandemia es tal que por los propios resultados obtenidos esta deficiencia de origen se puede obviar para cierto tipo de conclusiones de carácter más general. Evidentemente esta deficiencia es inasumible para un estudio epidemiológico serio pero no para el tipo de análisis que se pretende con este documento, que no deja de ser informal pese a una mantener una estructura ordenada y con ciertas semejanzas con un artículo científico.

13.- Análisis por grupos de edad y sexo de las defunciones ocurridas durante la primera ola de la pandemia

En este apartado haremos una somera disección de los fallecimientos producidos durante la primera ola de la pandemia por grupos de edad y por sexo. Tomaremos como periodo de referencia para el análisis de datos el correspondiente al primer periodo de exceso de defunciones tal y como se definió y determinó más arriba.

En resumen, los datos que se presentan ahora corresponden al periodo entre el **10 de marzo de 2020** y el **09 de mayo de 2020**.

En primera instancia nos limitaremos a analizar los datos para el agregado nacional, sin entrar en el detalle de lo acontecido en cada comunidad autónoma.

	edad < 65	edad 65-74	edad > 75	todos
hombres	7.645	9.681	38.489	55.815
mujeres	3.896	4.575	45.906	54.377
todos	11.773	14.438	85.042	111.253

Nótese que los desgloses por sexo siempre suman una cantidad inferior al total agregado correspondiente. Esto no es un error de programación sino una deficiencia, característica intrínseca por decirlo de una forma eufemística, de los datos de partida, en los que para una fecha dada las sumas por sexos no llegan al "todos" correspondiente, ni en el agregado total ni en los subtotales por franjas de edad.

Ante la disyuntiva de cómo continuar después de detectada esta discordancia, en vez de quedarnos con los números que representan las agregaciones "todos" (fila inferior y columna derecha), daremos por buenos los datos de detalle y calcularemos los nuevos agregados parciales y total.

La nueva tabla de distribución por sexos y edad queda entonces de la siguiente manera:

Expresándolo en número de defunciones:

	edad < 65	edad 65-74	edad > 75	Sum
hombres	7.645	9.681	38.489	55.815
mujeres	3.896	4.575	45.906	54.377
Sum	11.541	14.256	84.395	110.192

Representándolo como porcentajes:

	edad < 65	edad 65-74	edad > 75	Sum
hombres	6,94	8,79	34,9	50,7
mujeres	3,54	4,15	41,7	49,3
Sum	10,47	12,94	76,6	100,0

Sobre esta tabla hay que remarcar el hecho de que estamos obteniendo el total de fallecimientos, con

independencia de la causa, no sólo por la COVID-19, con lo cual por sí sola no puede arrojar mucha luz ya que no podemos separar los casos "Covid" de los "no-Covid". Lo que haremos para solventar esta carencia es comparar el reparto porcentual durante la ola con un periodo de referencia. En este caso no nos complicaremos con medias de largos periodos y lo compararemos con la distribución correspondiente al año 2019.

	edad < 65	edad 65-74	edad > 75	todos
hombres	33.536	33.964	119.760	187.260
mujeres	16.982	16.476	144.370	177.828
todos	52.825	51.759	268.454	373.038

Como ya apuntábamos antes, el problema de falta de coherencia entre los grupos "todos" y sus desgloses es inherente a los propios datos y volveremos a depurarlos de la misma manera en esta ocasión, conservando los desgloses y calculando nuevos subtotales y total agregado:

En número de fallecimientos:

	edad < 65	edad 65-74	edad > 75	Sum
hombres	33.536	33.964	119.760	187.260
mujeres	16.982	16.476	144.370	177.828
Sum	50.518	50.440	264.130	365.088

Como porcentajes:

	edad < 65	edad 65-74	edad > 75	Sum
hombres	9,19	9,30	32,8	51,3
mujeres	4,65	4,51	39,5	48,7
Sum	13,84	13,82	72,3	100,0

Coloquemos las tablas de porcentajes juntas una con la otra para que sea fácil compararlas:

- Primera ola de la pandemia:

	edad < 65	edad 65-74	edad > 75	Sum
hombres	6,94	8,79	34,9	50,7
mujeres	3,54	4,15	41,7	49,3
Sum	10,47	12,94	76,6	100,0

- Año 2019:

	edad < 65	edad 65-74	edad > 75	Sum
hombres	9,19	9,30	32,8	51,3
mujeres	4,65	4,51	39,5	48,7
Sum	13,84	13,82	72,3	100,0

Como se puede ver, con estos datos no es posible afirmar que la COVID-19 en términos de mortalidad haya afectado significativamente más a la población masculina que a la femenina y sólo se puede apreciar un desplazamiento de los fallecimientos hacia las franjas de edad mayores, algo que por otro lado es esperable puesto que la enfermedad afecta más severamente a las personas con patologías previas y éstas se encuentran

por lógica mayoritariamente entre los grupos de edad más avanzada.

. . . . . .

Apéndice - Referencias

(1) RStudio Team (2021). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA. URL: http://www.rstudio.com/

(2) 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/

(3) 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/

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

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

(6) H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

(7) 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