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

Juan Matorras Díaz-Caneja

13/01/2021

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 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 corrección 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_diagnostico_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 y Galicia. La razón de la selección de estas comunidades y no otras responde a consideraciones personales y no obedece a ningún criterio técnico.

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): **2021-01-13 10:48:22**

El análisis se ha llevado a cabo utilizando el software libre para análisis estadístico **R**, versión 4.0.3. (1)

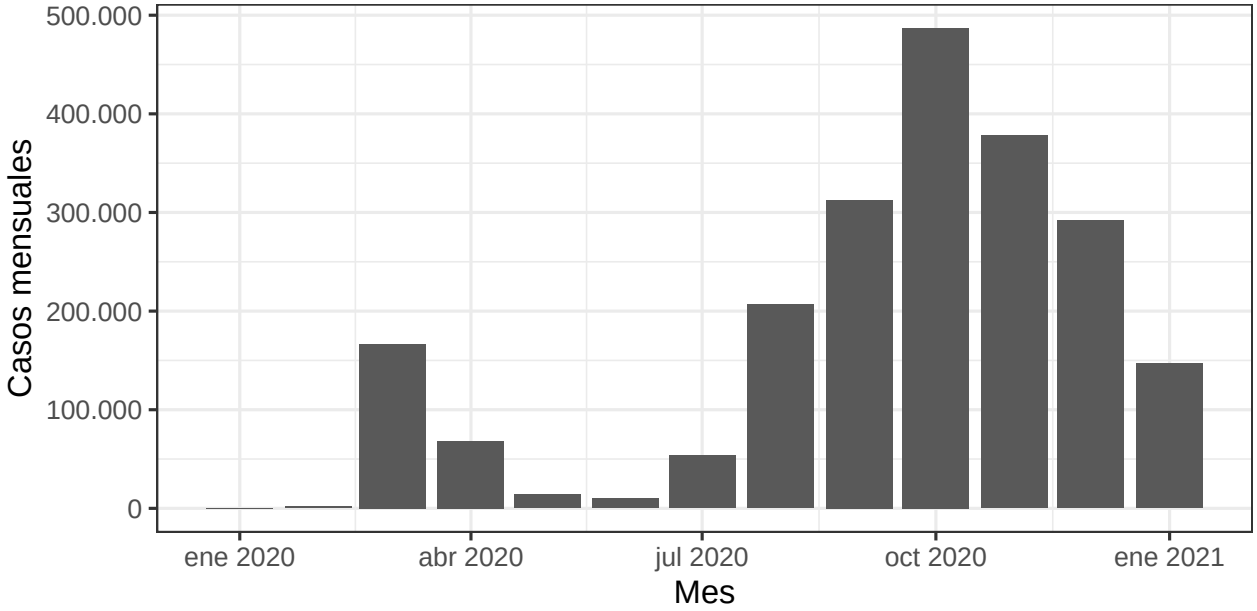
Se ha hecho uso también de los paquetes complementarios:

- **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) y (5)

- **data.table** con el objeto de manipular las tablas más eficientemente. (6)

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:

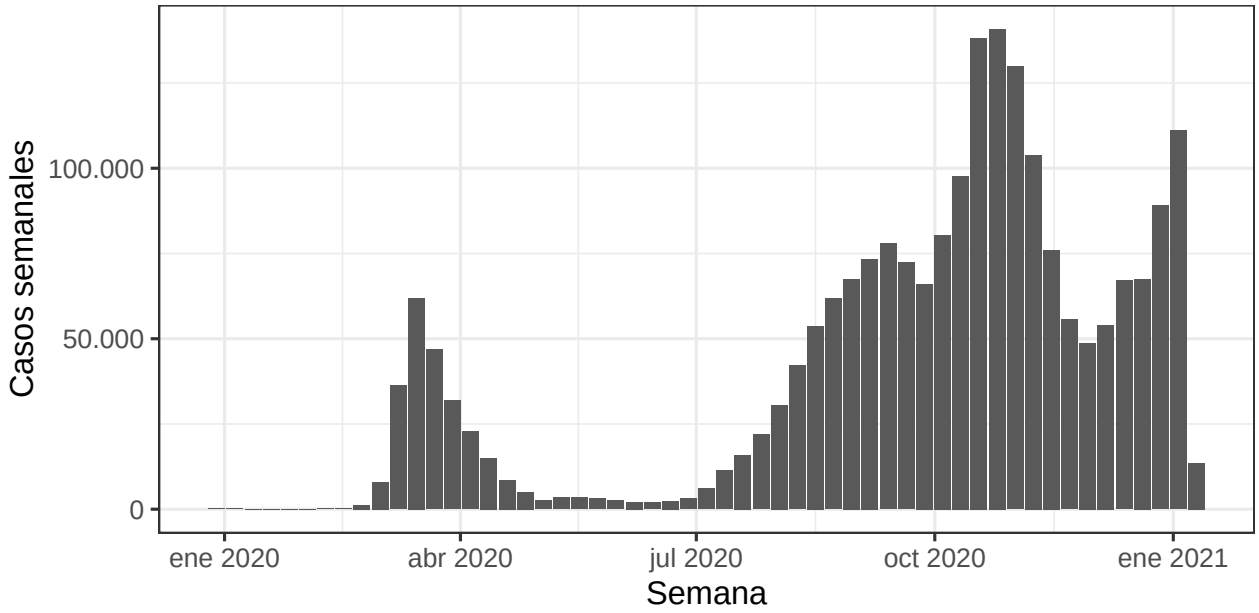
Mes	Casos
ene 2020	461
feb 2020	1.767
mar 2020	166.944
abr 2020	67.988
may 2020	14.124
jun 2020	9.896
jul 2020	53.471
ago 2020	206.678
sep 2020	312.704
oct 2020	486.873
nov 2020	377.977
dic 2020	291.800
ene 2021	146.736

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 **2.137.419**.

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 **4,516 %** hasta la fecha.

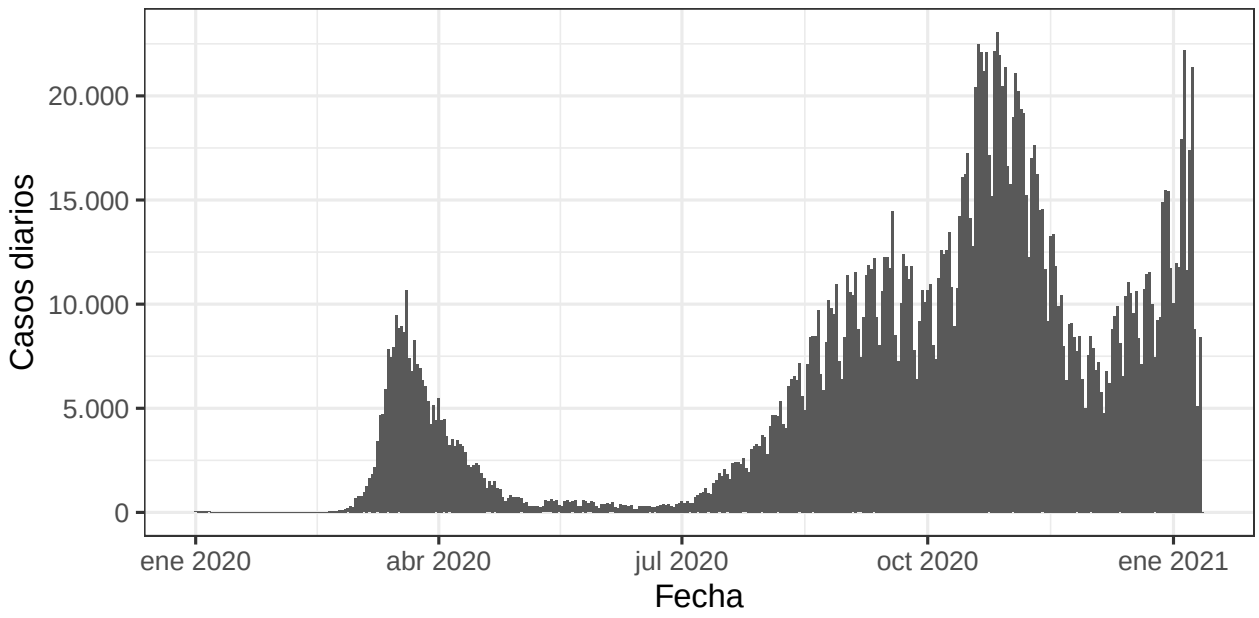
Incidencia semanal

- Evolución de número de casos identificados por semanas:

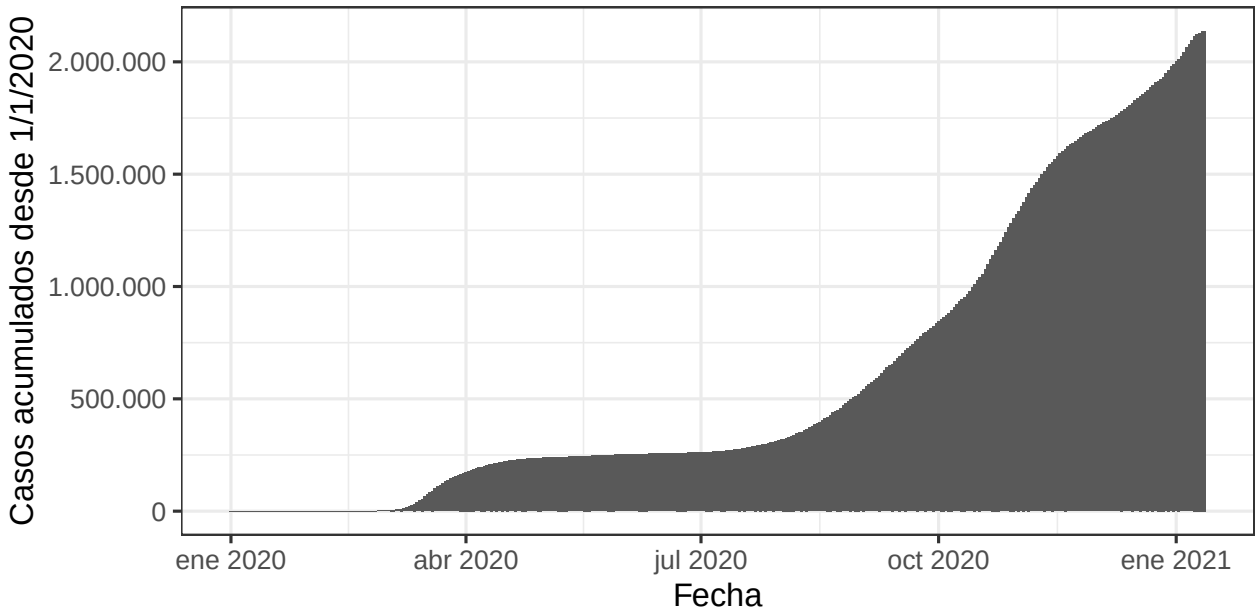


Incidencia diaria

- Curva epidémica de los casos notificados por días:

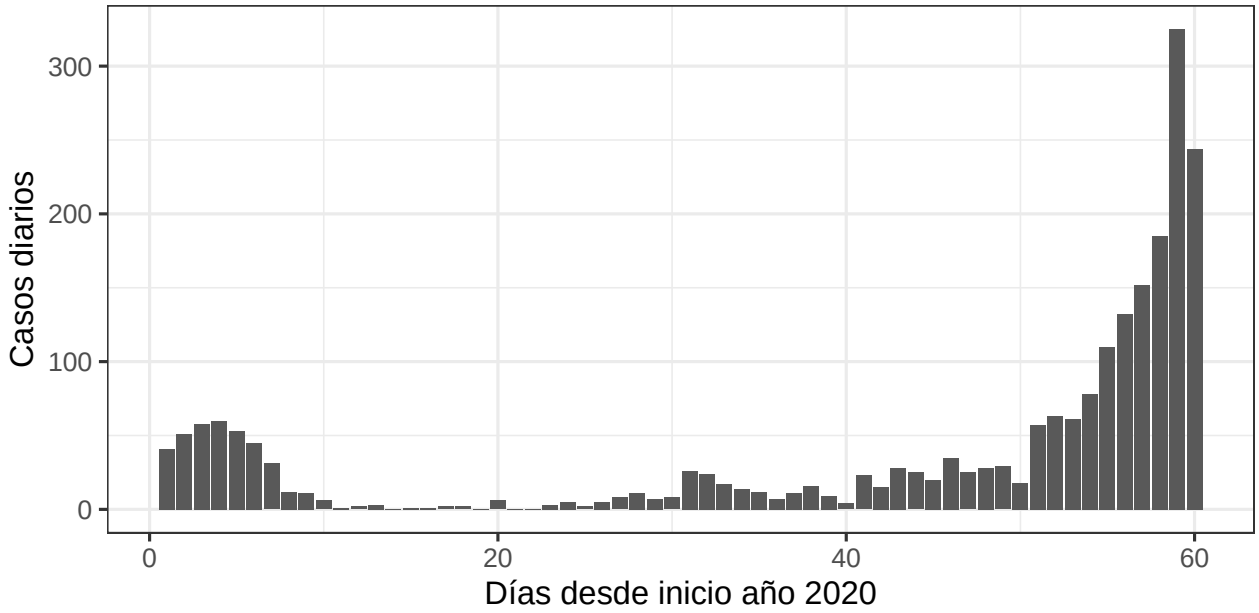


- Gráfico de casos acumulados a origen por días:



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:



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 declarados en los primeros días de 2020 que antes no existían y que más que una revisión de datos antiguos puede ser 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 comprativa donde se puede observar cómo han evolucionado el número de casos declarados para enero de 2020 en las distintas CCAA entre los dias 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 y de hecho se observan ya algunas rectificaciones en los datos correspondiente a Madrid. Estos datos aparentemente erróneos no se han eliminado ni sufrido ningún tipo de tratamiento especial y se han manipulado como si fuesen correctos.

Los casos reportados totales a lo largo de esos dos meses son **2.228**, 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: 821
- Periodo 51-60: 1.407

En 10 días se detectan **1,7** veces los casos que se habían producido en los 50 días anteriores.

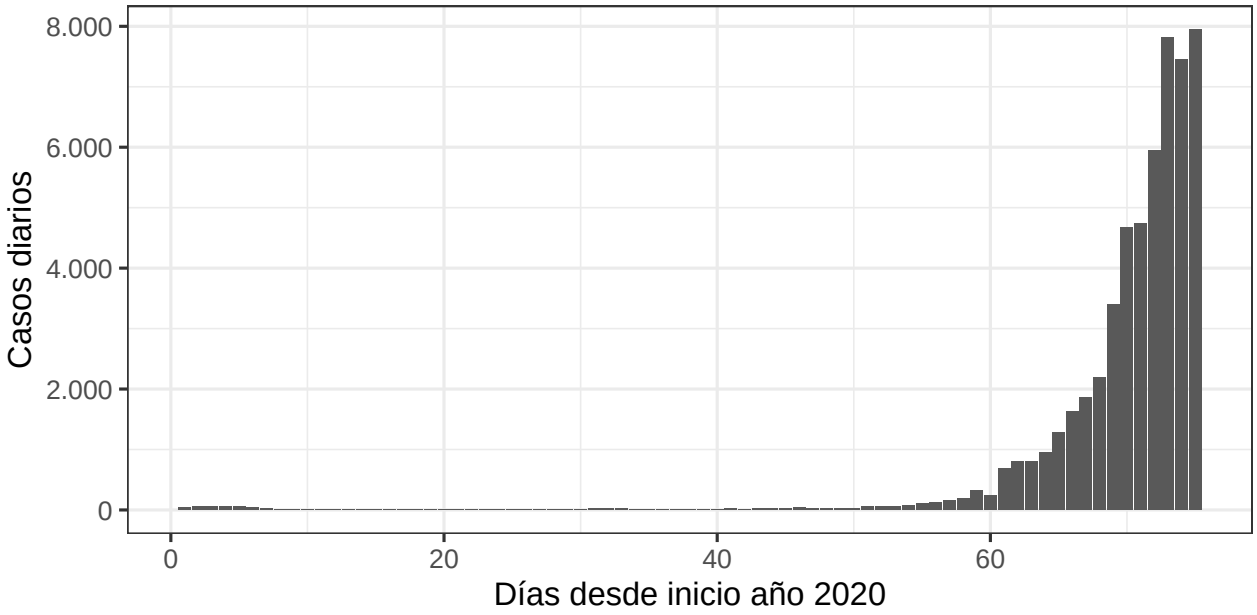
Subsiguiente evolución durante la primera quincena de marzo

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 **10.206** casos a añadir al total anterior, siendo éstos **4,6** 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 **41.986**, 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

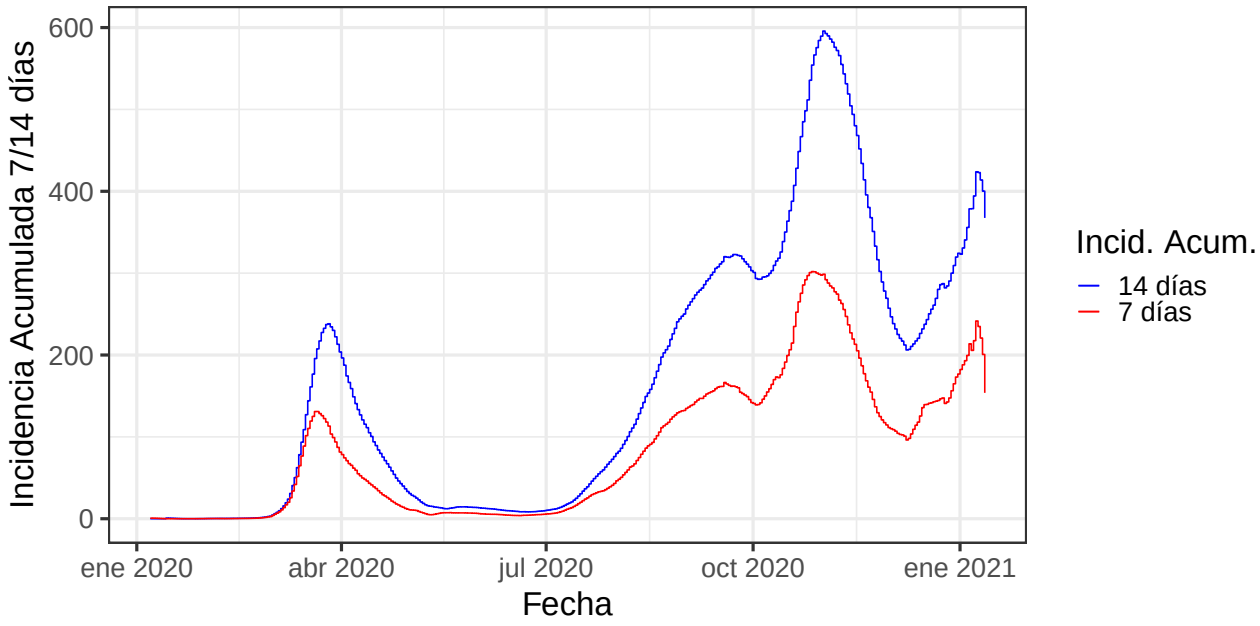
Pasemos ahora a calcular la incidencia acumulada por cada 100.000 habitantes en los 14 días previos a la declaración del estado de alarma que tuvo efecto

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.

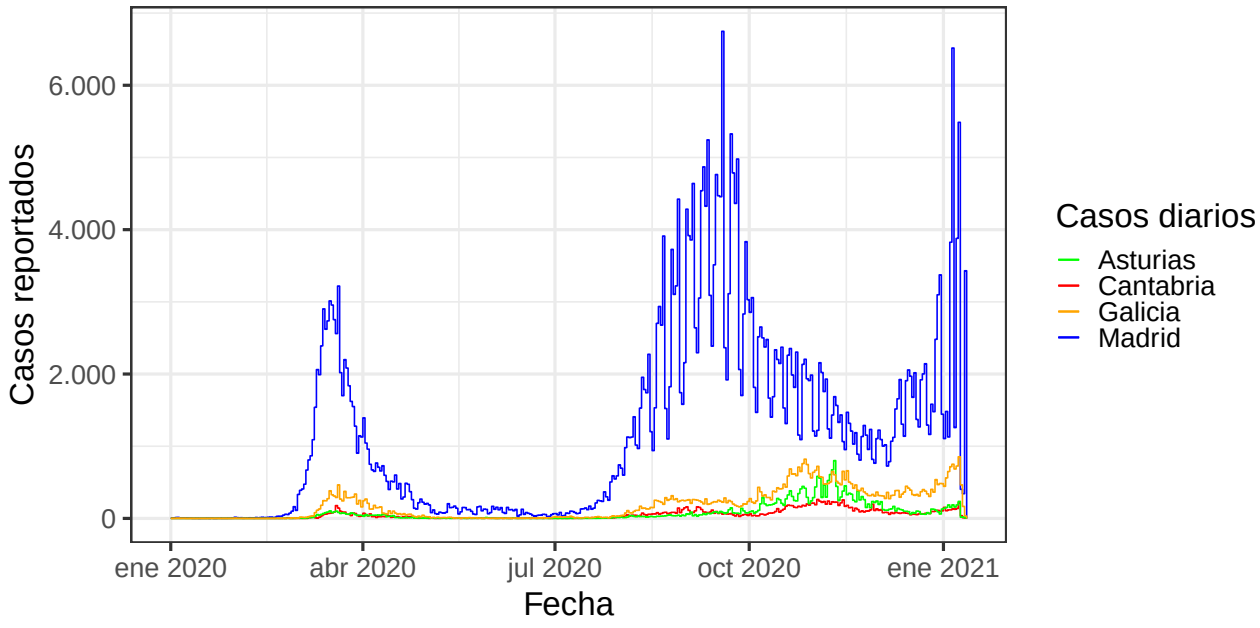
Evolución de la incidencia acumulada a lo largo de todo el periodo de análisis

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



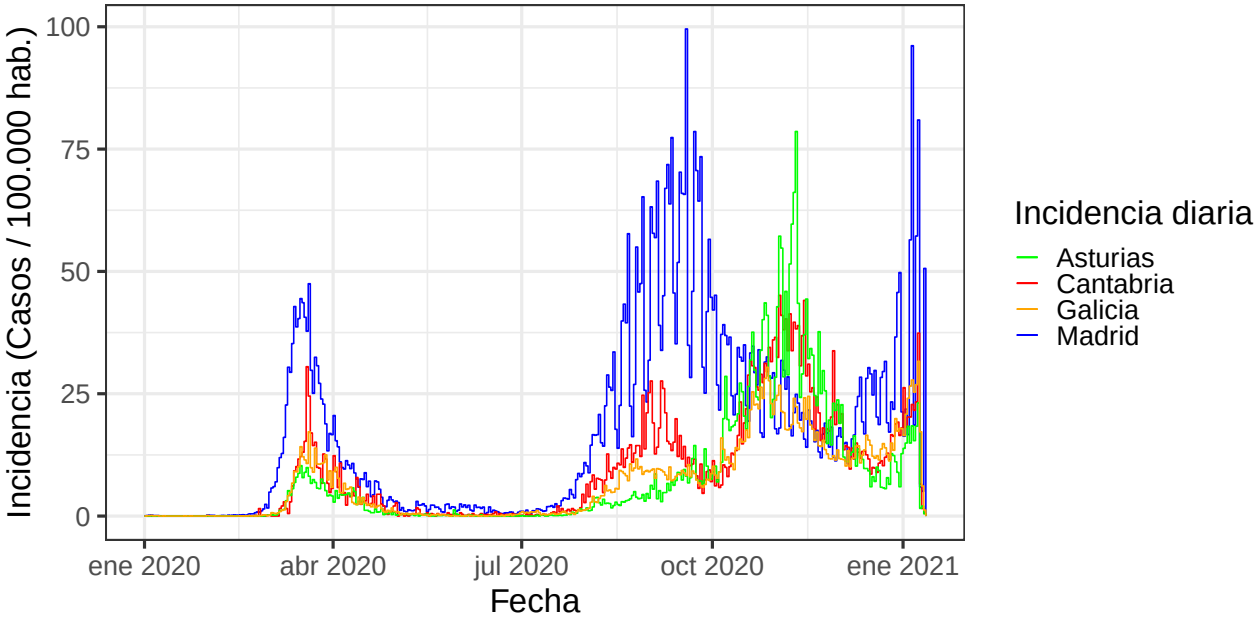
Comparación de la evolución del número de casos entre Madrid y otras CCAA

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

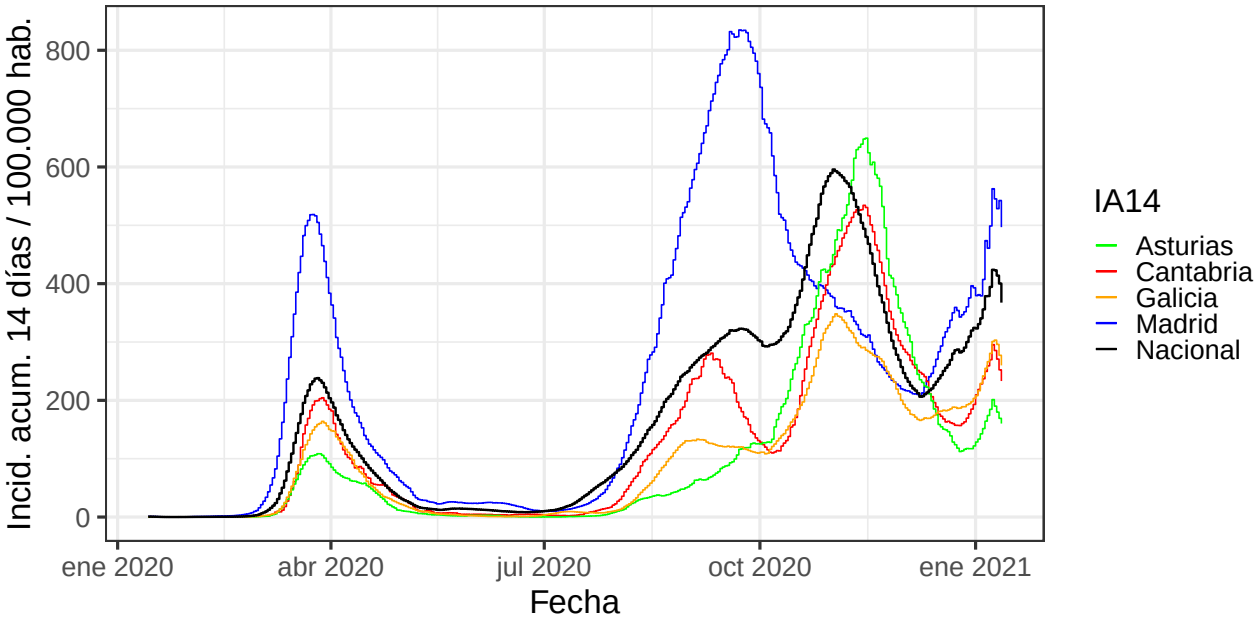


Como es lógico, los datos no son comparables en términos absolutos por la gran diferencia de población entre estas comunidades autónomas, por no entrar en la forma de vida en unas y otras, aunque sí que haya similitudes entre las de la cornisa cantábrica, y en cómo esto impacta en la dispersión de la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



Por 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:



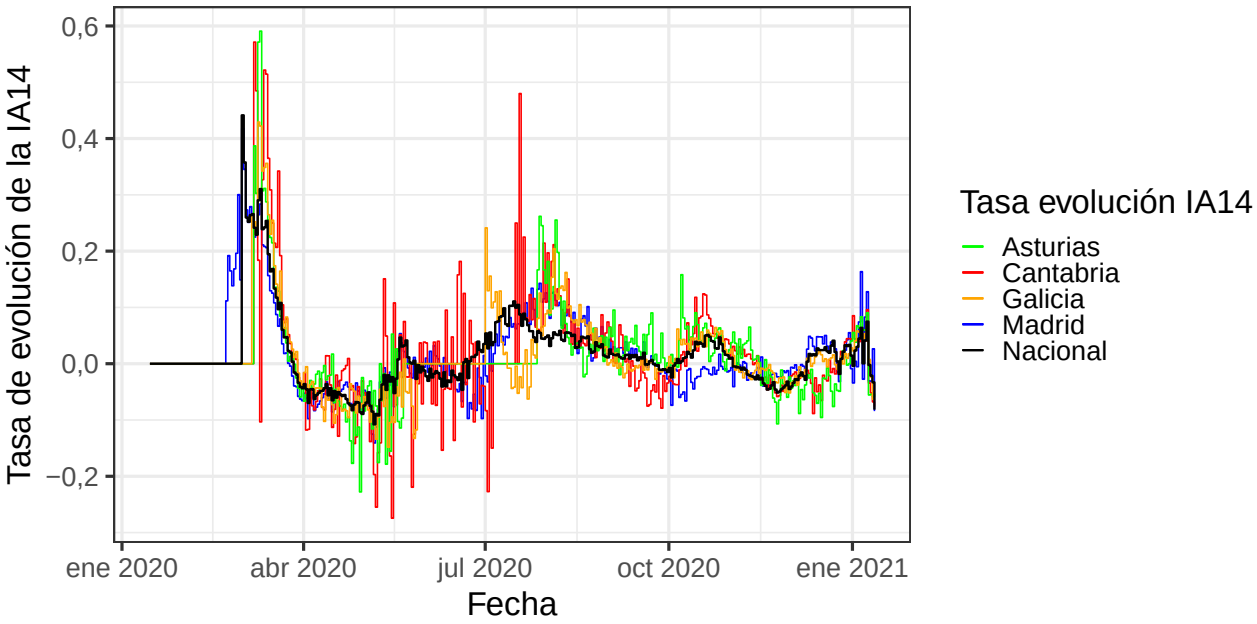
Evolución temporal de la tasa de variación de la incidencia acumulada en 14 días por cada 100.000 habitantes (IA14)

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 bajor, 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 ínstantáneo 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 referecia del número reproductivo es 1 mientras que para la tasa de evolución de la IA14 es 0.

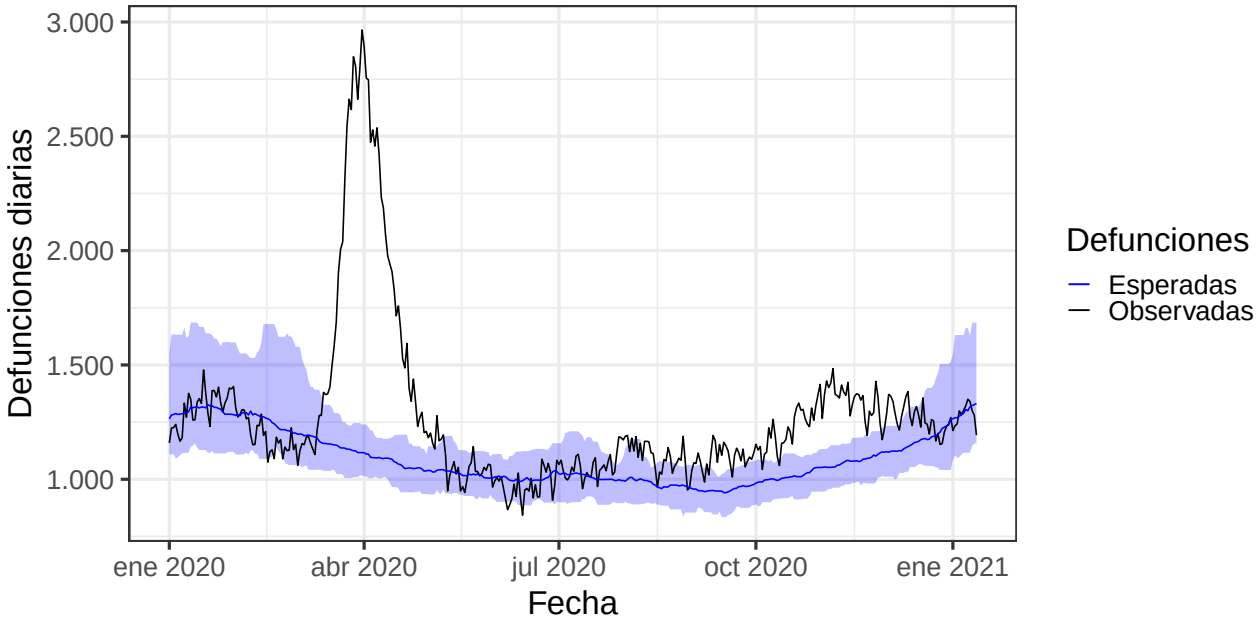
Por otro lado es lógica esta similitud entre las gráficas de la tasa de evolución de la IA14 y la del número reproductivo básico instántaneo, ya que números reproductivos altos se corresponden con evoluciones crecientes en la incidencia de la enfermedad mientras que números reproductivos por debajo de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

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): **2021-01-13 10:48:40**

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

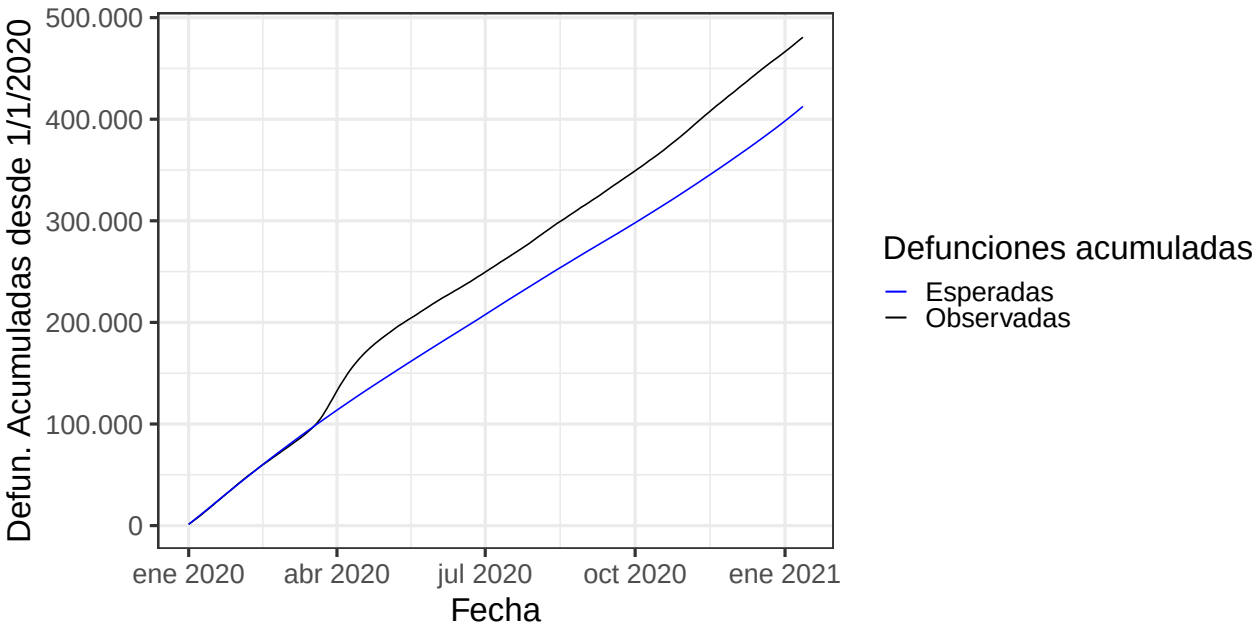


Como se puede ver, existe un periodo de mortalidad totalmente disparada a lo largo de los meses de marzo a mayo, con una punta con valores cercanos al triple de lo esperado, mientras que luego se aprecia otro periodo de desviación al alza, no tan acusado pero más prolongado en el tiempo y con una tendencia creciente a lo largo de varios meses antes de empezar a descender, que cubríría desde agosto hasta bastante avanzado diciembre.

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 distanciamento 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.

Más interesentante resulta la comparación directa de las cifras acumuladas hasta la fecha. En este caso tenemos, con los datos disponibles, una total de **480.640** defunciones observadas y **412.598** defunciones esperadas, resultando un exceso de **68.042** defunciones. Expresando dicho exceso en términos porcentuales, nos encontramos con un **16,5 %** 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: **197.518** personas
- Esceso de defunciones: **41.791** 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 (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	2020-12-23

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

Inicio	Fin
2020-03-10	2020-05-09
2020-07-20	2020-12-23

Los excesos de defunciones en estos 2 periodos son:

Inicio	Fin	Exceso de defunciones
2020-03-10	2020-05-09	44.587
2020-07-20	2020-12-23	26.193

Siendo el total agregado de exceso de defunciones de **70.780** personas.

Expresándolo en términos porcentuales, el exceso de defunciones es un **17,2 %** 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	44.587	66,9
2020-07-20	2020-12-23	26.193	16,2

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 2020, medido como exceso de mortalidad atribuible directa o indirectamente a la COVID por cada mil habitantes, es de **1,5**.

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 estadisticamente 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. Consecuentmente 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.

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 disyuntavia 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	35.238	35.821	126.753	197.812
mujeres	17.833	17.334	152.952	188.119
todos	55.518	54.551	284.270	394.339

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 calcuando nuevos subtotales y total agregado:

En número de fallecimientos:

	edad < 65	edad 65-74	edad > 75	Sum
hombres	35.238	35.821	126.753	197.812

	edad < 65	edad 65-74	edad > 75	Sum
mujeres	17.833	17.334	152.952	188.119
Sum	53.071	53.155	279.705	385.931

Como porcentajes:

	edad < 65	edad 65-74	edad > 75	Sum
hombres	9,13	9,28	32,8	51,3
mujeres	4,62	4,49	39,6	48,7
Sum	13,75	13,77	72,5	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	9,13	9,28	32,8	51,3
mujeres	4,62	4,49	39,6	48,7
Sum	13,75	13,77	72,5	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 másculina 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 entre los grupos de edad más avanzada.

. . . . . .

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

(6) 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