Datos Covid-19: Caso Chileno

Emiliano A. Moreno Miranda
Segundo Semestre 2020

Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

Introducción

En el mes de Marzo, año 2020, se registro en Chile el primer caso oficialmente confirmado de Covid-19, una nueva enfermedad viral, relacionada principalmente con problemas respitatorios. El brote inicial fue en China, pocos meses antes de llegar a Chile, cuando ya en otros paises como España e Italia, considerados de mayor desarrollo, la expansión del virus era causa de gran mortalidad. Sin tener claridad aún de la envergadura que alcanzaria la epidemia ni los tratamientos adecuados **(fuente)**, ya, en ese entonces, se hacia evidente que la propagación de la enfermedad era rápida, sugiriendo una alta capacidad de contagio y el grupo etario de mayor riesgo era la tercera edad **(fuente)**. Esta última hipotesis influenciada principalmente por los primeros datos de los países mencionados. Así los gobiernos del mundo se aprestaron a imponer medidas de confinamiento entre otras coerciones a la libertad, como principal medida contra la expasión del contagio.

Algunas de las medidas tomadas no tienen precedentes en la historia comtemporanea. También la cantidad de información *almacenada* desde el inicio de la situación hasta la fecha de elaboración del presente informe ha sido sustantiva dadas las capacidades tecnológicas actuales. De lo anterior surge la necesidad de, primero, convertir la información *almacenada* en conocimiento útil para la toma de deciciones y futuros eventos, así como evaluar el *impacto* de aquellas medidas en el manejo de la enfermedad, y también en otros aspectos de la vida humana.

En el caso de Chile el gobierno ha dispuesto una base de datos a través de la cuenta **GitHub del Ministerio de Ciencia**, en ella se encuentran distintos *productos de datos*, cada uno de los cuales consiste en un archivo en formato `.csv` con datos que documentan la realidad Chilena en el contexto del Covid-19. Con estos datos pretendemos poner a prueba las principales ideas de la opinión pública acerca de la enfermedad, es decir si los datos suministrados por la fuente mencionada permiten hacer dichas inferencias. También buscamos y proponemos nuevas alternativas para caracterizar y evaluar el *comportamiento* de las distintas regiones del país

Para nuestros propósitos, primero describimos verbalmente cada uno de los *productos de datos* que consideraremos a lo largo del texto (el repositorio mencionado es vasto y aquí se deben concentrar los esfuerzos). Luego, para proponer una evaluación del *comportamiento*, se sigue el procedimiento de una *transformación*, usando nuevas variables creadas a partir de combinaciones de las variables existentes. Finalmente, aplicamos la metodología de *ANOVA* junto a una validación por *prueba de hipótesis* para realizar inferencia sobre la hipótesis de que la tercera edad es la más afectada.

Descripción de los datos

Durante el informe haremos mención a los distintos productos de datos, suministrados por el Ministerio de Ciencia de Chile, esta primera sección la dedicamos a describir brevemente en que consiste el contenido de cada uno de estos productos. Sientase libre de volver a leer las descripciones cuanto considere necesario. Tenga en cuenta que las bases de datos estan actualizadas hasta el *16-11-2020*, fecha en que se comenzó con la elaboración del informe.

Producto de datos 03 - Casos totales por región:

Este producto da cuenta de los casos totales diarios confirmados por laboratorio en cada una de las regiones de Chile, según residencia, y concatena la información reportada por el Ministerio de Salud del país.

Producto de datos 09 - Pacientes COVID-19 en UCI por grupo de edad:

Set de 2 archivos que dan cuenta del número de pacientes en UCI por grupos etarios (<=39; 40-49; 50-59; 60-69; y >=70) y que son casos confirmados por COVID-19, reportados diariamente por el Ministerio de Salud, desde el 01-04-2020.

Producto de datos 10 - Fallecidos con COVID-19 por grupo de edad:

Set de 2 archivos que dan cuenta del número de personas fallecidas con COVID-19, agrupadas por rangos etarios (<=39; 40-49; 50-59; 60-69; 70-79; 80-89; y >=90) reportados diariamente por el Ministerio de Salud, desde el 09-04-2020.

Producto de datos 16 - Casos por genero y grupo de edad:

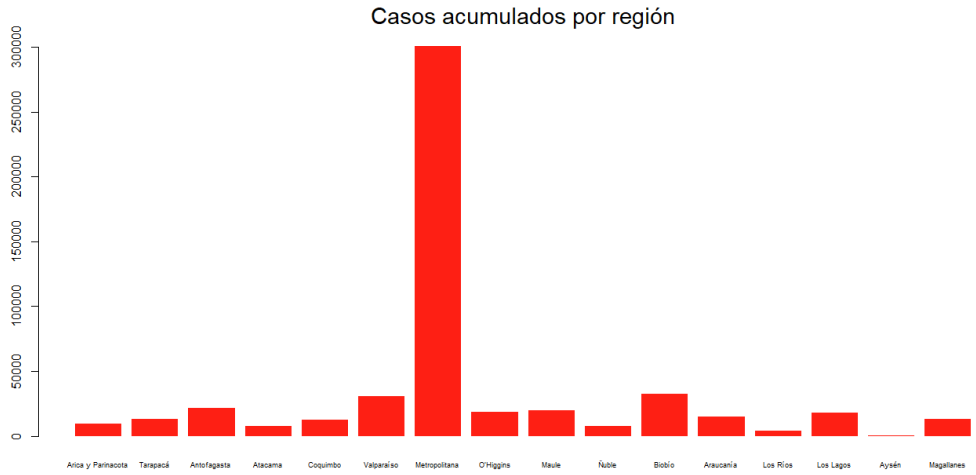
Archivo que da cuenta del número acumulado de casos confirmados distribuidos por género y grupo de edad, para cada fecha reportada. Este concatena la historia de los informes epidemiológicos publicados por el Ministerio de Salud del país.

Una mirada al comportamiento de las regiones

En esta sección revisaremos el comportamiento de la distintas regiones de Chile en la emergencia sanitaria producida por el Covid-19. Para ser preciso en el sentido que se le da a la palabra *comportamiento* en el texto, puede Ud. pensar en preguntarse si hubiese querido estar durante la emergencia en una región u otra. Una región con *mal* comportamiento, no es un lugar deseable, por el contrario, una región con *buen* comportamiento pareciera ser un lugar apropiado.

Esta primera mirada no pretende ser un análisis profundo, si no un punto de partida. Para tal efecto consideramos el *producto de datos* **03**.

Una manera natural de evaluar como se han *comportado* las regiones en relación al control de la enfermedad, podría ser la cantidad de casos totales. Para cada región tenemos la cantidad de casos acumulados a la fecha de corte del estudio, por ello usamos un gráfico de barra, el cual visualizamos a continuación.



Obs: Regiones ordenadas de Izq. a Derecha geográficamente en sentido N-S

Este gráfico es útil para visualizar los datos de manera general, pero es difícil apreciar el detalle de las regiones con más contagios. La siguiente tabla nos ayuda a comprender esto, ordenando las regiones por cantidad de contagios en forma descendente.

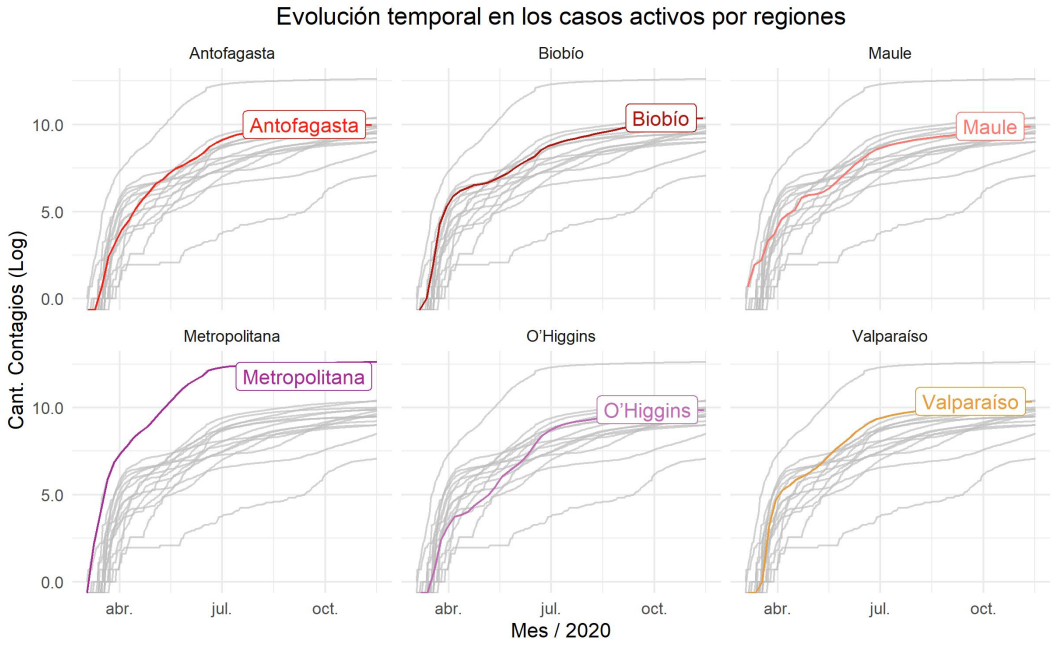
	Región	Casos Acumulados
7	Metropolitana	301207
11	Biobío	33110
6	Valparaíso	31467
3	Antofagasta	21879
9	Maule	20096
8	O'Higgins	19321

En síntesis, la región Metropolitana es la que mas casos presenta, entonces según este criterio es la región con peor desempeño en el manejo de la pandemia. Sin embargo sabemos que es la región mas poblada según los

datos del *Censo 2017* proporcionados por el **INE**, y también la primera en registrar casos del virus. Podemos considerar otras métricas para corregir ese hecho.

Una alternativa para evaluar el comportamiento de las regiones

Para comenzar el estudio de una alternativa para evaluar el *comportamiento* de las regiones en la pandemia de Covid-19, consideremos la evolución de los casos a través del intervalo de tiempo que considera nuestro informe, para visualizar esto utilizamos el siguiente gráfico de linea, poniendo énfasis en las regiones con mas casos registrados, tal como se determino en la *sección* anterior. Es importante notar que en el eje vertical se ha considerado una escala logarítmica para los datos, con el fin de facilitar la comparación.



Al respecto del la gráfica anterior, lo primero que nos llama la atención es que la pendiente de la *curva de contagios* (serie de tiempo) es diferente en cada región, por ejemplo, la curva de la región de Antofagasta es aproximadamente *logaritmica* mientras que la región de O'higgins parece tener una tendencia bastante más *lineal*. Esta diferencia también la podemos observar en las regiones que no hemos destacado.

Pensemos en la interpretación de esta *curva*: Mientras mayor es la pendiente, mayor es la cantidad de contagios nuevos por cada día. Parece razonable pensar que las regiones cuyos contagios evolucionan aceleradamente (pendiente grande), tienen un peor comportamiento, independientemente de la cantidad absoluta de los contagios, que una región donde el crecimiento es más estable (lineal).

Considerando lo anterior, vamos a proponer la medida Λ, dígase *lambda*, para evaluar el comportamiento de las regiones, donde Λ se define para cada región como la siguiente transformación de los datos:

$$\Lambda_i = ln\left(\frac{\kappa * \delta_i}{ln(P_i)} * Y_i\right)$$

Ahora dedicamos unas lineas a entender la *transformación* Λ. Primero, lo más sencillo, Y_i son los datos de la cantidad *total* de contagios en cada región indexada por la constante i. Es importante considerar esta información, pues uno de los objetivos de la transformación es enriquecer el criterio para el *comportamiento* establecido en la *sección* anterior.

Por segundo, tenemos a δ_i que se define como $\delta_i = \hat{\beta}_1$ donde $\hat{\beta}_1$ es el estimador de β_1 en el modelo de regresión lineal dado por:

$$Y_i = \beta_0 + \beta_1 * T_i + \epsilon_i$$

Además, $Y_i = \{Y_i^t : \ t \in T_i\}$ es la serie de tiempo para la evolución acumulada de contagios en la región indexada por i. Es importante hacer notar que el modelo presentado no tiene aspiración alguna en cuanto a su calidad predictiva respecto de la cantidad futura de contagios. La intuición tras δ_i es poder ponderar el valor de Λ por la *velocidad promedio* en que aumentan los contagios en un determinada región, esto beneficia a una región con muchos contagios pero un lento aumento y por el contrario castiga a otra región que podría tener pocos contagios pero un aumento repentino de casos.

Como tercer elemento, tenemos a $ln(P_i)$, donde P_i es la población total de la región obtenida desde la fuente citada. La función de este factor en la transformación Λ es beneficiar a las regiones más pobladas, donde, naturalmente, uno esperaría observar más casos de Covid-19. Puede Ud. preguntarse por qué no considerar solo P_i en vez de $ln(P_i)$. La respuesta es que el beneficio de este factor sería enorme para regiones populosas y no es la intención *"perdonar"* a aquellos lugares, desde luego los grandes centros urbanos tienen sus desafios y la transformación Λ debe reflejarlos.

Por cuarto y último, tenemos el factor κ, esto es simplemente un multiplicador para el efecto de δ. Teoricamente la única restricción que tenemos es $\kappa \neq 0$. Se sugiere considerar $\kappa \in \mathbb{N}$, donde valores tales que $\kappa > 1$ aumentan el efecto de δ. En este informe, consideramos solamente $\kappa = 1$, es decir, no se pretende potenciar la influencia de δ en el valor de Λ.

Tenemos entonces el razonamiento tras Λ, es momento de examinar como quedan las regiones ordenadas con este *nuevo criterio*. Sean de referencia las seis regiones con más casos determinadas en la sección anterior.

	Región	Lambda	Delta
7	Metropolitana	17.19275	1534.21208
6	Valparaíso	12.70929	151.51237
11	Biobío	12.66640	136.47475
3	Antofagasta	12.10972	110.55217
8	O'Higgins	11.78599	93.34963
9	Maule	11.77377	89.52006

Respecto de la tabla anterior, en primer lugar, se aprecia que se han mantenido las mismas seis regiones que fueron calificadas con peor comportamiento en la sección anterior.

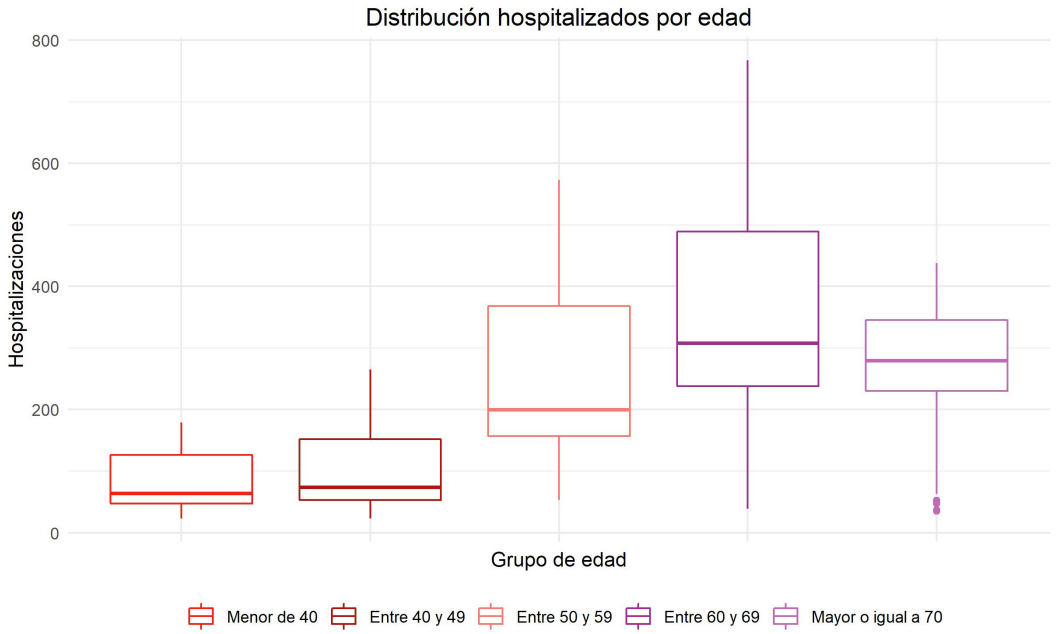
Lo segundo que notamos es que hubo un cambio en el ordenamiento. Con este nuevo criterio la región de Valparaiso tiene un peor comportamiento que la región del Bio-Bio. Situación similar con las regiones del Maule y O'higgins, en donde esta última resulta peor evaluada que siguiendo el criterio del mero conteo de casos. Antofagasta y la R.M. mantienen su posición relativa, aunque con menor diferencia, este último caso no es sorprendente, puesto que como se vio en el gráfico de *casos totales por región* el conteo absoluto de contagiados es mayor a cualquier otra región por un, extremadamente, alto margen.

Por último, rescatamos el hecho que la columna *Delta* (δ) también esta ordenada correlativamente de mayor a menor. Recordar que el criterio de ordenamiento de la tabla es la columna *Lambda* obtenida con la transformación Λ. Esto nos sugiere que la rapidez de contagios en una determinada región tiene un efecto importante en el nuevo criterio de evaluación.

ANOVA para casos por grupo etario

En esta sección utilizamos la técnica ANOVA para analizar los datos de la cantidad de hospitalizaciones y fallecimientos asociados a un diagnóstico de Covid-19. Esta técnica es apropiada cuando los datos están categorizados en grupos, también llamados *factores*. Por esta razón, corresponde aplicar en nuestros datos un *modelo ANOVA de 1 factor*, desde luego este factor es la edad del individuo a quien corresponde cada registro.

Primero estudiamos el caso de las hospitalizaciones, aprovechando la información almacenada en el *producto de datos* **09**. Luego, podemos resumir visualmente como se comporta la distribución de los datos, es decir , *mínimo, máximo, mediana* y los cuantiles `[0.25 y 0.75]`, todo esto para cada grupo de edad por separado. En consecuencia de este objetivo es apropiado considerar un *gráfico de cajon* como el que sigue.



El gráfico anterior nos sugiere que hay diferencias entre los distintos grupos de edad en cuanto a la cantidad de individuos hospitalizados, donde el grupo que más hospitalizaciones presenta es el que comprende edades entre `[60 y 69]` años. Para verificar si estas diferencias son significativas se propone el *modelo ANOVA de 1 factor*. La siguiente tabla resume los resultados de aquel análisis.

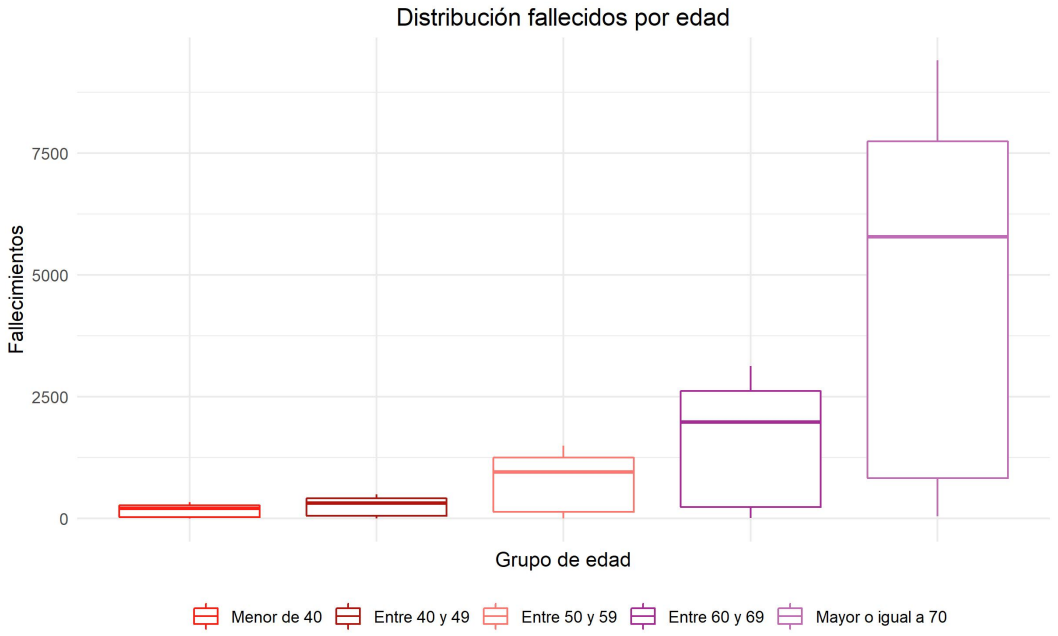
Grupo de edad	Estimación Media
Menor de 40	83.87
Entre 40 y 49	105.113
Entre 50 y 59	260.648
Entre 60 y 69	355.27
Mayor o igual a 70	265.991

Para comprender completamente la tabla anterior, necesitamos definir apropiadamente el concepto de *Estimación Media*. Este valor nos indica cuantas hospitalizaciones esperamos observar, en *promedio*, para cada grupo etario. Entonces, como vemos, en los grupos de mayor edad, se observa en promedio una mayor cantidad de hospitalizaciones. Naturalmente es válido preguntarse si estos valores son *significativos*, es decir, si estos valores reflejan la realidad o son producto de variaciones propias de un fenómeno no determinístico. Para esto se realiza el procedimiento de una *prueba de hipótesis* mediante el cual se ha validado la *significancia* de estos resultados.

Para quien se interese en el detalle técnico del análisis, la hipótesis nula H_0 corresponde a la igualdad de media. H_0 se rechaza con un $Valor_p \approx 0$ en todos los grupos etarios, luego, la diferencia de medias es significativa.

En lo que sigue, se realiza un análisis muy similar, esta vez considerando la cantidad de fallecidos por grupo etario, en vez de las hospitalizaciones. Para tal efecto se considera el *producto de datos* **10**. Es importante mencionar que las categorias de edad suministradas por el Ministerio de Ciencia son diferentes entre los productos de datos *09 y 10*. Por esto se ha aplicado una limpieza de datos, combinando las 3 categorias de mayor edad, con el fin de trabajar con las mismas categorias, tanto en el análisis de hospitalizaciones como en el de fallecimientos. Esto ha de facilitar la interpretación de los resultados.

Primero, comenzamos con el resumen visual equivalente:



Podemos ver como la tendencia se mantiene, es decir a más edad, más casos. Para estos datos, la tendencia es aún más fuerte, puesto que, a diferencia del caso de las hospitalizaciones, el grupo de edad más avanzada es el que más fallecimientos concentra.

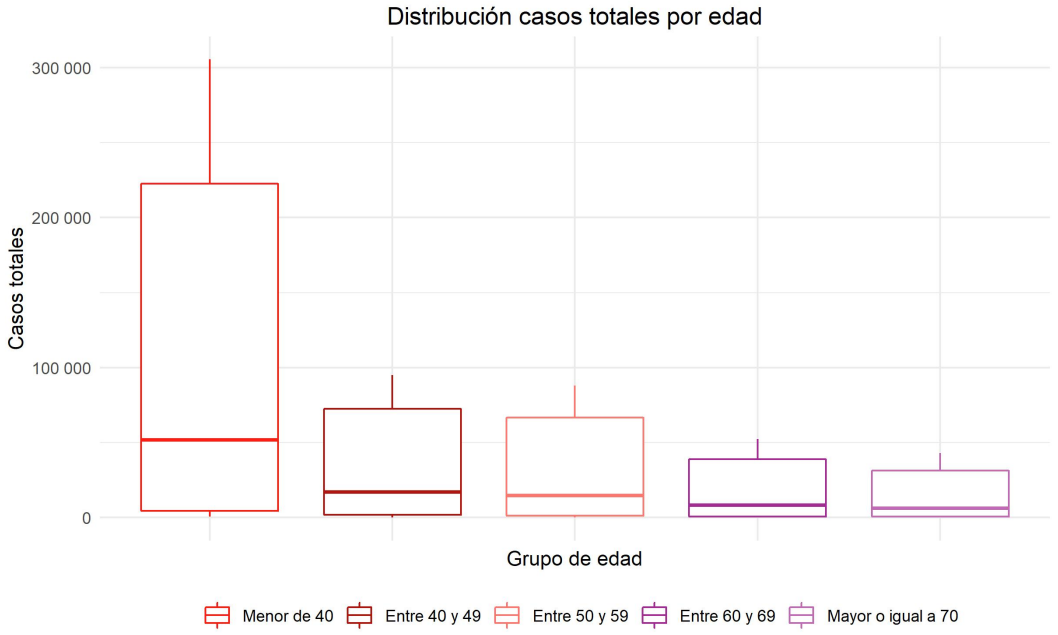
Para apoyar estos hallazgos de manera numérica utilizamos nuevamente el *modelo ANOVA de 1 factor*. La interpretación es equivalente a lo discutido en el caso de las hospitalizaciones. Los resultados se sumarizan en la siguiente tabla.

Grupo de edad	Estimación Media
Menor de 40	170.86
Entre 40 y 49	260.419
Entre 50 y 59	776.032
Entre 60 y 69	1617.523
Mayor o igual a 70	4821.748

Los valores presentados en la tabla anterior se consideran significativos luego de realizar las pruebas de hipótesis correspondientes, tal como ya se discutió. Estos hallazgos confirman que en promedio, durante el intervalo de tiempo analizado en el informe, han habido muchos mas decesos en el rango etario mayor o igual a 70 años.

Para poner en perspectiva los hallazgos anteriores, vale la pena considerar como es la distribución de la cantidad total de casos separando por grupo de edad. Esta información se obtiene desde el *producto de datos* **16**.

Para estos datos también ha sido necesaria una limpieza de la base para que se ajuste a las necesidades particulares de este texto. Si bien la organización original es más fina en cuanto a sus categorías y permite un análisis más estratificado, el objetivo que acá se persigue es poder comparar con lo que ya se presentó hasta este punto. Consecuentemente se visualiza esta información por medio de un *gráfico de cajón*, facilitando la comparación con los gráficos anteriores.



Como se puede ver, a pesar de haber determinado que la mayor cantidad de hospitalizaciones y fallecimientos se concentra en los grupos de edad más avanzada, acá tenemos que la mayor cantidad se contagios se presenta en el grupo de menor edad. Esto nos sugiere que los síntomas de la enfermedad son más agresivos entre pacientes longevos.

Síntesis y proyecciones

En primer lugar, al considerar la transformación Λ, se logro establecer un nuevo criterio para evaluar el comportamiento de las regiones respecto del manejo de la pandemia causada por el Covid-19. En este nuevo ordenamiento, ya no solo miramos la cantidad total de casos, lo cual puede confuso debido a diferencias demográficas u otras, si no que ampliamos la perspectiva considerando la rapidez en el aumento de contagios y la población total de la región. Con lo anterior se pudo determinar que regiones como Valparaiso y O'higgins resultaron peor evaluadas, de esto podemos concluir que el solo número de casos acumulados a la fecha del estudio no captura completamente el comportamiento de estas regiones respecto del brote viral. En el caso de la región Metropolitana, vimos que se mantuvo en el primer lugar, aún considerando el criterio Λ, esto sugiere que la cantidad de casos acumulados si es indicativo del comportamiento de esta región. De todas maneras, el hecho de aplicar la transformación ha aportado en el sentido de poder comparar el comportamiento de la R.M respecto de otras regiones en base a un criterio más estable.

Aunque el criterio Λ amplia la perspectiva para evaluar a las regiones, este sigue considerando solo información relacionada a los contagios. Como extensión, puede considerarse a futuro incluir nuevos datos en la formula de Λ. Un ejemplo de esto sería tomar el factor κ como $\kappa(t)$ donde t es algún parámetro de interés, como el número de camas UCI disponibles, cantidad de profesionales de la salud habilitados, número de centros asistenciales, entre otros.

En segundo lugar, al estudiar la prevalencia de la infección en diferentes grupos etarios, se determinó que efectivamente los grupos de edad mas avanzada han concentrado la mayor cantidad de eventos tanto de hospitalizaciones como de fallecimientos. Luego de haber validado estas diferencias a través de procedimientos

estadísticos se puede decir que en el caso Chileno, también son estos grupos, de edad mas avanzada, los que mas riesgo presentan si es que contraen Covid-19.

Por último destacar la disposición del Ministerio de Ciencia al facilitar estos datos. Sin embargo, como vimos en el caso de la alternativa propuesta para evaluar el comportamiento de las regiones, no solo es suficiente considerar información *inmediatamente* relacionada al Covid-19. Como se propuso en la proyección para la transformación Λ, se pueden enriquecer las métricas al cruzar y combinar los datos Covid-19 con otras fuentes de información.

Bibliografía

Datos Covid 19, **Ministerio de Ciencia, Chile**, *2020*

Adultos mayores reciben el golpe más duro del coronavirus, **Washington Post**, *2020*

No hay un medicamento específico para el coronavirus, **NPR.ORG**, *2020*

Sintesis de Resultados Censo 2017, **INE**, *2017*

Función Lineal, **Wikipedia**, -

Logaritmo, **Wikipedia**, -

Modelo ANOVA de 1 factor, **reliawiki.org**, -