Estadística Aplicada II

Colman, Mariano; Graffigna, Santiago; Soria, Lucas

Universidad de mendoza

TRABAJO PRACTICO FINAL

Correlación entre la tasa de crecimiento de casos activos de COVID-19 y días hasta alcanzar su punto máximo

Tabla de contenidos

[Introducción: 2](#_Toc42895404)

[Marco Teórico: 3](#_Toc42895405)

[Objetivo: 4](#_Toc42895406)

[Descripción del sistema de datos: 5](#_Toc42895407)

[Procedencia: 5](#_Toc42895408)

[Técnicas de recolección de datos: 5](#_Toc42895409)

[Definición de las variables: 5](#_Toc42895410)

[Presentación de los datos: 6](#_Toc42895411)

[Países que aplicaron medidas de distanciamiento social: 7](#_Toc42895412)

[Países que no aplicaron medidas de distanciamiento social: 8](#_Toc42895413)

[Análisis de los datos: 9](#_Toc42895414)

[Análisis de los casos activos de cada país: 9](#_Toc42895415)

[Análisis de los casos activos por millón de habitantes de cada país 15](#_Toc42895416)

[Conclusión: 21](#_Toc42895417)

[Bibliografía: 22](#_Toc42895418)

[Anexos: 22](#_Toc42895419)

# Introducción:

El presente trabajo se realizó en modalidad de trabajo final, para la cátedra de Estadística Aplicada II, de la carrera de Ingeniería en informática de la Universidad de Mendoza. El objetivo del mismo es la aplicación de los conocimientos obtenidos durante el cursado de la materia en un caso práctico.

El análisis realizado en este trabajo pretende demostrar la existencia de correlación entre la tasa de crecimiento de los casos activos de COVID-19 y la cantidad de días transcurridos hasta alcanzar el punto máximo de dichos casos en algunos países del mundo durante la pandemia de coronavirus a junio del año 2020. Se analizo la tasa de crecimiento de casos activos de coronavirus en distintos países, algunos de los cuales tomaron medidas de aislamiento social, como Alemania, Argentina y Uruguay, y otros que no lo hicieron, como Estados Unidos, Italia y Suecia. Los datos fueron obtenidos de la Universidad de Virginia.

# Marco Teórico:

El coronavirus o COVID-19 es una enfermedad infecciosa causada por el virus SARS-CoV-2, la cual produce síntomas similares a los de la gripe, tales como: fiebre, tos seca, disnea (dificultad respiratoria), mialgia (dolores o molestias musculares) y fatiga. En los casos más graves puede producir neumonía, síndrome de dificultad respiratoria aguda, sepsis o choque séptico. Los síntomas aparecen entre los dos y catorce días a partir de la exposición al virus.

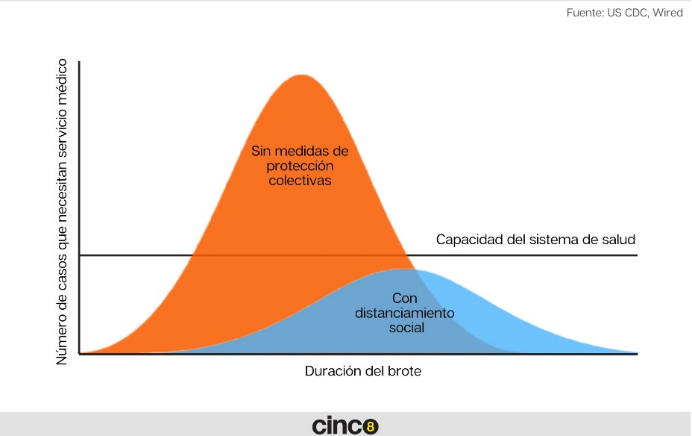
Al día de la realización de este análisis no existe un tratamiento específico y las medidas terapéuticas principales consisten en aliviar los síntomas y mantener las funciones vitales del paciente. Esto lleva a que cerca del 3.75% de los casos de COVID-19 resulten en muerte, según la OMS (Organización Mundial de la Salud).

La transmisión de este virus se produce por gotas de Flügge o microgotas, pequeñas gotas de secreciones que se expulsan de forma inadvertida por la boca y nariz al realizar acciones como toser, hablar o incluso respirar. La infección se da cuando el virus transportado en las mismas entra en contacto con las membranas mucosas del cuerpo presentes en ojos, nariz y boca. El virus puede permanecer viable durante largos períodos de tiempo en objetos inanimados que luego pueden ser manipulados por personas que, al tocarse la cara, se contagian.

Los métodos de prevención más comunes y recomendados son evitar el contacto cercano con personas, el uso de mascarilla, el lavado frecuente y eficiente de las manos, o su desinfección con alcohol en gel y evitar tocarse la cara. Para prevenir que una persona contagiada afecte a otras, la medida más efectiva es el aislamiento obligatorio durante catorce días, tiempo que requiere la incubación del virus. Una persona infectada puede ser asintomática o sintomática. En el primer caso, se indica la reclusión en sus hogares, para evitar ser transmisores del virus. Las personas que presentan síntomas, además del aislamiento obligatorio, pueden llegar a requerir ser hospitalizadas para su asistencia médica, dependiendo de su situación.

Al declarase esta enfermedad como una pandemia distintos países impusieron restricciones a sus habitantes, con el fin de controlar su expansión. Algunas de estas medidas fueron cuarentenas voluntarias u obligatorias, medidas de distanciamiento social, el uso de mascarillas y protocolos para el movimiento de la población.

La imposición del aislamiento social tiene como fin disminuir el pico epidémico, para que los servicios de salud puedan reaccionar de la mejor forma al punto máximo de casos activos, pero tiene como contrapartida que prolonga el tiempo en que se observan los casos positivos en la población. Al objetivo de esta medida, se la denomina coloquialmente “aplanar la curva” y se lo grafica de la siguiente manera:



Fuente: Periódico CINCO8, US CDC. Disponible online

# Objetivo:

En base al marco teórico planteado en el apartado anterior, con este trabajo se pretende demostrar que optar por una cuarentena obligatoria, tiene como resultado el “aplanamiento” de la curva y que, en el caso de no aplicarse dicha medida, la curva no sufriría ese efecto.

Para poder demostrar el objetivo se plantea analizar la existencia de una correlación entre la tasa de crecimiento de casos activos desde el paciente cero hasta su pico y los días que demora en alcanzarlo cada país. Se asume que el efecto de aplanamiento prolonga en el tiempo la llegada al pico de casos y disminuye la magnitud del mismo.

Se espera encontrar que existe dicha correlación y que los países que no optaron por el aislamiento social llegaron antes al máximo de casos activos, y que al mismo tiempo tuvieron un número mucho mayor de personas contagiadas.

# Descripción del sistema de datos:

## Procedencia:

Los datos proceden del sitio web de la NSSAC (Network Systems Science and Advanced Computing), división del BII (Biocomplexity Institute and Initiative) de la Universidad de Virginia. Disponibles en https://nssac.bii.virginia.edu/covid-19/dashboard/ al día del 6 de junio de 2020.

## Técnicas de recolección de datos:

La NSSAC recopilo los datos de distintos “open source datasets”, entre los que se encuentran ASM (American Society for Microbiology), CDC (Centers for Disease Control and Prevention), FDA (Food and Drug Administration), JHU CSSE (Johns Hopkins University Center for Systems Science and Engineering) y WHO (World Health Organization).

Además, se generaron datos a partir de la población de cada país, dato obtenido del respectivo sitio web de cada país en Wikipedia.

# Definición de las variables:

Para poder cumplir el objetivo se deben obtener los casos activos de los países a analizar: Alemania, Argentina, Estados Unidos, Italia, Suecia y Uruguay. Para ello se precisa obtener los casos confirmados de dichos países, tanto de aquellos que han aplicado aislamiento social obligatorio, como de los que no. Además de los casos confirmados se necesita información sobre la cantidad de personas que lograron recuperarse de la enfermedad y la cantidad de personas que perecieron combatiéndola. A partir de estos datos se puede despejar la cantidad de casos activos de COVID-19 en cada país.

Dichos datos son registrados una vez al día, a partir del primer día en donde se presenta el primer caso confirmado de COVID-19, llamado caso 0 (cero). Se tomarán los casos activos hasta el día en que cada país llegue a su respectivo máximo.

Las variables casos activos y días son variables cuantitativas. Cuando se analice la correlación entre la rapidez en que crecen los casos activos y los días hasta el pico de casos, se cualificará la correlación de la siguiente manera:

🡪 Muy mala correlación o correlación inexistente

🡪 Mala correlación o correlación de baja/moderada

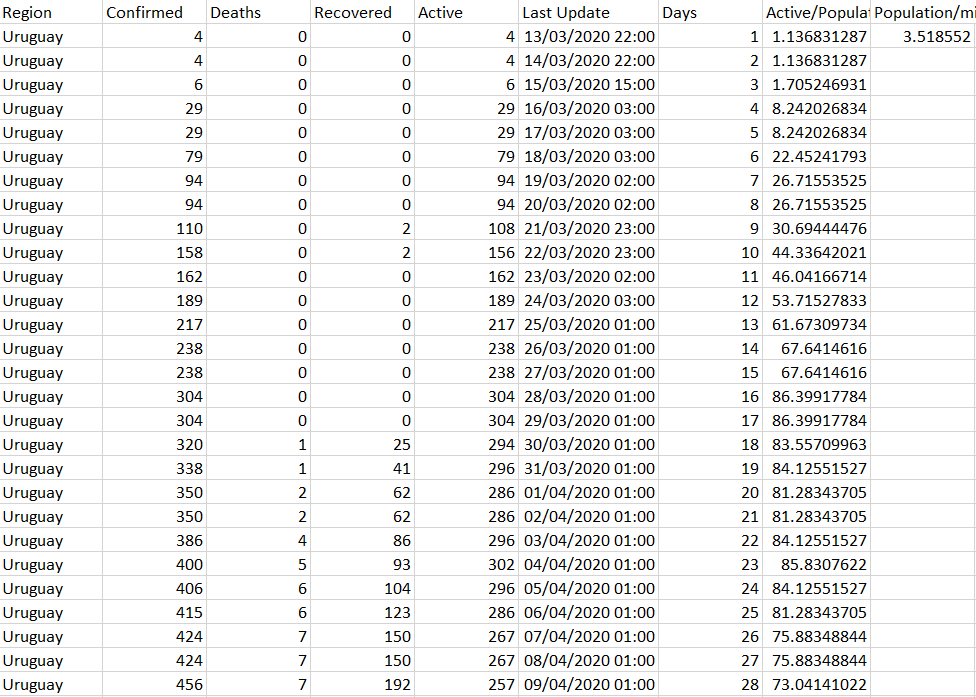
🡪 Buena correlación o correlación moderada/alta

🡪 Muy buena correlación o correlación excelente

Otra variable que se tendrá en cuenta es la población de cada país, para poder utilizar la variable de casos activos por millón de habitantes.

# Presentación de los datos:

Los datos están almacenados en un libro de Microsoft Excel donde se le dedica una página a cada país que se decidió analizar, todas las paginas siguen el formato indicado por la siguiente imagen:



Captura de una sección de la información contenida en la hoja de Uruguay

## Países que aplicaron medidas de distanciamiento social:

Los gráficos representan los casos activos a lo largo del tiempo. En el eje de las ordenadas se encuentran los casos activos y en el eje de las abscisas se localizan los días que transcurrieron desde el primer caso en cada país.

Casos activos de COVID-19 en Alemania Casos activos de COVID-19 en Argentina

Casos activos de COVID-19 en Uruguay

Se puede apreciar en los gráficos que Alemania tardó 71 días en llegar a su punto máximo, el cual es 72969 casos activos. Uruguay tardó solamente 17 días, con máximo de 304 casos. Y Argentina al día 96 desde el caso 0 todavía no llega a su punto máximo, obteniendo 14317 casos activos.

## Países que no aplicaron medidas de distanciamiento social:

Los gráficos representan los casos activos a lo largo del tiempo para los países que no aplicaron la medida. En el eje de las ordenadas se encuentran los casos activos y en el eje de las abscisas se localizan los días que transcurrieron desde el primer caso en cada país.

Casos activos de COVID-19 en Italia Casos activos de COVID-19 en Suecia

Casos activos de COVID-19 en Estados Unidos

En los últimos tres gráficos se puede apreciar Italia tardó 81 días en llegar a su punto máximo, el cual es 108257 casos activos. Por otro lado, tanto Suecia como Estados Unidos no han alcanzado su punto máximo. Suecia posee 34260 casos activos tras 128 días y Estados Unidos tiene 1395568 tras solamente 85 días, convirtiéndolo en el país con mayor cantidad de casos activos de los 6 presentados.

Si bien se toma a Italia como un país que no tomo medidas de distanciamiento social, esto no es cierto. Pero debido a que sus gobernantes decidieron optar por aplicar la medida de cuarentena obligatoria cuando su sistema de salud ya empezaba a saturarse, la tomaremos como un país que decidió no optar por estas medidas.

# Análisis de los datos:

## Análisis de los casos activos de cada país:

Para lograr el objetivo de establecer una correlación entre la tasa de crecimiento de casos activos desde el paciente cero hasta su pico y los días que demora en alcanzarlo cada país, primero se debe definir la tasa de crecimiento para cada país. Para ello se realiza una regresión lineal entre los casos activos y días transcurridos.

Una vez hacemos la regresión tenemos los siguientes resultados por país:

|  |  |
| --- | --- |
| Alemania |  |
| Coeficiente de determinación R^2 | 0.616452243 |
| Coeficiente de correlación r | 0.785144728 |
| Ordenada al origen | -17318.43058 |
| Pendiente | 841.5749497 |

Casos activos de COVID-19 en Alemania

Como se puede observar, el ajuste de los datos a una recta da como resultado la ecuación: .

Si bien el ajuste tiene como ordenada al origen a un número negativo de casos (situación que no puede suceder ya que implica que puede haber una cantidad negativa de personas), solo es de interés la pendiente, la cual indica la tasa de crecimiento que se buscaba.

Analizando el coeficiente de determinación R^2 podemos ver que hay una relación lineal, aunque el ajuste no sea muy bueno. La misma situación se presenta en los demás países.

|  |  |
| --- | --- |
| Argentina |  |
| Coeficiente de determinación R^2 | 0.788316646 |
| Coeficiente de correlación r | 0.887871976 |
| Ordenada al origen | -2412.418421 |
| Pendiente | 118.369452 |

Casos activos de COVID-19 en Argentina

|  |  |
| --- | --- |
| Uruguay |  |
| Coeficiente de determinación R^2 | 0.975076762 |
| Coeficiente de correlación r | 0.987459752 |
| Ordenada al origen | -48.61029412 |
| Pendiente | 20.13970588 |

Casos activos de COVID-19 en Uruguay

|  |  |
| --- | --- |
| Italia |  |
| Coeficiente de determinación R^2 | 0.844389705 |
| Coeficiente de correlación r | 0.918906799 |
| Ordenada al origen | -30976.59784 |
| Pendiente | 1599.274142 |

Casos activos de COVID-19 en Italia

|  |  |
| --- | --- |
| Suecia |  |
| Coeficiente de determinación R^2 | 0.895802745 |
| Coeficiente de correlación r | 0.946468565 |
| Ordenada al origen | -7114.194513 |
| Pendiente | 257.8049779 |

Casos activos de COVID-19 en Suecia

|  |  |
| --- | --- |
| Estados Unidos |  |
| Coeficiente de determinación R^2 | 0.98126726 |
| Coeficiente de correlación r | 0.99058935 |
| Ordenada al origen | -80840.95546 |
| Pendiente | 18713.3582 |

Casos activos de COVID-19 en Estados Unidos

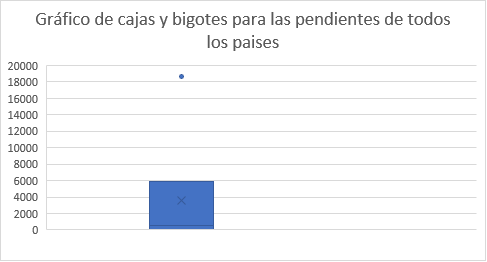
Observando estos datos se puede notar que los países que no han aplicado medidas de distanciamiento social tienen pendientes más pronunciadas y poseen un coeficiente de determinación R^2 mayor, por lo que tienen una mejor aproximación a una recta. La mayor pendiente obtenida para los países que no aplicaron medidas de distanciamiento significa que en ellos los casos de COVID-19 crecen a una tasa mayor.

Luego de obtener la pendiente de la recta del ajuste lineal de todos los países, se analiza si existe correlación con los días que transcurren hasta el máximo de casos.

|  |  |
| --- | --- |
| Correlación entre pendiente y días |  |
| Coeficiente de determinación R^2 | 0.006339531 |
| Coeficiente de correlación r | 0.079621169 |
| Ordenada al origen | 2297.758796 |
| Pendiente | 16.24261224 |

Gráfico de las pendientes en relación con la cantidad de días después del primer caso

Después de analizar las variables, se evidencia una muy mala correlación entre las mismas. Al graficarlas se observa que el punto correspondiente a Estados Unidos se encuentra alejado del resto de los datos. Por lo que se decide probar si este valor es uno anómalo dentro de los datos de los países. Para demostrarlo se hace un gráfico de cajas y bigotes.

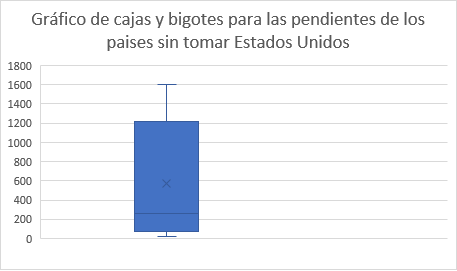


Como se puede apreciar en el gráfico, Estados Unidos representa un valor anómalo. Entonces, se decide repetir el análisis sin incluir los valores correspondientes a este país, como se puede ver a continuación.

|  |  |
| --- | --- |
| Correlación entre pendiente y días |  |
| Coeficiente de determinación R^2 | 0.010547248 |
| Coeficiente de correlación r | 0.102699794 |
| Ordenada al origen | 436.4936469 |
| Pendiente | 1.665890567 |

Gráfico de las pendientes en relación con la cantidad de días después del primer caso sin Estados Unidos

Se puede observar que la correlación entre pendiente de las rectas y los días que transcurren hasta el máximo de casos mejora, pero sigue siendo una muy mala correlación. Se vuelve a realizar el gráfico de cajas y bigotes y no se observan datos anómalos.



Sin embargo, como la correlación no es satisfactoriamente buena, se decide realizar una modificación en la variable casos activos para incorporar nueva información. Se resuelve crear una variable nueva: casos activos por millón de habitantes, definida en el apartado de “definición de variables” como el cociente entre los casos activos y la población del país correspondiente expresada en millones.

La ventaja que presenta esta variable es que permite comparar los casos de COVID-19 entre países con distinta cantidad de habitantes. De esta manera, al repetir el análisis se espera tener una mejor correlación

## Análisis de los casos activos por millón de habitantes de cada país

Repitiendo el proceso anterior se obtiene los siguientes datos:

|  |  |
| --- | --- |
| Alemania |  |
| Coeficiente de determinación R^2 | 0.616452243 |
| Coeficiente de correlación r | 0.785144728 |
| Ordenada al origen | -208.2811351 |
| Pendiente | 10.12125117 |

Casos activos de COVID-19 por millón en Alemania

|  |  |
| --- | --- |
| Argentina |  |
| Coeficiente de determinación R^2 | 0.788316646 |
| Coeficiente de correlación r | 0.887871976 |
| Ordenada al origen | -53.6824113 |
| Pendiente | 2.634019684 |

Casos activos de COVID-19 por millón en Argentina

|  |  |
| --- | --- |
| Uruguay |  |
| Coeficiente de determinación R^2 | 0.975076762 |
| Coeficiente de correlación r | 0.987459752 |
| Ordenada al origen | -13.81542581 |
| Pendiente | 5.723861942 |

Casos activos de COVID-19 por millón en Uruguay

|  |  |
| --- | --- |
| Italia |  |
| Coeficiente de determinación R^2 | 0.844389705 |
| Coeficiente de correlación r | 0.918906799 |
| Ordenada al origen | -511.6631347 |
| Pendiente | 26.41638133 |

Casos activos de COVID-19 por millón en Italia

|  |  |
| --- | --- |
| Suecia |  |
| Coeficiente de determinación R^2 | 0.895802745 |
| Coeficiente de correlación r | 0.946468565 |
| Ordenada al origen | -699.4226738 |
| Pendiente | 25.34575722 |

Casos activos de COVID-19 por millón en Suecia

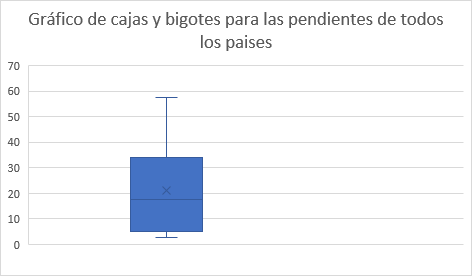
|  |  |
| --- | --- |
| Estados Unidos |  |
| Coeficiente de determinación R^2 | 0.98126726 |
| Coeficiente de correlación r | 0.99058935 |
| Ordenada al origen | -248.192188 |
| Pendiente | 57.45242977 |

Casos activos de COVID-19 por millón en Estados Unidos

Como se puede observar en el siguiente cuadro, se consigue una mejor correlación entre las variables de interés, pero la misma sigue siendo mala. Tomando la variable de casos por millón de habitantes, se consigue eliminar el valor anómalo que generaba Estados Unidos, como se puede observar en el gráfico de la dispersión de las variables, donde se ve que el punto no esta tan alejado del resto de los datos. Esto puede confirmarse en el gráfico de cajas y bigotes, ya que éste muestra que no hay datos atípicos.

|  |  |
| --- | --- |
| Correlación entre pendiente y días |  |
| Coeficiente de determinación R^2 | 0.098736743 |
| Coeficiente de correlación r | 0.314224033 |
| Ordenada al origen | 7.315332957 |
| Pendiente | 0.175317371 |

Gráfico de las pendientes en relación con la cantidad de días después del primer caso

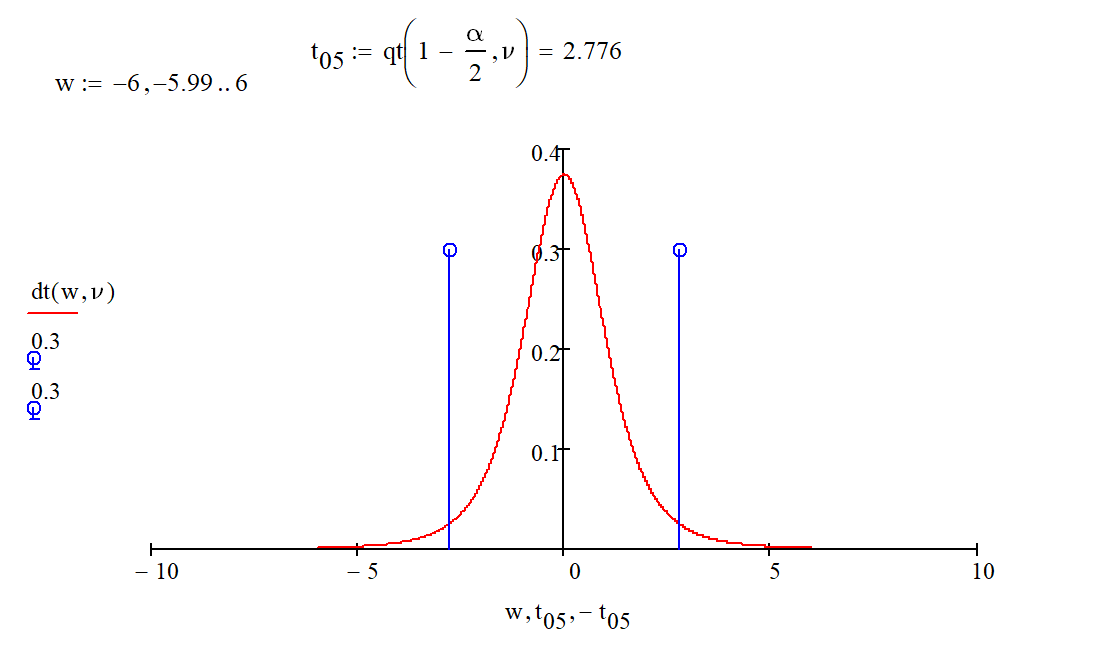


Dado que el coeficiente de correlación es bajo y que la pendiente de la recta generada es cercana a 0 (cero), se decide comprobar si esta es o no significativamente distinta de 0 (cero). Para demostrar si efectivamente las pendientes de las regresiones no dependen linealmente de los días transcurridos hasta el máximo de casos activos, se plantea la siguiente prueba de Hipótesis:

1. Hipótesis Nula:

Hipótesis Alternativa:

1. Nivel de significancia:
2. Criterio:



Siendo:

* = 6

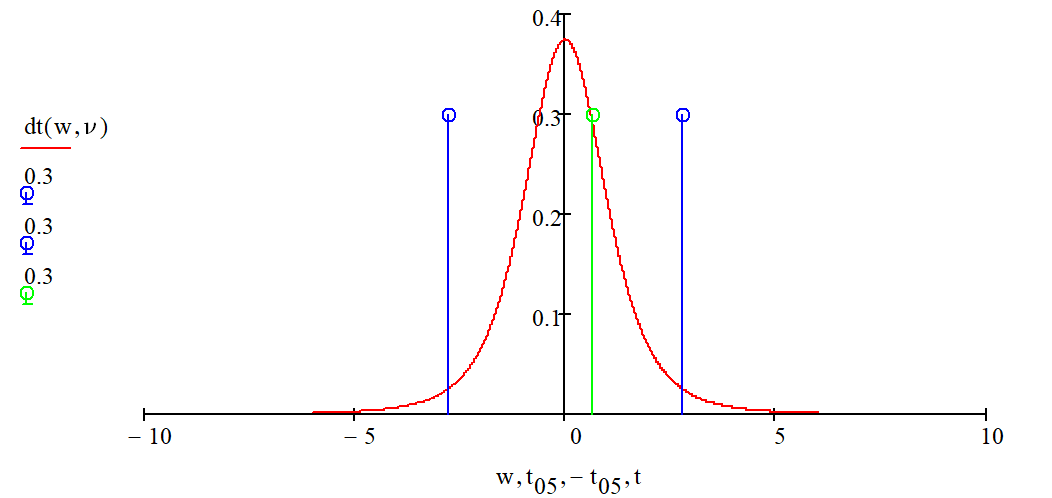
Para esta prueba se considera un nivel de significancia de 0.05, 4 grados de libertad en la distribución de t-student. Dado que es una prueba a dos colas, se obtiene que la hipótesis nula será aceptada si el estadístico t cumple la siguiente condición:

1. Cálculos:

Estadístico t 🡪

Siendo:

* b = Estimación de la pendiente
* = Coeficiente de la regresión lineal
* = Error estándar de estimación
* = Varianza muestral
* = tamaño muestral



1. Decisión:

En base a los resultados obtenidos en la prueba, se acepta la hipótesis nula de que b no es significativamente distinto de 0 (cero). Por lo que la línea de regresión es horizontal y la media de las pendientes no depende linealmente de los días transcurridos.

# Conclusión:

A partir de los datos analizados, se puede desprender que dado que se obtuvo un valor de r muy bajo entre las distintas variables analizadas y que la pendiente de su regresión lineal no es significativamente distinta de 0 (cero), no existe correlación alguna entre la tasa de crecimiento de los casos activos de COVID-19 por millón de habitantes y la cantidad de días transcurridos entre el caso 0 (cero) y el pico de casos activos. Por lo que se puede afirmar que, con los datos obtenidos, no es posible comprobar que los países que no optaron por aplicar medidas de cuarentena obligatoria tempranas, alcanzaron el pico de casos activos antes que los que si aplicaron estas medidas.

Por lo tanto, no se pudo probar el efecto de “aplanamiento” de la curva, sin embargo, se pudo observar la disminución de los casos activos para los países que decidieron aplicar medidas de aislamiento obligatorio.

Una de las razones por las que se concluye esto, es que al momento del presente trabajo, en algunos países los casos activos pueden no haber alcanzado su punto máximo, por lo que no se conoce con certeza cuál será el valor de este punto o cuánto tardará en llegar al mismo.

Asimismo, es probable que la hipótesis no se haya cumplido debido a diferencias culturales entre los países, que hace que en algunos casos sea difícil el sostenimiento del aislamiento social, ya que los habitantes tienden a mantener contacto físico estrecho entre ellos, como lo es en Argentina o Italia.

Además, las políticas adoptadas por los países tendieron a priorizar en algunos casos la salud de la población y en otros el sostenimiento de la actividad económica, como por ejemplo en Estados Unidos, que no aplicaron medidas de distanciamiento social por razones económicas.

# Bibliografía:

* Gabaldón, Juan Carlos. *El dilema de la estrategia contra el COVID-19*. <https://www.cinco8.com/perspectivas/el-dilema-de-la-estrategia-contra-el-covid-19> [Consulta: 2 de junio de 2020]
* Wikipedia. *Alemania.* <https://es.wikipedia.org/wiki/Alemania> [Consulta: 10 de junio de 2020]
* Wikipedia. *Argentina.* <https://es.wikipedia.org/wiki/Argentina> [Consulta: 10 de junio de 2020]
* Wikipedia. *Estados Unidos.* <https://es.wikipedia.org/wiki/Estados_Unidos> [Consulta: 10 de junio de 2020]
* Wikipedia. *Italia.* <https://es.wikipedia.org/wiki/Italia> [Consulta: 10 de junio de 2020]
* Wikipedia. *Suecia.* <https://es.wikipedia.org/wiki/Suecia> [Consulta: 10 de junio de 2020]
* Wikipedia. *Uruguay.* <https://en.wikipedia.org/wiki/Uruguay> [Consulta: 10 de junio de 2020]
* Biocomplexity Institute and Initiative.*COVID-19 Surveillance Dashboard.* <https://nssac.bii.virginia.edu/covid-19/dashboard/> [Consulta: 6 de junio de 2020]
* Wikipedia. *COVID-19.* <https://es.wikipedia.org/wiki/COVID-19> [Consulta: 4 de junio de 2020]
* Universidad de Mendoza. *Estadística Aplicada II*. <https://virtual.um.edu.ar/course/view.php?id=208> [Consulta: 12 de junio de 2020]
* Wikipedia. *National responses to the COVID-19 pandemic*. <https://en.wikipedia.org/wiki/National_responses_to_the_COVID-19_pandemic#Lockdowns> [Consulta: 12 de junio de 2020]

# Anexos:

Para poder leer las hojas de Excel correspondientes a los datos crudos de los casos de COVID-19 de cada país deberá oprimir la tecla Ctrl + Click (en el enlace que desee abrir) y abrirlo desde una cuenta institucional de la Universidad de Mendoza.

* [Alemania](https://docs.google.com/spreadsheets/d/1lDGtrlJSadcxzyV0jJv8TIW56EdQWVih/edit#gid=924607894)
* [Argentina](https://docs.google.com/spreadsheets/d/1lDGtrlJSadcxzyV0jJv8TIW56EdQWVih/edit#gid=1493531653)
* [Estados Unidos](https://docs.google.com/spreadsheets/d/1lDGtrlJSadcxzyV0jJv8TIW56EdQWVih/edit#gid=1494780784)
* [Italia](https://docs.google.com/spreadsheets/d/1lDGtrlJSadcxzyV0jJv8TIW56EdQWVih/edit#gid=1211447645)
* [Suecia](https://docs.google.com/spreadsheets/d/1lDGtrlJSadcxzyV0jJv8TIW56EdQWVih/edit#gid=2102525546)
* [Uruguay](https://docs.google.com/spreadsheets/d/1lDGtrlJSadcxzyV0jJv8TIW56EdQWVih/edit#gid=301625910)