Métodos estadísticos
07/03/2022

Tarea #1
Estadística descriptiva y programación con R

Acosta Imandt Daniel
Barón Bárcenas Iván
Garduño Martínez Eduardo

1. Censo de Población y Vivienda 2020.

Cada 10 años el Instituto Nacional de Estadística y Geografía (INEGI) levanta el Censo de Población y Vivienda con el objetivo de conocer varias características de los habitantes de México y sus viviendas a nivel nacional, estatal, municipal, por localidad, por grupos de manzanas y hasta por manzana. Descarga la base de datos con los Principales resultados por localidad (ITER) de 2020 de los Estados Unidos Mexicanos y genera la siguiente información:

1.1.

Una base de datos a nivel municipio extrayendo directamente los sub-totales para cada municipio y usa esta base para responder las siguientes preguntas.

- ¿De qué tamaño es tu base de datos? (renglones × columnas y peso en MB).

  La nueva base de datos tiene unas dimensiones de: 2,469 renglones x 232 columnas y un peso de 22.5 Mb

- Obtén la población total de la República Mexicana y busca alguna fuente confiable para corroborar tu suma.

  Existe una población total en México de 126,014,024 personas, este dato puede ser corroborado en el comunicado oficial del INEGI

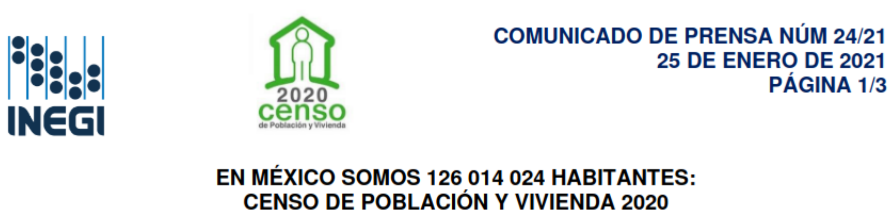

Figura 1: Población mexicana según el censo de población y vivienda 2020

- No filtres y agregues sobre las localidades para cada municipio. ¿Por qué esto sería una pésima idea?

  Ya que hay municipios con 1 o 2 habitantes y en la descripción de todos los habitantes, todos los campos son nulos y la suma de estos no dan el total de habitantes, esto causaría una pérdida de datos sobre los habitantes totales, además de que el resultado ya se encuentra en directamente en la base de datos.

1.2.

El porcentaje de población económicamente activa para cada una de las 32 entidades federativas, así como para las 16 alcaldías que componen la CDMX. Presenta los resultados mediante gráficas de barras y describe lo que te parezca más importante. **Importante:** Calcula estos porcentajes tomando como base la población de 12 años y más en cada estado u alcaldía.

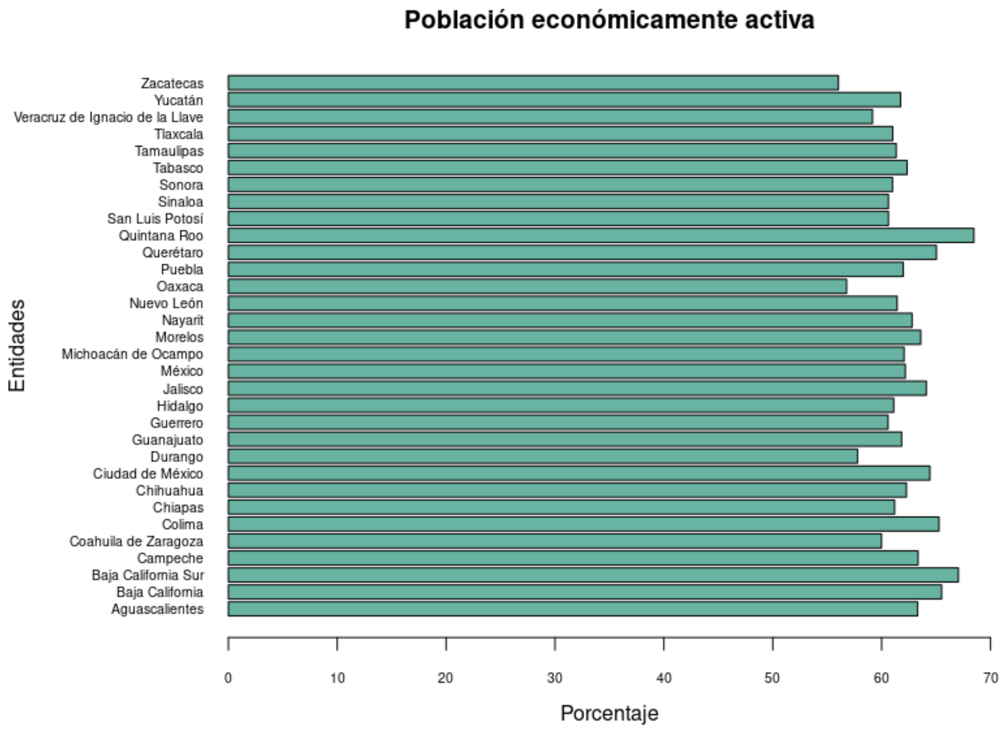

Figura 2: Población económicamente activa por entidad

A partir del histograma de población económicamente activa por entidad, obtenemos resultados interesantes y es que Quintana Roo es el estado con mayor porcentaje de personas con edad para ser considerada económicamente activa, teniendo cerca del $\%70$ de su población económicamente activa, es decir 7 de cada 10 personas.

El caso contrario a esto es Zacatecas y Puebla, estos estados cuentan con menor el menor porcentaje de población económicamente activa, con un poco menos del $\%60$, considerando que el máximo es de casi del $\%70$ la diferencia no es tan abrupta.

Por ultimo notamos que todas las entidades tienen entre el $55\,\%$ y el $70\,\%$ de personas en edad para laborar, se encuentran haciéndolo, como consideramos a la base de la población a la gente mayor de 12 años, este procentaje puede subir si consideramos una base de población diferente, específicamente con una edad mayor ya que en México hay un gran porcentaje de personas jóvene y niños.

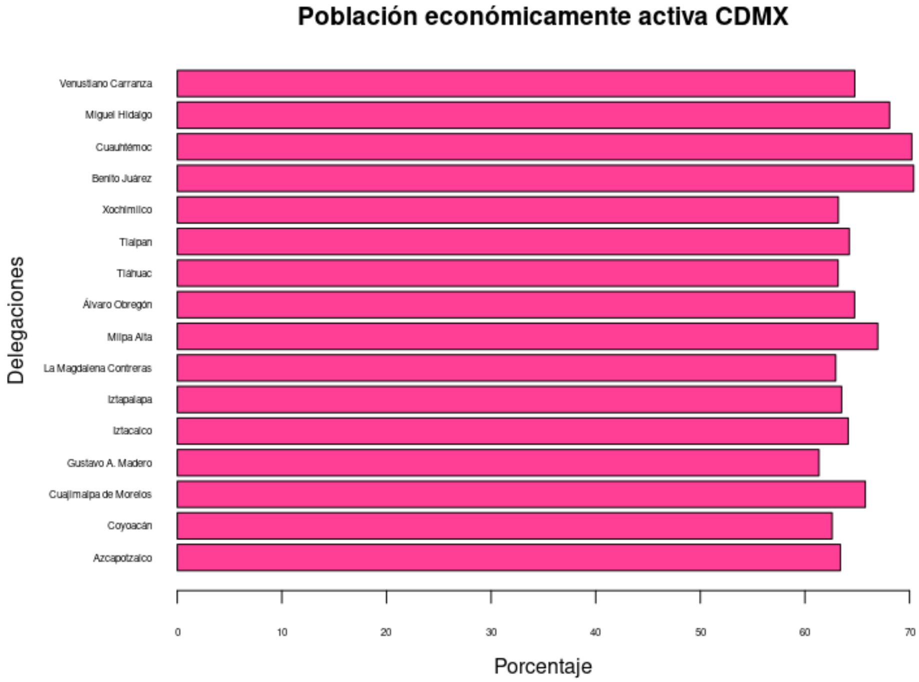

Figura 3: Población económicamente activa por delegación

Al ver el histograma de la población económicamente activa por delegación, vemos que todas las delegaciones cuentan con un porcentaje de población económicamente activa de entre el $60\,\%$ y el $70\,\%$, de dónde la delegaciones Cuauhtémoc y Benito Juárez son las más altas con un un porcentaje un poco arriba del $\%70$, mientras que la delegación Gustavo A Madero es la que tiene el porcentaje más bajo, un un poco más del $\%60$.

Cómo podemos ver en la Ciudad de México el porcentaje de personas activas económicamente se encuentra dentro de los mismos rangos que los niveles nacionales, aunque aquí estaca que en las delegaciones Benito Juárez y Cuauhtémoc cuentan con un porcentaje muy alto.

1.3.

El porcentaje de población de 5 años y más que habla alguna lengua indígena así como el porcentaje de población de 15 años y más analfabeta. Lo anterior para cada uno de municipios que componen nuestro país. Los porcentajes se tienen que calcular con respecto a las poblaciones correspondientes de cada municipio (de forma análoga a lo que se hizo en el inciso anterior). ¿Existe alguna relación entre ambas variables? (realiza un diagrama de dispersión)

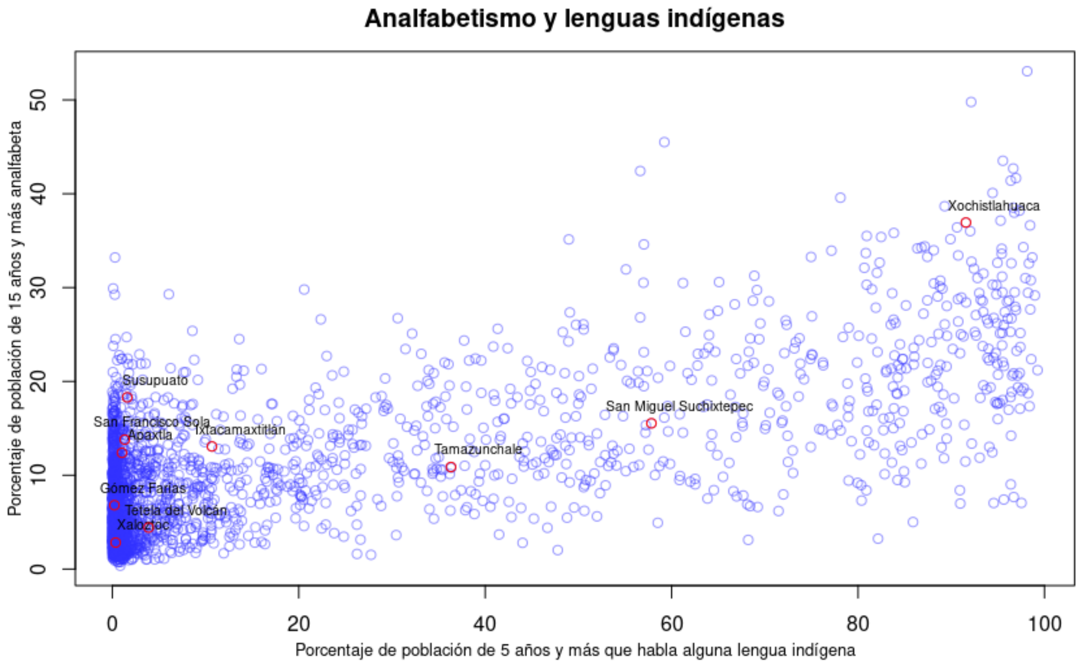

Figura 4: Comparación de personas que hablan una lengua indígena y analfabetas

Al analizar la *Figura 4* notamos que existe cierta correlación entre las personas que hablan alguna lengua indigna y aquellas que no saben leer ni escribir, ya que si el porcentaje de la población que habla una lengua indígena aumenta, suele haber un incremento en el analfabetismo dentro de la mismo municipio, como es el caso del municipio de "Xochistiahuaca", en donde la gran parte de la población habla alguna lengua indígena y cerca del $40\,\%$ es analfabeta.

Por otro lado notamos que en la mayoría de los municipios no se suele hablar ninguna lengua indígena y no solo eso, sino que el porcentaje de analfabetismo es muy bajo en estos mismos municipios, sin embargo existen excepciones, en dónde hay municipios dónde no se habla alguna lengua indígena y sin embargo hay un alto porcentaje de analfabetismo.

Por lo que se debería de promover más la educación y la importancia de la lectura y escritura en las localidades en donde gran parte de la población habla alguna lengua indígena, esto se puede hacer creando más escuelas en estas localidades.

2. Pandemia de COVID-19 en la CDMX.

2.1.

Descarga los datos de la pandemia a nivel nacional: filtra 1) sólo las variables necesarias y 2) la información de los casos positivos a COVID19 (`CLASIFICACION_FINAL` valores 1, 2, y 3) únicamente para las personas que residen en la CDMX.

2.2. Casos, hospitalizaciones y fallecimientos

1. A partir de la fecha de inicio de síntomas (variable `FECHA_SINTOMAS` en formato as.Date) obtén la gráfica del número de casos por mes.

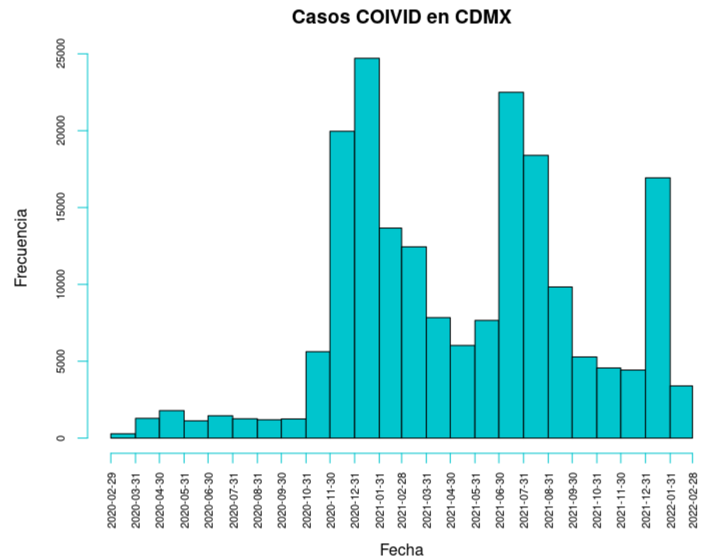

Figura 5: Casos por mes de covid-19 para CDMX

2. Suma 15 días a la fecha de inicio de síntomas y considerando la variable `TIPO_PACIENTE` obtén el número de hospitalizaciones por mes.

   Nota: Para este ejercicio tomamos en cuenta como paciente hospitalizado a los pacientes con un 1 en el `TIPO_PACIENTE`

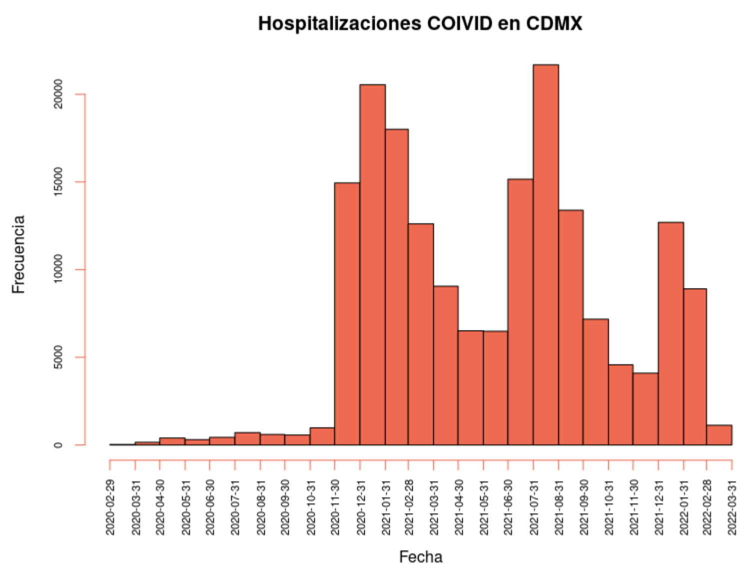

Figura 6: Casos por mes de COVID-19 para CDMX +15 días a la fecha de incio síntomas

3. Con esta información calcula:

$$\text{porcentaje de \ hospitalizaciones por mes} = \frac{\text{hospitalizados por mes}}{\text{infectados} \times \ \text{mes}}$$

Genera una gráfica de serie de tiempo con esta información (en el eje x el mes y en el y los %).

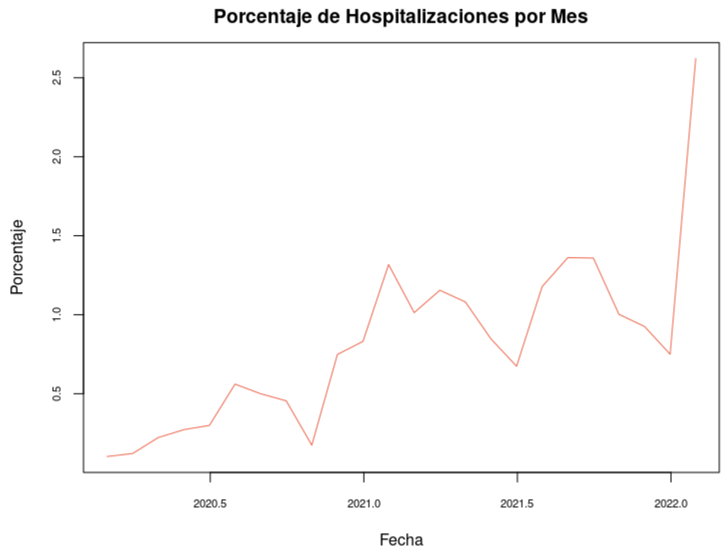

Figura 7: Hospitalizaciones por mes de COVID-19 en CDMX

4. Haz lo mismo con las defunciones. En este caso se cuenta con la fecha de defunción, por lo que no tendrás que estimarla.

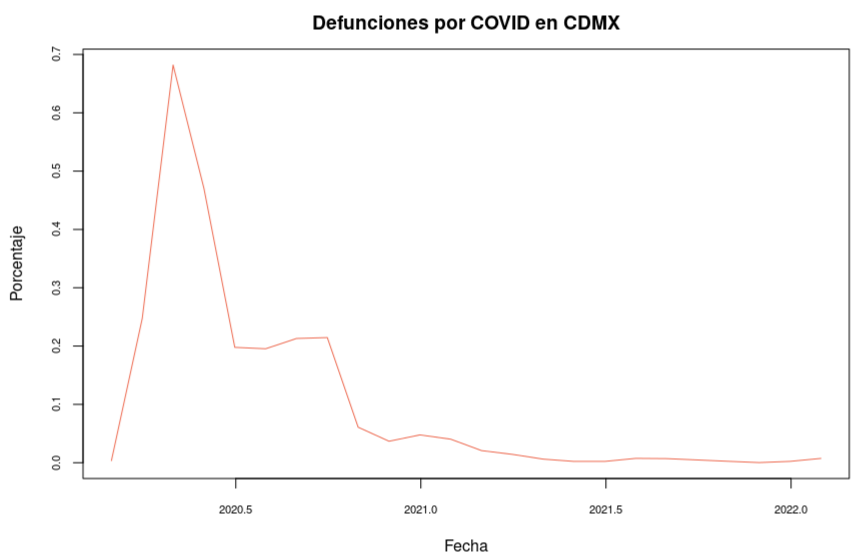

Figura 8: Defunciones por mes de COVID-19 en CDMX

5. Observa bien las tres gráficas y describe lo que observas.

   Tanto para los casos de COVID registrados como para las hospitalizaciones en Ciudad de México que podemos observar en la *Figura 5* y la *Figura 6* tienen una correlación muy evidente, además de que se pueden apreciar claramente las tres olas de infecciones de COVID, el fin de la tercera ola tiene poco menos de un mes de haber terminado.

En cuánto a las series de tiempo de la *Figura 7*, podemos ver que al principio de la pandemia el porcentaje de hospitalizaciones ha ido incrementando, especialmente en este último mes, de primer momento esto puede ser alarmante, sin embargo puede significar todo lo contrario, ya que este último mes ha habido muy pocos casos registrados y de estos son casos graves que requieren hospitalización. Además más adelante veremos que estas últimas hospitalizaciones no son fatales en la mayoría de los casos.

En la *Figura 8* podemos ver que el porcentaje de mortalidad del COVID en un inicio incremento drásticamente, sin embargo ha disminuido notablemente y llegando a estar cerca del $0\,\%$ actualmente, a pesar de que el porcentaje de las hospitalizaciones aumentó, las defunciones por COVID se mantienen a la baja, esto puede ser debido a que la mayor parte de la población se encuentra vacunada con ambas dosis y cómo sabemos la vacuna reduce la mortalidad considerablemente así como el estado de enfermedad grave. Considerando los pocos casos que existen por el momento y la muy baja mortalidad del COVID podemos decir que la pandemia se está acercando a su fin.

2.3. Hospitalizaciones y grupos de edad.

1. Usando las edades de los casos, define los grupos de edad [0, 20), [20, 40), [40, 60), 60 y mayores.
2. Obtén el porcentaje de hospitalizaciones por mes para cada uno de estos grupos de edad y gráfica las las cuatro series de tiempo.

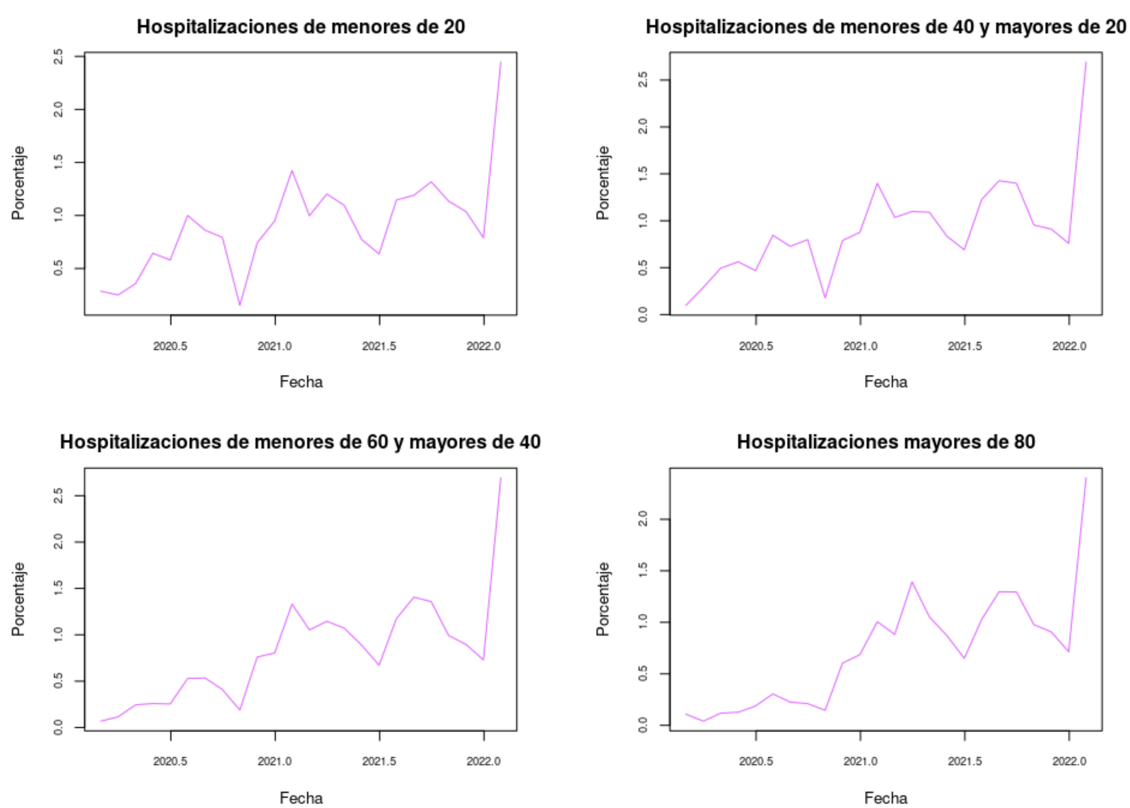

Figura 9: Porcentaje de Hospitalizaciones por COVID en diferentes rangos de edades.

3. Describe lo que observas.

   Lo primero que me salta a la vista es que todas las series de tiempo se parecen mucho entre ellas o comparten muchos rasgos importantes, como el descenso importante en el porcentaje de hospitalizaciones a mediados del 2020 así como el incremento de estas mismas en este último mes.

   Sorprendentemente el porcentaje de hospitalizaciones de las personas menores de 20 es muy similar al de los mayores de 60, incluso para los personas mayores de 60 años, el porcentaje es mucho menor. La única explicación que encuentro es que para una persona relativamente joven se sienta mal y sospeche que tenga COVID es porque lo más seguro es que sí tenga COVID, por lo que el porcentaje de jóvenes hospitalizados por COVID, es muy alto Además de que algo similar ha de suceder en los otros grupos de edad, ya que las personas mayores a 60 cuentan con el porcentaje menor.

3. Repaso de probabilidad y simulación (Juego de Casino)

Un juego de casino consiste en que un jugador pagará m para participar y debe escoger un número n del 1 al 6. Posteriormente, se lanzarán 3 dados (justos) de forma independiente. Por cada dado que haya salido con el número seleccionado n, el casino le dará al jugador m como recompensa. Tres jugadores proponen las siguientes estrategias:
Jugador 1: Siempre pagar m = \$10 y escoger el número$n$ = 6.
Jugador 2: Siempre pagar m= \$115 y escoger el número n = 6.
Jugador 3: Siempre pagar m = \$10 y escoger al azar el número n.
Sea X_i las ganancias del i-ésimo jugador. Encuentra de forma teórica y vía simulaciones del juego para cada una de las estrategias anteriores:

3.1. Obtén la función de masa de probabilidad de X_i para i = 1, 2 y 3.

Primero sacaremos su función de masa de probabilidad vía simulaciones, una vez creado el programa con las reglas del juego y las diferentes estrategias para cada jugador, procederemos a correrlo 50000 veces ,ya que por la ley de los grandes números podemos decir que se va a aproximar al valor real.Una vez hecho esto procederemos a graficar los resultados, que vienen a continuación para cada jugador.
Ahora para el jugador 1 salió que su FMP fue

$$\mathbb{P}(X = 0) \approx 0{,}5812 \quad \mathbb{P}(X = 10) \approx 0{,}34636 \quad \mathbb{P}(X = 20) \approx 0{,}068 \quad \mathbb{P}(X = 30) \approx 0{,}00434$$

Para el jugador 2 salió que su FMP fue

$$\mathbb{P}(X = 0) \approx 0{,}57628 \quad \mathbb{P}(X = 115) \approx 0{,}34894 \quad \mathbb{P}(X = 230) \approx 0{,}07 \quad \mathbb{P}(X = 345) \approx 0{,}00478$$

Para el jugador 3 salió que su FMP fue

$$\mathbb{P}(X = 0) \approx 0{,}5758 \quad \mathbb{P}(X = 10) \approx 0{,}3496 \quad \mathbb{P}(X = 20) \approx 0{,}06983 \quad \mathbb{P}(X = 30) \approx 0{,}00475$$

De lo que podemos observar notamos que , sin importar las reglas que haya seguido cada jugador, las probabilidades para los distintos valores es prácticamente la misma, por lo que al juego le es indiferente las estrategias de los distintos jugadores.

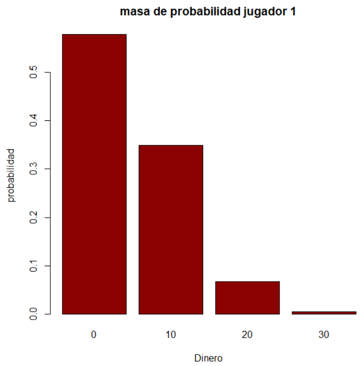

Figura 10: jugador 1

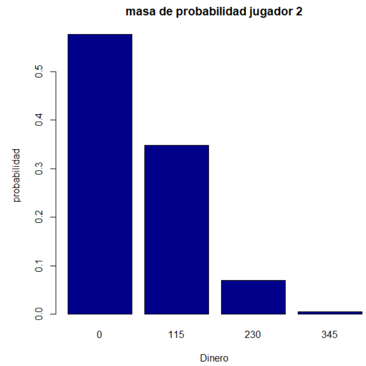

Figura 11: jugador 2

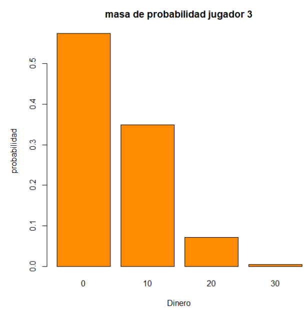

Figura 12: jugador 3

Ahora encontraremos su función de masa de probabilidad de manera teórica.
Para eso vamos a recordar un poco el problema,escogemos un número del 1 al 6, y se lanzan 3 dados , cada uno con $\frac{1}{6}$ de probabilidades de que salga ese número , por ultimo se ve cuantos si salieron ese número y por cada uno se le da lo que apostó.
Por lo que sea $x =$ veces que cae el dado en un número en especifico, entonces tenemos lo siguiente.

$$\mathbb{P}(X=0) = \frac{\cancel{3!}(\frac{5}{6})^3}{0!\cancel{3!}} = \frac{125}{216} \approx 0{,}5787$$

$$\mathbb{P}(X=1) = \frac{3!(\frac{1}{6})(\frac{5}{6})^2}{1!2!} = \frac{25}{72} \approx 0{,}347$$

$$\mathbb{P}(X=2) = \frac{3!(\frac{1}{6})^2(\frac{5}{6})}{1!2!} = \frac{5}{72} \approx 0{,}06944$$

$$\mathbb{P}(X=3) = \frac{\cancel{3!}(\frac{1}{6})^3}{0!\cancel{3!}} = \frac{1}{216} \approx 0{,}00463$$

Notamos que $\frac{125}{216} + \frac{25}{72} + \frac{5}{72} + \frac{1}{216} = 1$.También observamos que los valores obtenidos teóricamente y por simulación son bastante parecidos.
Aquí hace falta recalcar que no importa la estrategia que escogió cada jugador, su masa de probabilidad va a ser la misma ya que las ocurrencias de los dados son independientes y además suponemos que los dados no están cargados.

3.2. $\mathbb{E}[X_i]$ para i=1,2 y3

Ahora sacaremos sus esperanzas de forma teórica.Notamos que solo hay cuatro valores posibles para X y además ya tenemos sus probabilidades por el ultimo inciso, por lo que ya podemos sacar su esperanza de la siguiente forma:

$$\mathbb{E}[X] = (0 \times \frac{125}{216}) + (1 \times \frac{25}{72}) + (2 \times \frac{5}{72}) + (3 \times \frac{1}{216}) = \frac{1}{2}$$

Ahora solo hay que escalar este resultado por el dinero apostado por cada jugador, por lo que sus esperanzas quedan de la siguiente manera.

$$\mathbb{E}[X_1] = \frac{1}{2} \times \$10 = \underline{\$5}$$

$$\mathbb{E}[X_2] = \frac{1}{2} \times \$115 = \underline{\$57{,}5}$$

$$\mathbb{E}[X_3] = \frac{1}{2} \times \$10 = \underline{\$5}$$

Ahora por otro lado vamos a sacar la esperanza de forma computacional, para eso vamos a usar `mean(f)` donde `f` es nuestra simulación, así para el primer jugador salió una esperanza de 4.9972, para el segundo jugador de 57.2 y por ultimo para el tercer jugador tuvimos una esperanza de 5.0184.
Como podemos ver nuestros valores experimentales le pegan bastante bien a los que sacamos de manera teórica y mientras más simulaciones hagamos estos valores se van a ir acercando más, esto por la ley de los grandes números.

3.3. $\mathbb{V}(X_i)$ para i=1,2 y 3

Ahora veremos la varianza de forma teórica.Para eso recordamos que en el caso discreto esta viene dada por:

$$\mathbb{V}(X) = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

$$\mathbb{V}(X_1) = 10^2(\frac{25}{72}) + 20^2(\frac{5}{72}) + 30^2(\frac{1}{216}) - 25 = \frac{125}{3} \approx \underline{41{,}666}$$

$$\mathbb{V}(X_2) = 115^2(\frac{25}{72}) + 230^2(\frac{5}{72}) + 345^2(\frac{1}{216}) - 25 \approx \underline{5487{,}3766}$$

$$\mathbb{V}(X_3) = 10^2(\frac{25}{72}) + 20^2(\frac{5}{72}) + 30^2(\frac{1}{216}) - 25 = \frac{125}{3} \approx \underline{41{,}666}$$

Ahora vamos a sacar la varianza de manera experimental ,para eso vamos a usar la función `var(f)` donde `f` es nuestra simulación ,haciendo esto conseguimos las siguientes varianzas para el jugador 1: 41.66883, para el jugador 2:5573.04 y para el jugador 3: 41.96408.

De nuevo logramos notar que los valores obtenidos teóricamente y los experimentales se acercan demasiado.

3.4. Concluye tus observaciones y compara los resultados teóricos y simulados de cada una de las estrategias propuestas

Al haber simulado el juego del casino logramos notar algo clave desde el principio y es que la probabilidad de que salgas perdiendo todo lo que apostaste en un juego es de $\frac{125}{216}$ lo cual es mayor a $\frac{1}{2}$, por lo que claramente sales perdiendo más dinero del que ganas, al sacar la esperanza vemos que al jugar demasiadas veces este valor es un medio del total apostado, es decir si apuestas $\$1000$ terminaras regresando a casa con algo parecido a $\$500$.
Ahora por otro lado vemos que tanto la estrategia del jugador 1 y del jugador 3 son diferentes,tanto su masa de probabilidad, su esperanza y su varianza son exactamente iguales, esto se debe a que los dos

apostaron la misma cantidad de dinero y aunque el jugador escoja aleatoriamente un número ,estamos suponiendo que los dados son justos, entonces la probabilidad de que caiga ese número en especifico es de $\frac{1}{6}$,notamos que el jugador 2 tiene la misma masa de probabilidad que los otros jugadores, pero tanto su esperanza como su varianza cambian, esto se debe a que el jugador 2 se arriesga más y le mete más dinero a cada juego.
Por ultimo comparamos los resultados obtenidos teóricamente y vía simulación ,donde notamos que estos valores se parecen bastante y conforme aumentamos el número de juegos el error va disminuyendo, esto sabemos que se debe a que conforme vamos aumentando el número de ensayos, entonces su promedio se ira acercando al valor esperado.

En pocas palabras si te ofrecen participar en un juego así no lo hagas , terminaras perdiendo mucho dinero sin importar la estrategia que escoges y además al hacer simulaciones de este juego conforme más hagas su comportamiento será más parecido al valor teórico.

4. Repaso de Estadística y Simulación

Como referencia para este ejercicio, ve el capitulo 7 de "El juego del Calamar" que se encuentra en Netflix (Ver Fragmento). El objetivo será simular la dinámica implementada en el juego "El puente de cristal". Consideremos que hay N = 18 pares de peldaños, y J = 16 jugadores en total. Para este ejercicio, vamos a quitar todas las variantes tales como el tiempo y que dos jugadores puedan estar en un mismo peldaño. Además, asumiremos que todos los jugadores son valientes (no dudarán en avanzar adivinando el siguiente peldaño) y tienen memoria perfecta, es decir si el jugador j con $j \leq J$ ha avanzado hasta el n-ésimo peldaño con $n \leq N$, entonces, los jugadores posteriores a él podrán avanzar sin problema hasta el peldaño n.

Vía simulaciones estima los siguientes valores
De el resultado obtenido
Realizamos 500000 simulaciones y los resultados de estas son:

Valores	0	1	2	3	4	5	6	7	8	9	10
Frec (F_i)	372	1487	5670	16498	35382	60546	83737	92724	83350	60865	35443

Valores	11	12	13	14	15	16
Frec (F_i)	16177	5902	1498	307	38	4

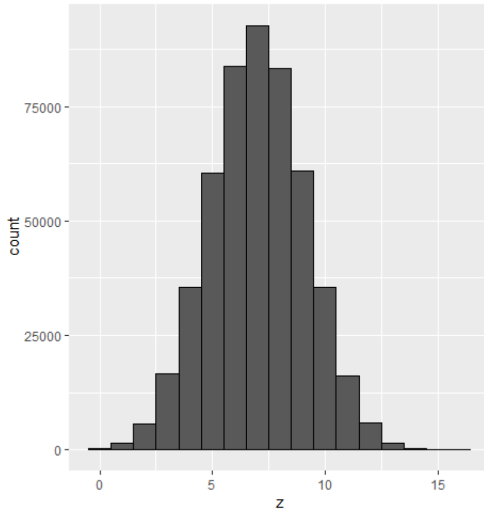

Figura 13: Histograma de los Resultados

4.1. La probabilidad de que al menos crucen la mitad de los jugadores.

Con base en las simulaciones que hicimos obtuvimos que la respuesta es

$$(\sum_{i=8}^{16} F_i)\frac{1}{500000} = (201737)\frac{1}{500000} = 0{,}403474 = 40{,}3474\,\%$$

4.2. ¿Cuántos jugadores pasan en promedio?

Con base en las simulaciones que hicimos obtuvimos que la respuesta es

$$(\sum_{i=0}^{16}(i)F_i)\frac{1}{500000} = (3500261)\frac{1}{500000} = 7{,}000522 \quad \text{personas}$$

4.3. La probabilidad de crucen menos jugadores que el promedio.

Con base en las simulaciones que hicimos obtuvimos que la respuesta es

$$(\sum_{i=0}^{7}(i)F_i)\frac{1}{500000} = (296416)\frac{1}{500000} = 0{,}592832 = 59{,}2832\,\%$$

4.4. El promedio de jugadores que cruza condicionado a que cruzaron más de 5

Con base en las simulaciones que hicimos obtuvimos que la respuesta es

$$(\sum_{i=6}^{16}(i)F_i)\frac{1}{\sum_{i=6}^{16} F_i} = (2993682)\frac{1}{380045} = 7{,}877178 \quad \text{personas}$$

4.5. Da un intervalo de confianza del 90 % de cuántos jugadores cruzarán

Con base en las simulaciones que hicimos obtuvimos que la respuesta es para

$$\mathbb{P}(4 \leq X \leq 9) = (\sum_{i=4}^{9}(i)F_i)\frac{1}{500000} = (416604)\frac{1}{500000} = 0{,}833208$$

$$\mathbb{P}(4 \leq X \leq 10) = (\sum_{i=4}^{10}(i)F_i)\frac{1}{500000} = (452047)\frac{1}{500000} = 0{,}904094$$

Por lo tanto, un intervalo que posee al menos el $90\,\%$ de confianza es el intervalo $[4, 10]$

Apéndice

Todos los códigos en esta liga, para asegurar su reproducibilidad enlace, así mismo en estas ligas se pueden descargar las bases de datos.
https://www.inegi.org.mx/programas/ccpv/2020/Datos$_abiertos$.
https://www.gob.mx/salud/documentos/datos-abiertos-152127