

Análisis estadístico del impacto social y psicológico del uso de Internet

Naiara Pérez Leizaola 1ºC2

Noelia Saugar Villar 1ºC2



**RESUMEN:**

En este análisis estadístico, a través de diferentes métodos, gráficos, tablas, etc. hemos visto de qué manera puede afectar el uso de internet a la depresión, ya que el experimento llevado a cabo del que hemos sacado los datos estudiaba el impacto social y psicológico que internet puede llegar a tener en la vida de las personas.

En esta investigación se les ofreció internet gratuito a 169 personas de 73 hogares durante uno o dos años, a cambio de que su tiempo y actividad online fuese monitorizada y de que hiciesen ciertos tests al inicio y al final de la prueba. El ensayo reveló que los que pasaban más tiempo utilizando internet estaban más deprimidos al finalizar la investigación. A esa misma conclusión hemos llegado nosotras utilizando los datos sacados del experimento para analizarlos estadísticamente.

Comenzamos con la estadística descriptiva, poniendo en manifiesto las características de los datos y sintetizándolos en un número reducido de parámetros.

Mediante tablas y gráficos como box-whiskers, analizamos el tipo de distribución y su posible asimetría y anomalía, además de los estadísticos de las variables (media, curtosis, etc)

Tras el análisis descriptivo, estudiamos las distribuciones en el muestreo, y la inferencia sobre una población, calculando los intervalos de confianza para la media y la varianza.

Seguimos con el análisis de varianza (ANOVA), donde estudiamos el efecto de los factores F1 y F2 y su interacción doble con las variables, determinando si dicha interacción es significativa o no mediante el análisis de los intervalos LSD.

Por último, mediante la regresión lineal observamos que las variables que más efecto tienen la una sobre la otra son las de depresión antes y después, ya que cuanto mayor es la depresión antes del experimento, mayor es esta después. Para llegar a esta conclusión, hemos usado matrices como la de correlación y gráficos de dispersión, entre otros.

**BASE DE DATOS:**

El número de observaciones es de 169 datos.

En todas las variables cuantitativas continuas faltan uno o dos datos, y en las variables cualitativas no falta ninguno.

Variables cualitativas:

-F1: Género(hombre o mujer)

-F2: Edad(adolescente o adulto)

Variables cuantitativas continuas:

-X1: Uso de Internet(horas medias por semana)

-X2: Depresión antes

-X3: Depresión después

-X4: Ingresos del hogar($)

El link de la base de datos es el siguiente:

<https://dasl.datadescription.com/datafile/depression-and-the-internet/?_sfm_cases=50+1000&sort_order=_sfm_cases+asc+num&sf_paged=13>

**CUESTIONES:**

1. ESTADÍSTICA DESCRIPTIVA
   1. Construye una tabla de frecuencias con la variable F2. Comenta los resultados

Tabla 1. Tabla de frecuencias de la variable

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Clase | Valor | Frecuencia | Frecuencia relativa | Frecuencia acumulada | Rel. acum. |
| 1 | Adolescente | 48 | 0,2840 | 48 | 0,2840 |
| 2 | Adulto | 121 | 0,7160 | 169 | 1,0000 |

Podemos observar que de los dos tipos de edad que hay, adolescente y adulto, el adulto aparece con mucha más frecuencia, con 121 sujetos y frecuencia relativa de 0,716; mientras que en los adolescentes se puede advertir una frecuencia menor, con 48 sujetos y una frecuencia relativa de 0,284.

* 1. Construye un gráfico de barras con la variable F2. Comenta los resultados, indicando lo que representa la escala vertical.

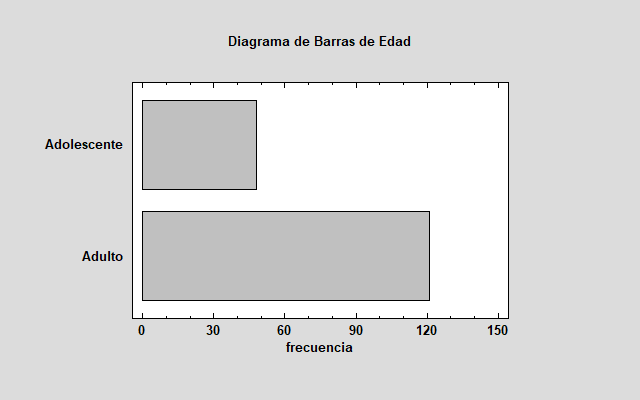
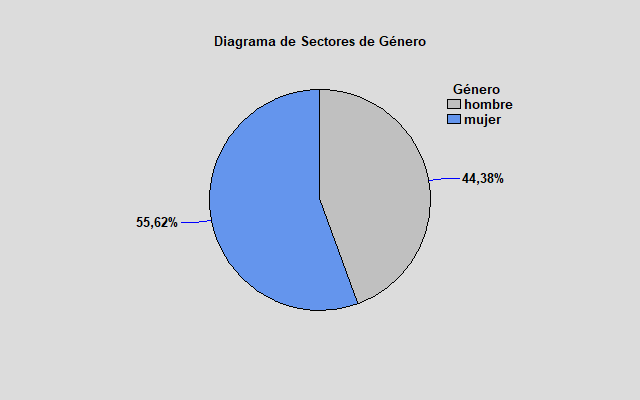


Figura 1. Gráfico de barras con la variable

En el gráfico de barras podemos ver, como hemos observado anteriormente, que la frecuencia de los sujetos adultos es un poco mayor que 120, 121, la cual es más del doble de la frecuencia de los sujetos adolescentes, que se encuentra entre 40 y 50, 48.

* 1. Construye un gráfico de tartas con la variable F1. Comenta los resultados

Figura 2. Gráfico de tartas con la variable 

En este gráfico tenemos los porcentajes del género de los sujetos. Observamos que la mayoría son mujeres, que componen el 55,62%, aunque no por mucha diferencia; mientras que los hombres constituyen el 44,38%.

* 1. Construye una tabla de frecuencias cruzadas (tabla de contingencia) con las variables F1 y F2. Comenta los resultados. Explica la diferencia entre frecuencia absoluta y relativa, y la diferencia entre frecuencia marginal y condicional.

Tabla 2. Tabla de frecuencias cruzadas con las variables y

|  |  |  |  |
| --- | --- | --- | --- |
|  | Hombre | Mujer | Total por fila |
| Adolescente | 20  11,83% | 28  16,57% | 48  28,40% |
| Adulto | 55  32,54% | 66  39,05% | 121  71,60% |
| Total por col. | 75  44,38% | 94  55,62% | 169  100,00% |

Gracias a la tabla de contingencia podemos observar la frecuencia de cada combinación de valores posible entre ambas variables. Al igual que hemos visto en los gráficos de los otros ejercicios sabemos que en la muestra hay 48 adolescentes (28,40%) y 121 adultos (71,60%) y que hay 75 hombres (44,38%) y 94 mujeres (55,62%). Pero en esta tabla también observamos que dentro de la categoría de adolescente, 20 son hombres (11,83%) y 28 son mujeres (16,57%). Dentro de la categoría de adulto vemos que 55 son hombres (32,54%) y 66 son mujeres (39,05%). También se puede observar, tomando como punto de partida que sean mujeres y hombres, cuántos son adolescentes y cuántos adultos, en cuyo caso llegaríamos a los mismos resultados.

Las frecuencias absolutas son el número de sujetos de la muestra que son de una categoría dada, por ejemplo, la frecuencia absoluta de hombres en este caso es de 75. Las relativas son el porcentaje que constituyen los sujetos de la muestra de cierta categoría, en este caso la de mujeres es 55,62%. Las frecuencias marginales son las totales de las categorías y las condicionales, las de los casos de una categoría teniendo en cuenta que ya son de otra de otra variable, como los hombres adolescentes, 20.

* 1. Variables continuas: general
     1. Indica en una tabla los principales estadísticos de las 4 variables continuas: mínimo, máximo, rango, rango intercuartílico, media, mediana, varianza, desviación típica, coeficiente de asimetría estandarizado y coeficiente de curtosis estandarizado.

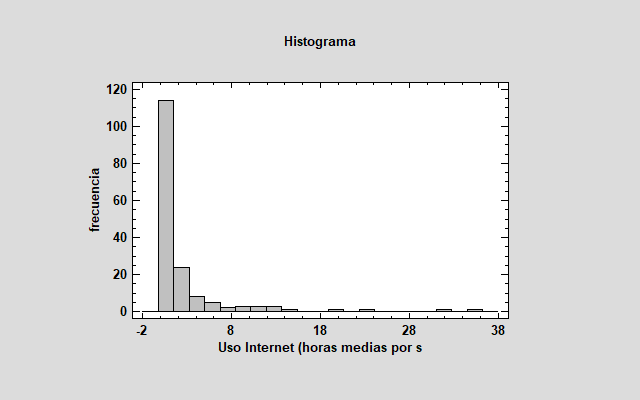
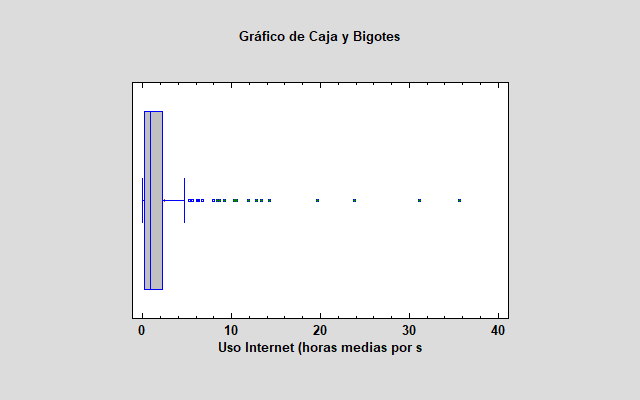
Tabla 3. Tabla de principales estadísticos de X1, X2, X3, X4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| VARIABLES | Uso de internet | Depresión antes | Depresión después | Ingresos del hogar |
| Media | 2,42865 | 0,73479 | 0,617604 | 54,4055 |
| Mediana | 0,855428 | 0,666626 | 0,5 | 52,5 |
| Desviación típica | 4,93701 | 0,485862 | 0,463439 | 22,8239 |
| Varianza | 24,37406 | 0,23606 | 0,214775 | 520,9304 |
| Mínimo | 0 | 0 | 0 | 5,0 |
| Máximo | 35,6186 | 2,5332 | 3,0 | 85,0 |
| Rango | 35,6186 | 2,5332 | 3,0 | 80,0 |
| Rango intercuartílico | 2,02226 | 0,666687 | 0,464233 | 48,5417 |
| Coeficiente de Curtosis Estandarizado | 54,8658 | 2,4279 | 14,5486 | -2,95847 |
| Coeficiente de Asimetría Estandarizado | 21,9192 | 5,08022 | 9,87873 | 0,251344 |

* + 1. Verificar si alguna de las variables tiene una mediana bastante distinta de la media. En caso afirmativo, ¿a qué es debido?

La variable de uso de internet tiene la media y la mediana bastante distintas, siendo la primera 2,42865 y la otra 0,855428, ya que no es de distribución normal (tiene datos anómalos y no es simétrica).

* 1. Coloca en una misma figura: un histograma de la **variable X1** y a la derecha su correspondiente gráfico de caja-bigotes (box-whiskers).



Figuras 3 y 4. Histograma y gráfico caja-bigotes de la variable X1.

* + 1. Comenta brevemente ambos gráficos. ¿Qué indica la escala vertical del histograma? ¿Qué se pretende representar en el gráfico box-whiskers?

La escala vertical del histograma indica la frecuencia de la variable, en este caso el uso de Internet.

En el gráfico de box-whiskers se pretende representar la zona donde se concentran los casos, además de intuir la posible asimetría e identificar los datos anómalos. También es útil para comparar distribuciones.

* + 1. A la vista de los resultados, comenta la pauta de variabilidad de los datos: si la distribución es aproximadamente simétrica o asimétrica (leve, moderada o fuerte), si hay asimetría positiva o negativa. Justifica la respuesta.

Como podemos ver en el histograma y el gráfico de box-whiskers, hay una clara asimetría en esta variable la cual es bastante fuerte y obviamente positiva, ya que en el gráfico de box-whiskers la separación entre el tercer cuartil y el valor máximo es mucho mayor que la distancia entre el primer cuartil y el valor mínimo , lo cual hace que la asimetría sea positiva.

* + 1. Comenta si hay datos claramente anómalos que deberían descartarse del estudio.

Podemos observar que sí que hay datos que son claramente anómalos y se podrían descartar ya que se encuentran muy lejos de donde la mayoría.

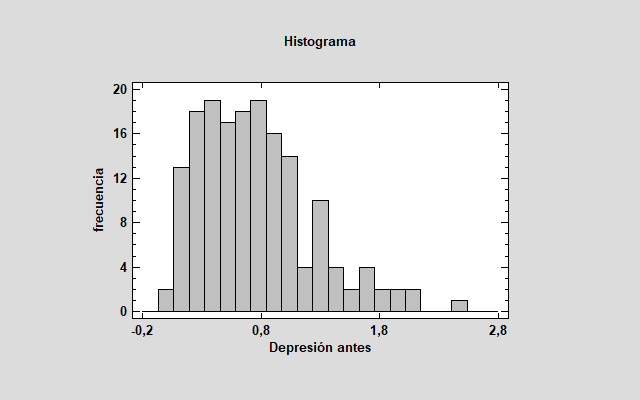
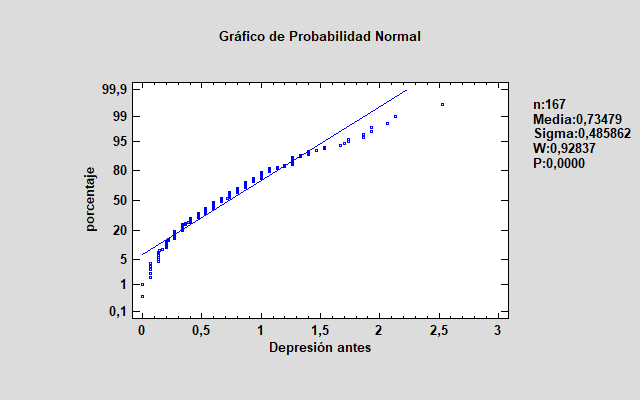
* + 1. A partir de los resultados obtenidos y teniendo en cuenta los coeficientes de asimetría y curtosis, ¿es razonable asumir un modelo de distribución normal para los datos?

Claramente, esta variable no se distribuye normalmente porque sus coeficientes de asimetría y curtosis están muy alejados del intervalo en el que están comprendidos habitualmente ([-2, 2]). También podemos observar que en el gráfico hay una clara asimetría positiva, lo que contribuye a que no se pueda distribuir normalmente.

* + 1. Ventajas e inconvenientes entre ambas técnicas gráficas, ¿cuál de las dos en este caso piensas que aporta una información más útil en este caso? Justifica tu respuesta.

El diagrama de box-whiskers es usado frecuentemente para comparaciones, además de que no necesitan muchos datos para mostrar de forma clara los resultados; mientras que en el histograma son necesarios mínimo 40 datos para su correcto uso. Dicho esto, creo que el histograma muestra con mayor claridad la asimetría y los datos anómalos, ya que tenemos un número de datos elevado.

* 1. Coloca en una misma figura: un histograma de la **variable X2** y a la derecha su correspondiente papel probabilístico normal.

Figuras 5 y 6. Histograma y gráfico de probabilidad normal de la variable X2.

* + 1. Comenta brevemente en qué consiste el papel probabilístico normal y cuál es su utilidad.

Consiste en un gráfico que tiene la línea de donde deberían estar los puntos para considerar que la variable que estamos observando se distribuye normalmente. Si los puntos no se alejan mucho de la línea normal, podemos considerar que la variable tiene una distribución normal.

* + 1. A la vista de los resultados, comenta la pauta de variabilidad de los datos: si la distribución es aproximadamente simétrica o asimétrica (leve, moderada o fuerte), si hay asimetría positiva o negativa. Justifica la respuesta.

Como podemos ver en el histograma y el gráfico de probabilidad normal, hay una asimetría moderada y positiva, ya que en el histograma hay una cola a la derecha y en el gráfico de probabilidad normal los puntos forman una leve curva.

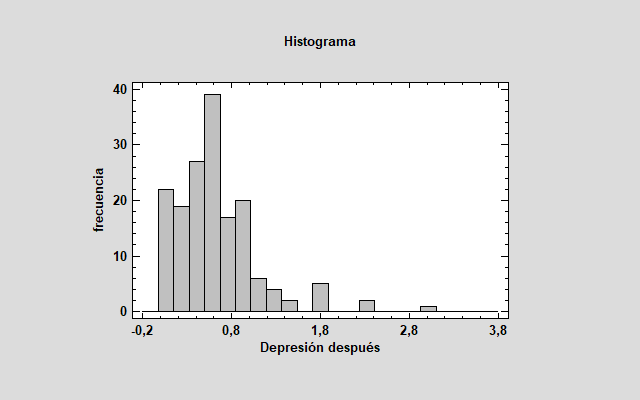
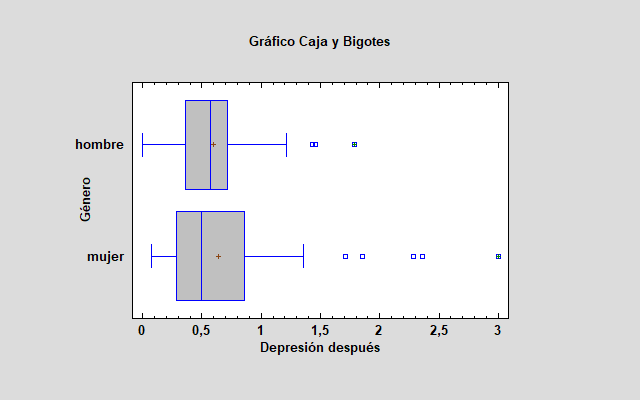
* + 1. Comenta si hay datos claramente anómalos que deberían descartarse del estudio.

Hay varios datos que se podrían descartar, aquellos que se alejan mucho de la recta en el gráfico de probabilidad normal.

* + 1. A partir de los resultados obtenidos y teniendo en cuenta los coeficientes de asimetría y curtosis, ¿es razonable asumir un modelo de distribución normal para los datos?

Esta variable no se distribuye normalmente porque sus coeficientes de asimetría y curtosis no se encuentran dentro del intervalo en el que están comprendidos habitualmente ([-2, 2]), aunque el coeficiente de curtosis se aproxima bastante al intervalo normal (2,4279), lo cual indica que aunque hay datos anómalos, estos no son muchos.

* 1. Coloca en una misma figura: un histograma de la **variable X3** y a su derecha, un gráfico de caja-bigotes múltiple, en función del factor F1.

Figuras 7 y 8. Histograma y gráfico caja-bigotes múltiple de la variable X3.

* + 1. Comenta brevemente en qué consiste el gráfico de caja-bigotes múltiple.

Es un gráfico de box-whiskers habitual, solo que se basa en una variable cualitativa, por lo que separa, en este caso los casos de hombres y mujeres, para poder compararlos con facilidad.

* + 1. A la vista de los resultados, comenta la pauta de variabilidad de los datos (para cada una de las variantes del factor F1): si la distribución es aproximadamente simétrica o asimétrica (leve, moderada o fuerte), y si hay asimetría positiva o negativa. Justifica la respuesta.

En ambos casos se ve que los gráficos son muy parecidos; en los hombres es asimétrica leve ya que la separación entre el tercer cuartil y el valor máximo el ligeramente mayor que la separación entre el primer cuartil y el valor mínimo, mientras que en las mujeres, al ser mayor la diferencia entre ambas separaciones, la asimetría es moderada. Ambos tienen asimetría positiva.

* + 1. Comenta si hay datos claramente anómalos que deberían descartarse del estudio.

En el caso del hombre parece que no hay casi ningún caso claramente anómalo que debería descartarse, en cambio en el de la mujer si que hay varios claramente anómalos, que se encuentran muy alejados del resto de datos.

* + 1. A partir de los resultados obtenidos y teniendo en cuenta los coeficientes de asimetría y curtosis, ¿es razonable asumir un modelo de distribución normal para los datos?.

Podemos asumir que la variable no se distribuye normalmente, porque los coeficientes se salen del rango los dos y porque en los gráficos se ve claramente que son asimétricos, por lo tanto no pueden ser normales.

* 1. Para describir gráficamente la pauta de variabilidad de la **variable X4**, elige el gráfico que consideres que aporta más información (histograma, papel probabilístico normal o gráfico de caja-bigotes).

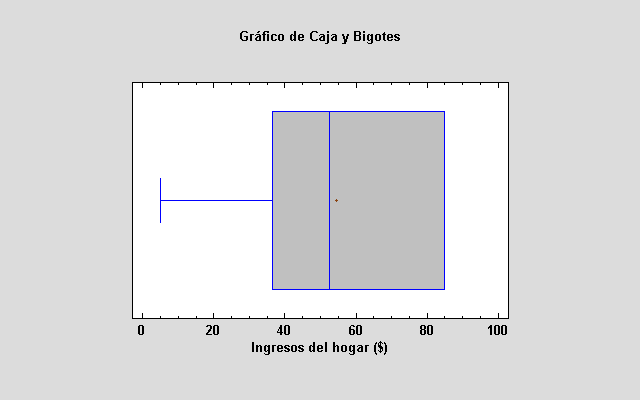


Figura 9. Gráfico de caja-bigotes de la variable X4.

* + 1. ¿Por qué has elegido este gráfico? Justifica la respuesta.

Porque es el que mejor muestra la asimetría negativa, mientras que en los otros gráficos resulta un poco confuso.

* + 1. Comenta las características más relevantes que observes en la representación.

Lo primero que llama la atención es la fuerte asimetría negativa que presenta el gráfico, cuya separación entre el tercer cuartil y el valor máximo ni siquiera llega a ser visible. Además, no hay ningún dato anómalo y la media y la mediana se aproximan bastante.

* 1. Con la variable X1: selecciona aleatoriamente la mitad de los datos (N/2), y multiplícalos por 2. El resto de valores quedan igual. Construye a continuación un histograma con todos los datos (N), comenta el resultado.

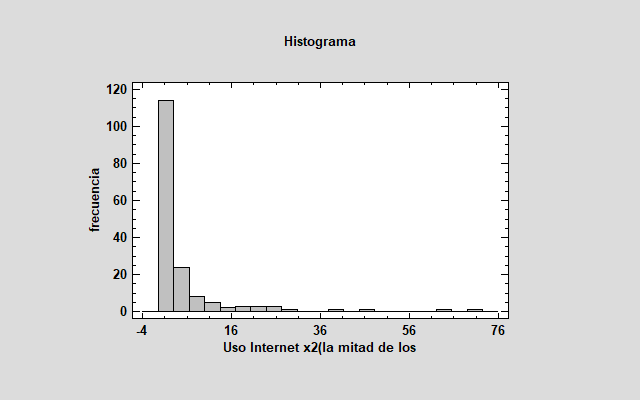


Figura 10. Histograma de la variable con la mitad de los datos partidos por 2.

Como vemos en el histograma, multiplicar los datos por dos no ha cambiado la distribución de la variable, ya que esta sigue siendo una variable exponencial. Este cambio solo ha hecho que los valores del eje x se multipliquen por dos.

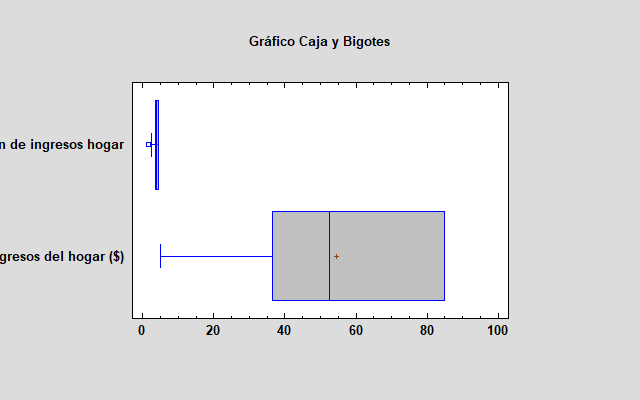
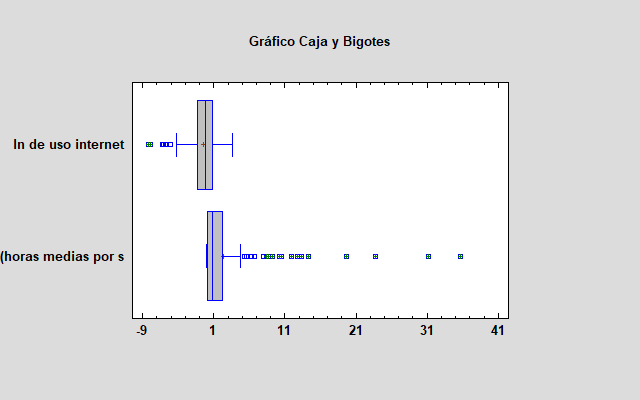
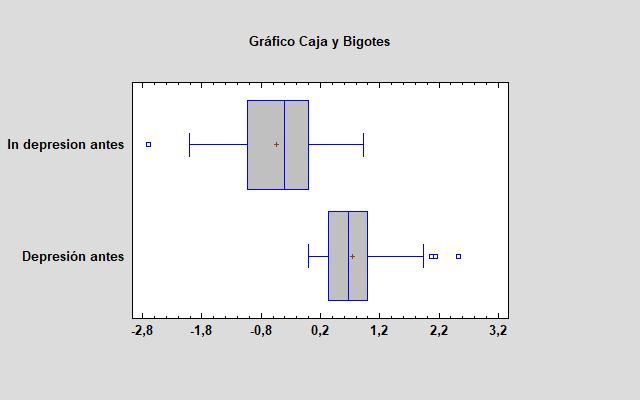
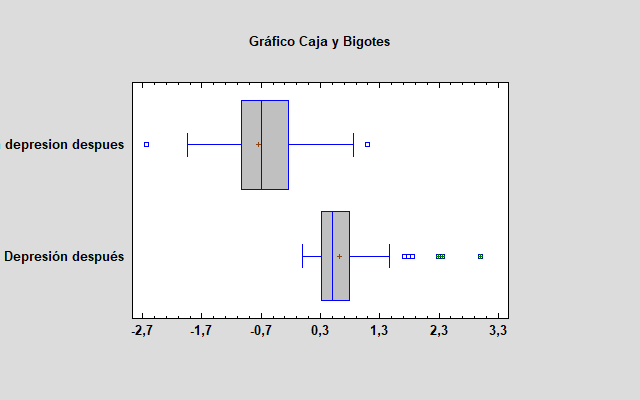
1. DISTRIBUCIONES DISCRETAS Y CONTINUAS
   1. Dadas las variables discretas de los datos que has escogido, ¿su distribución sigue alguno de los tipos estudiados en la asignatura? Justifica tu respuesta y, en caso de que sea afirmativa, indica los valores de los parámetros que describen dicha distribución de probabilidad.

Las variables cualitativas (género y edad) son discretas codificadas, pero no siguen ninguna de las distribuciones que hemos estudiado.

* 1. Dadas las variables continuas de los datos que has escogido, ¿su distribución sigue alguno de los tipos estudiados en la asignatura y que son diferentes a la distribución normal? Justifica tu respuesta y, en caso de que sea afirmativa, indica los valores de los parámetros que describen dicha distribución de probabilidad.

La variable X1 tiene una distribución exponencial, ya que a medida que crecen los valores de la variable disminuye la frecuencia de estos. El parámetro de esta distribución es α, cuyo valor en este caso es 1,88, que se puede calcular utilizando la fórmula P(X≤x)=1-e^(-α.x).

* 1. Aplica logaritmos a los valores de las variables continuas de tu conjunto de datos, ¿siguen alguna distribución de las que se han estudiado en la asignatura?

Figuras 11, 12, 13 y 14. Gráfico caja-bigotes del logaritmo neperiano de las variables X1, X2, X3, X4.

Podemos observar que al hacer el logaritmo de los datos, estos se vuelven más simétricos, lo que hablando sobre distribuciones, hace que se acerquen más a ser una distribución normal, sin importar cual fuese su distribución anteriormente, lo cual se ve claramente que pasa en los gráficos de box-whiskers. Por tanto, de las que hemos distribuciones que hemos dado en la asignatura, podríamos deducir que es normal, aunque no sean exactamente normales, es la que más se acerca.

* 1. ¿Cuáles de las variables que tienes en tu conjunto de datos siguen una distribución normal? ¿Cuáles son sus parámetros? Justifica tu respuesta.

La variable X2 sigue una distribución normal con una muy leve asimetría que hace que se salga de los rangos de los coeficientes de curtosis y asimetría, pero en el gráfico podemos ver que su distribución podría considerarse como normal. Sus parámetros son m=0,73479, ya que esa es su media, y σ=0,485862, ya que esa es su desviación típica.

* 1. Si dos de tus variables siguen una distribución normal, suma sus valores por parejas, con lo que tendrás una nueva variable. Indica qué distribución sigue ésta y qué parámetros la caracterizan. Si no tienes dos distribuciones normales en tu estudio, genera aleatoriamente 100 valores de una distribución N(media=5, sigma=4) y 100 valores de una distribución N(media=2, sigma=3), súmalos por parejas e indica la distribución y parámetros de la nueva variable. Tienes que justificar tu respuesta tanto de manera teórica como apoyándote en las salidas de Statgraphics que lo demuestran.

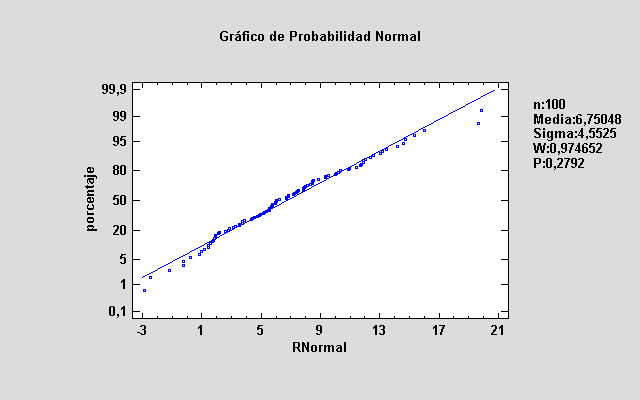


Figura 15. Gráfico de probabilidad normal de la suma de las variables del enunciado.

Tabla 4. Tabla de principales estadísticos de la variable

|  |  |
| --- | --- |
| Recuento | 100 |
| Promedio | 6,75048 |
| Desviación estándar | 4,5525 |
| Coeficiente de variación | 67,4396% |
| Mínimo | -2,91362 |
| Máximo | 19,8852 |
| Rango | 22,7988 |
| Sesgo Estandarizado | 1,85301 |
| Curtosis Estandarizada | 0,402435 |

La nueva variable tiene distribución normal, de media 6,75048 y de desviación típica 4,5525. Según la teoría, la suma de dos variables de distribución normal independientes entre sí (X1 y X2) da otra variable de distribución normal (Y) cuya media y varianza son las sumas de las medias y varianzas de las dos variables independientes. Y=N(mX1 + mX2, (σX1^2 +σX2^2)^1/2)

Además, si analizamos el gráfico de probabilidad normal en Statgraphics, vemos que los datos coinciden con la recta, y el coeficiente de asimetría y de curtosis se encuentran dentro del intervalo normalizado (-2,2).

1. DISTRIBUCIONES EN EL MUESTREO - INFERENCIA SOBRE UNA POBLACIÓN

**3.1.** Asumir que la variable X1 sigue una distribución normal de media igual a la media muestral, y desviación típica igual al valor muestral. Si se toma aleatoriamente una muestra de 5 datos de esta población y se calcula su media, calcular el intervalo de confianza que comprendería el 95% de estos valores.

X1:(/x=2’42865, s=4’93701 )

N muestra aleatoria = 5

⍺=5%

t4 (0’05) = 2’776

[/x - t4(0’05)\*s/, /x + t4(0’05)\*s/] =

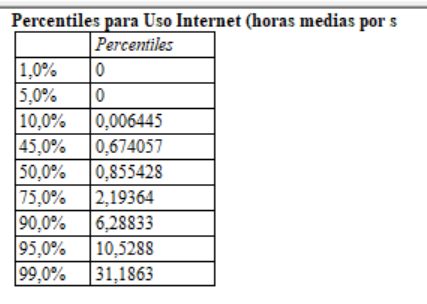
[2’42865 - 2’776\*(4’93701/), 2’42865 + 2’776\*(4’93701/)]

[-3’7, 8’5578] = Intervalo de confianza de la m de la población, con un nivel de confianza del 95%

**3.2.**  Asumiendo que X1 sigue una distribución normal, calcular el percentil 45 de la distribución (Z45). Resuelve el contraste de hipótesis H0: m = Z45 frente a la alternativa: H1: m ≠ Z45. Considera como tamaño de la muestra el número de observaciones (individuos) de la matriz de datos.

N=169

Tabla 5. Tabla de percentiles de la variable X1..



Hipótesis: Z45=0,674057=m

Comprobación:

t168(=0.05)=1.962

|tcalc|===4.617479>1.962, por lo tanto, no se acepta la hipótesis.

**3.3.**  Asumir que la variable X1 sigue una distribución normal de media igual a la media muestral, y desviación típica igual al valor muestral: X ≈ N(m; σ). Si se toma aleatoriamente una muestra de 5 datos de esta población y se calcula su varianza, calcular el intervalo de confianza que comprendería el 95% de estos valores.

X1:(/x=2’42865, s=4’93701 )

N muestra aleatoria = 5

⍺=5%

P()=95%

=0.484

=11.143

[;]=

[;]=

[2.958;14.193] = Intervalo de confianza de la de la población, con un nivel de confianza del 95%

**3.4.**  Asumir que la variable X1 sigue una distribución normal de media igual a la media muestral, y desviación típica igual al valor muestral: X ≈ N(m; σ). Si se toma aleatoriamente una muestra de 5 datos de esta población y se calcula su varianza, ¿cuál es la probabilidad de que se superior a 2·σ?

X1:(/x=2’42865, s=4’93701 )

N muestra aleatoria = 5

P()=P((N-1)\*>(N-1)\*)=P(>4\*)=P(>1.6204)=[0.75;0.9]

P(>1.064)=0.9

P(>1.923)=0.75

**3.5.** Si se toman dos muestras de tamaño 10 de la variable X1, ¿cuál es la probabilidad de que la varianza de la segunda muestra sea más del doble que la primera?

Misma población, por lo tanto =

N (tamaño muestras) = 10

P(>2)=P(>)=P()=0’05

Probabilidad de 0’05 de que la sea más del doble que

**3.6.** Obtener un intervalo de confianza para la media de X1 a nivel poblacional, con un nivel de confianza del 95%.

Intervalos de confianza del 95,0% para la media: 2,42865 +/- 0,75428 [1,67437; 3,18293]

¿Qué interpretación tiene en la práctica el intervalo obtenido?

Para que la media sea aceptada debe encontrarse en el intervalo obtenido, con riesgo del 5%. Por tanto podemos deducir que aunque la media se desvíe un poco la media que tenemos siempre que esté en ese intervalo será aceptable.

**3.7.** En relación a la pregunta anterior, plantea una hipótesis nula sobre la media poblacional que sea compatible con los valores muestrales observados de X1

Hipótesis: m=2, ya que [1,67437< m < 3,18293]

t168(=0.05)=1.962

|tcalc|===0.006678 <1.962, por lo tanto, se acepta la hipótesis.

**3.8.** Obtener un intervalo de confianza para la varianza de X1 a nivel poblacional, con un nivel de confianza del 95%.

Intervalos de confianza del 95,0% para la desviación estándar: [4,45822; 5,53197]

¿Consideras más adecuado indicar este intervalo con un nivel de confianza del 95% o bien del 99%?

Con el nivel de confianza del 99% conseguiremos datos más exactos que con el 95%, pero esto reduciría la posible población y lo más adecuado sería el nivel de confianza del 95% ya que a pesar de ser un poco más flexible, tendría más facilidad que en el otro caso de ser aprobado, sin que creciese mucho el riesgo.

**3.9.** Asumiendo independencia entre X1 y X2, ¿hay suficiente evidencia para afirmar que dichas variables constituyen una muestra aleatoria extraída de una población con distinta varianza?

No, ya que como podemos ver, por ejemplo en la distribución chi cuadrado, que puede haber dos muestras independientes que tengan la misma varianza poblacional.

1. ANÁLISIS DE LA VARIANZA

**4.1.** Realizar un análisis de la varianza (ANOVA) para estudiar el efecto de los factores F1 y F2, y el de su interacción doble, en la variable X1. Considerar un nivel de significación del 1%. ¿El efecto de la interacción doble es estadísticamente significativo?

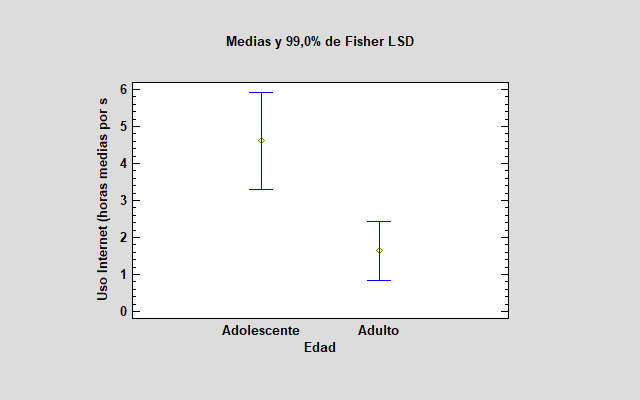


Figura 16. Gráfico de medias con intervalos LSD de la variable X1 según la edad.

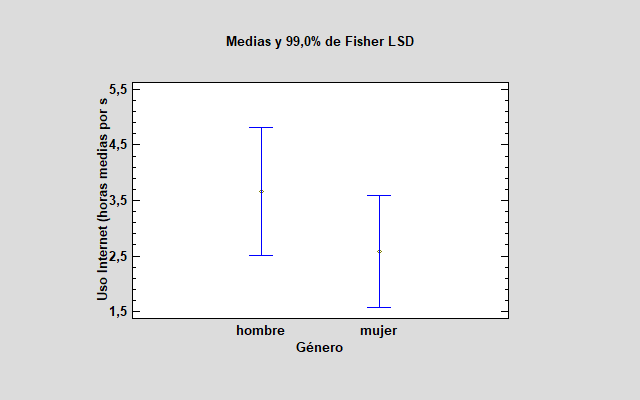


Figura 17. Gráfico de medias con intervalos LSD de la variable X1 según el género.

Como el p-value de F2 (edad) es 0.0005<0.01(), el efecto de la edad es significativo estadísticamente; además, como el p-value de la interacción es 0.1408>0.01(), la interacción no es estadísticamente significativa.

El gráfico de los intervalos LSD es coherente con esta conclusión, ya que los intervalos no se solapan. El factor edad tiene, por lo tanto, un efecto significativo en el uso medio de internet (horas/semana). Este efecto es lineal y negativo, ya que al unir los intervalos con una línea recta, esta tiene pendiente negativa; además, los adolescentes pasan más tiempo usando internet que los adultos.

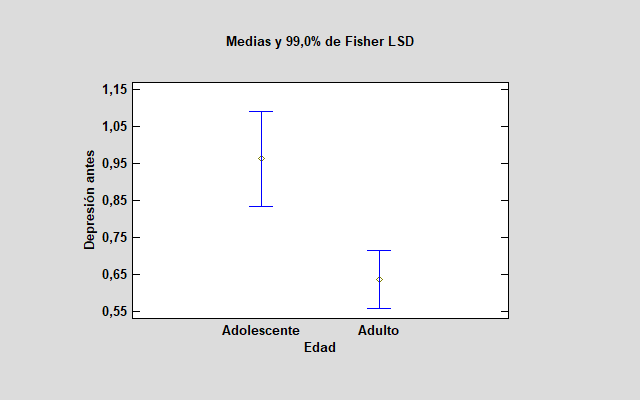
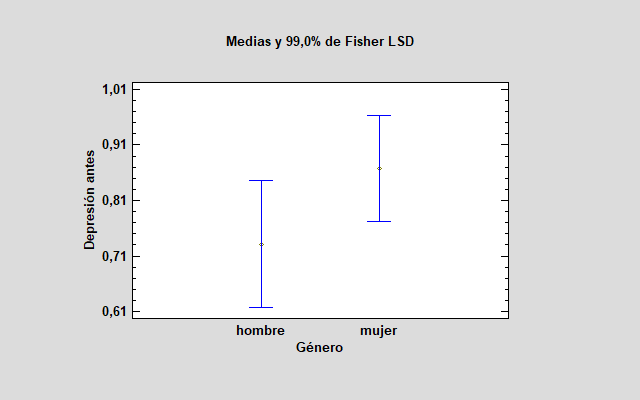
**4.2.** Repite el ANOVA anterior con la variable X2.

Figura 18. Gráfico de medias con intervalos LSD de la variable X2 según la edad.

Figura 19. Gráfico de medias con intervalos LSD de la variable X2 según el género.

Como el p-value de F2(Edad) es 0,0001 el cual es menor que 0.01(), el efecto de la edad en esta variable es estadísticamente significativo, a diferencia del efecto del género, porque el valor de p-value para F1(Género) es 0,0956, el cual es mayor que 0.01(), lo que hace que el efecto no sea estadísticamente significativo. Por lo tanto, la interacción en conjunto no es estadísticamente significativa, al igual que en la variable anteriormente observada.

Al no ser solapados los intervalos en el caso de la edad y sí ser solapados en el caso del género, vemos que el gráfico de estos es coherente con la conclusión. La edad, por lo tanto, sí que es un factor significativo en la depresión antes. Al igual que en la variable anteriormente observada, vemos que esta también tiene un efecto lineal negativo, con una pendiente negativa en el gráfico. También podemos ver que el nivel de depresión antes del experimento es mayor en adolescentes que en adultos.

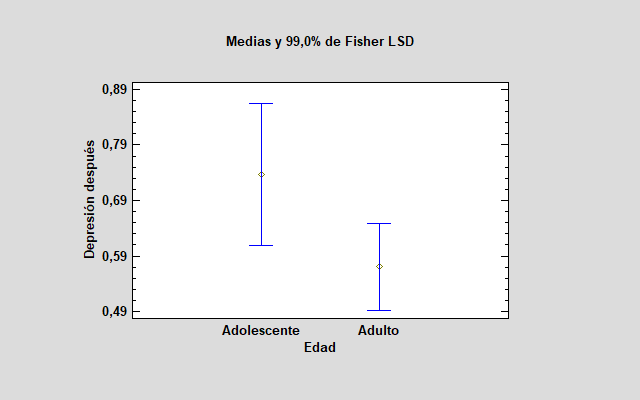
**4.3.** Repite el ANOVA anterior con la variable X3.

Figura 20. Gráfico de medias con intervalos LSD de la variable X3 según la edad.

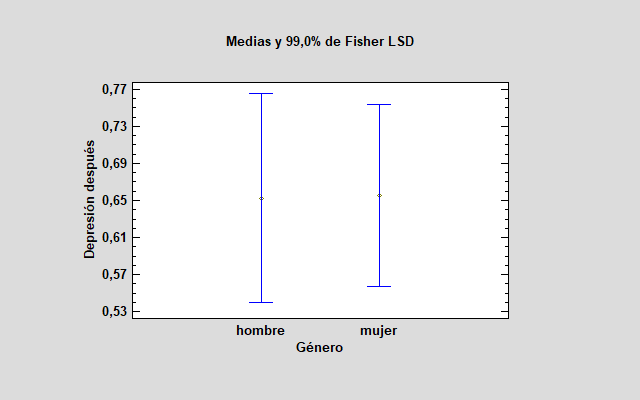


Figura 21. Gráfico de medias con intervalos LSD de la variable X3 según el género.

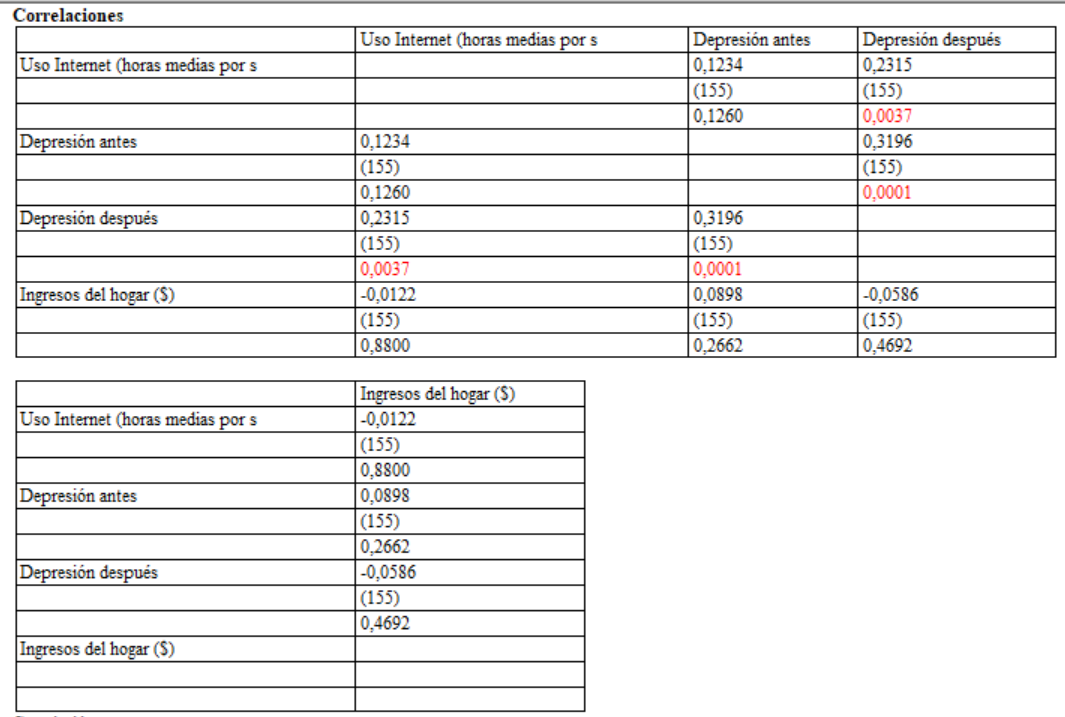
El factor de la edad en la variable de la depresión después del experimento no es estadísticamente significativo, ya que el p-value es 0.0424 el cual es mayor que 0.01(); del mismo modo, el factor del género en la variable no es estadísticamente significativo, ya que el p-value es 0,9727, el cual es mayor que 0.01(). Por lo tanto la interacción no podría considerarse estadísticamente significativa, al igual que en las dos variables observadas anteriormente.

Podemos observar que en los dos gráficos se solapan los intervalos, por lo que sabemos que no es estadísticamente significativo en ningún caso, lo cual concuerda con la conclusión previamente sacada con el p-value. A pesar de ello en la edad se ve una pendiente lineal negativa, y en el género se ve una línea prácticamente horizontal entre los intervalos. También vemos que la depresión después es ligeramente mayor en los adolescentes que en los adultos y es prácticamente la misma en el caso de hombres y mujeres.

1. REGRESIÓN LINEAL

5.1. Obtener una matriz de correlación con las variables X 1 , X 2 , X 3 y X 4 . ¿Qué se deduce de esta matriz? Explicar por qué es simétrica, y cuál es el valor de los elementos de la diagonal principal

Tabla 5. Matriz de correlación con las variables X1, X2, X3 y X4..



El coeficiente de correlación lineal indica la relación lineal entre 2 variables. Por eso, los elementos de la diagonal principal son 1 (son el mismo punto), y al matriz es simétrica ya que, por ejemplo, la correlación entre X1 y X2 es la misma que entre X2 y X1.

5.2. A partir de la matriz anterior, identifica la pareja de variables con mayor grado de correlación.

La pareja de variables con mayor grado de correlación son la depresión antes (X2) y la depresión después (X3); r(X2,X3)= 0’3196

5.2.1. Realiza un gráfico de dispersión entre ambas. ¿Qué se deduce de este gráfico?

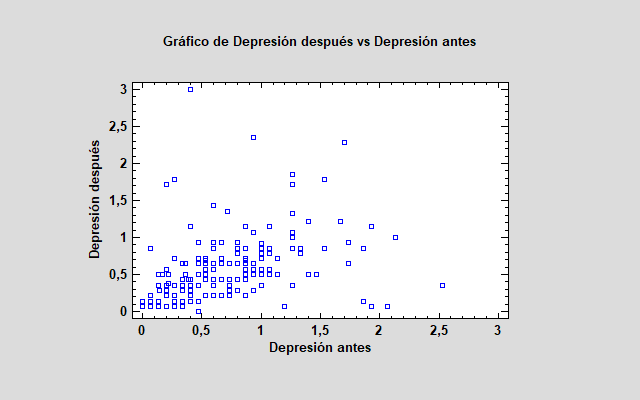


Figura 22. Gráfico de dispersión entre las variables X2 y X3.

Los individuos con mayor depresión antes tienen mayor depresión después, como se puede ver en el gráfico, ya que al aumentar el valor de la depresión antes(eje x), aumenta el valor de la depresión después(eje y). A pesar de que haya algunos valores que se disipan un poco de la linealidad, en general esto se cumple.

5.2.2. Describe la relación entre ambas variables (relación lineal o cuadrática, correlación positiva o negativa, fuerte/moderada/débil, etc.).

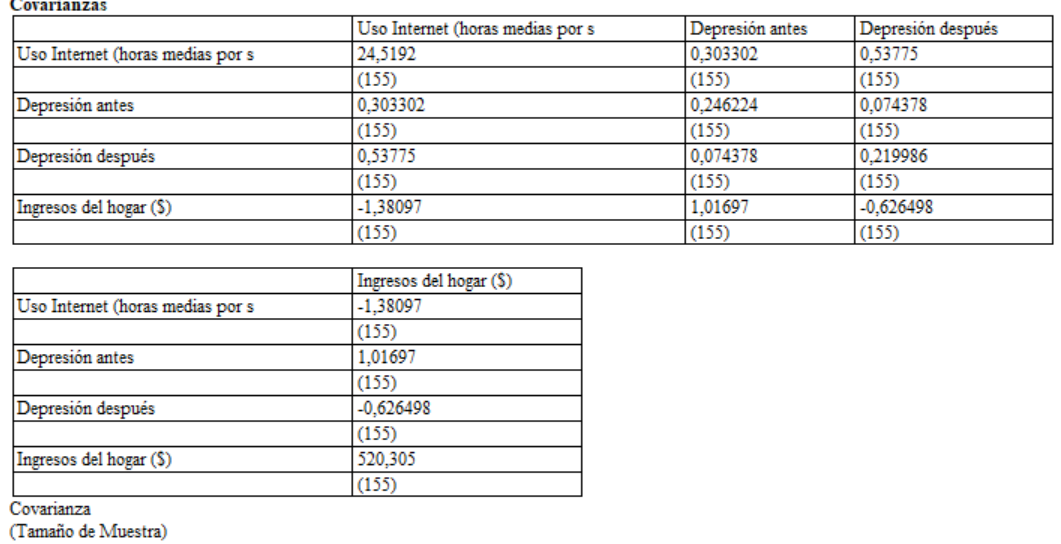
El diagrama pone de manifiesto una relación lineal positiva entre las dos variables estudiadas, que se refleja en una nube de puntos cuyo eje principal tiene un sentido creciente. Esta es débil, ya que a pesar de que se pueda observar la relación lineal positiva, los valores se encuentran bastante dispersos.

5.2.3. Comenta la posible causalidad de la correlación: a partir de la interpretación física de ambas variables, ¿es posible en este caso sospechar que la correlación observada se debe a una relación causa-efecto, a una dependencia parcial, o bien a una interdependencia entre las dos variables?

En esta correlación, podemos ver que hay una dependencia parcial , ya que como observamos en el gráfico de dispersión, a medida que crece la depresión antes, crece la depresión después, pero no están completamente relacionados. Esto es lógico, ya que es normal que alguien que tenga depresión, continúe teniéndola independientemente del internet o tenga una mayor depresión, con más probabilidad que alguien que antes no la tenía.

5.3. Obtener la matriz de varianzas-covarianzas de las 4 variables. ¿Qué utilidad tiene esta matriz?

Tabla 6. Matriz de varianzas-covarianzas de las 4 variables.



Esta matriz, en el eje diagonal muestra las varianzas de las variables y fuera de la diagonal la covarianza entre dos de ellas.

5.4. Entre las 4 variables X 1 a X 4 , elige aquella que podría considerarse como variable respuesta, es decir, que es función de alguna otra variable explicativa. A dicha variable la vamos a llamar Y. A partir de la matriz de correlación, identifica la variable con mayor correlación con Y. Realiza un análisis de regresión lineal simple que permita predecir los valores de Y en función de X:

Variable respuesta: X3 (depresión después)

Variable explicativa: X2 (depresión antes).

5.4.1. Inserta en el trabajo el gráfico de dispersión de Y en función de X junto con la recta de regresión ajustada, y el intervalo de la predicción, con un nivel de confianza del 95%. Escribe la ecuación matemática del modelo: Y = a + b · X.

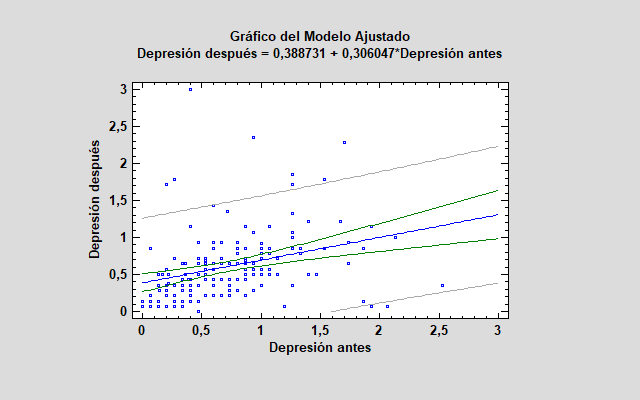


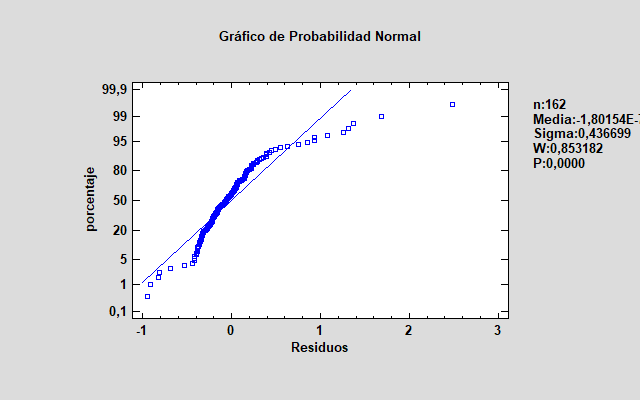
Figura 23. Gráfico de dispersión de Y en función de X junto con la recta de regresión ajustada.

Y = 0.388731+0.306047·x

5.4.2. ¿Cuál es la interpretación práctica de los coeficientes “a” y “b” del modelo?

a (0’388731) se refiere a la depresión resultante cuando la depresión inicial de un individuo es nula; mientras que b (0’306047) es el incremento de la depresión final por cada incremento de la depresión inicial.

5.4.3. Guarda los residuos del modelo y represéntalos sobre un papel probabilístico normal. ¿Qué se deduce?

 Figura 24. Papel probabilístico de los residuos.

Se deduce gracias al papel probabilístico normal que los residuos no siguen una distribución normal, ya que los puntos del gráfico se alejan bastante de la línea que seguiría un distribución normal. También se observa que los coeficientes de curtosis y asimetría se alejan bastante de los intervalos de una distribución normal.

5.4.4. Guarda los residuos del modelo y represéntalos en función de X. ¿Se sospecha que pueda existir un efecto cuadrático? Explica cómo se puede verificar si dicho efecto resulta estadísticamente significativo.

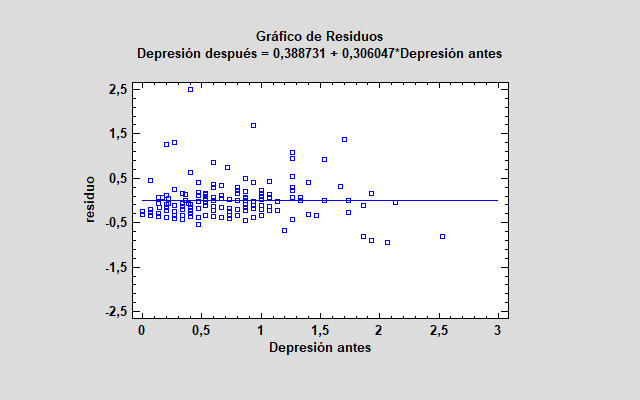


Figura 25. Gráfico de residuos.

Se puede observar que existe un ligero efecto cuadrático, que no es muy perceptible, pero sí que se ve que se eleva ligeramente al principio y luego desciende, creando así una parábola.

5.4.5. Obtener un intervalo de la predicción de Y cuando X vale su primer cuartil (con un nivel de confianza del 95%). ¿Qué interpretación práctica tiene este resultado?

E(Y/X=0,333313) = 0.388731+0.306047·0,333313 = 0.490740443

==0.439132807

0.490740443±2

0.490740443±2\*0.439132807

[-0.387525171 ; 1.369006057]

Este intervalo son los límites entre los que fluctuará aproximadamente la depresión posterior en el 95% de los casos en los que la depresión anterior media sea 0,33.

1. APARTADOS ADICIONALES

6.1. Realizar un ANOVA con más de dos factores e interpretar adecuadamente los resultados.

(no es posible hacerlo por falta de variables cualitativas)

6.2. Estudiar la hipótesis de homocedasticidad en ANOVA (realizando un nuevo ANOVA tomando como variable respuesta los residuos al cuadrado).

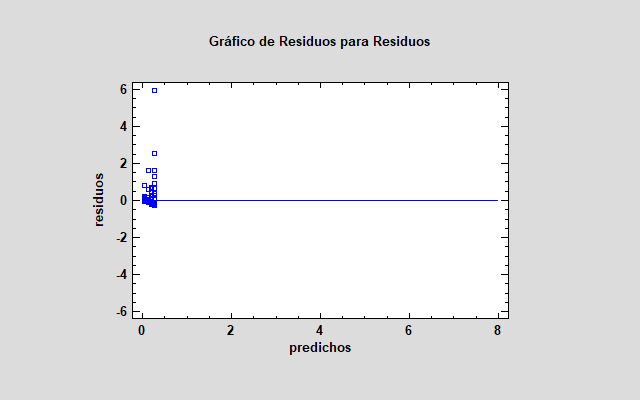


Figura 26. Gráfico de residuos para residuos.

Esta gráfica sirve para detectar una violación de la hipótesis de homocedasticidad (o igualdad de varianzas). En ciertas ocasiones ocurre que la variabilidad de los datos aumenta a medida que aumenta la magnitud del dato. Esto suele suceder en instrumentos de medición, donde el error del instrumento de medición es proporcional a la escala de lectura. En situaciones como esta, la gráfica de residuos frente a predichos se ensanchará como un embudo hacia la derecha, como en este caso.

Es decir, la varianza de los errores no es constante a lo largo del tiempo.

6.3. Realizar un análisis de regresión lineal múltiple para predecir Y en función del resto de variables explicativas disponibles.

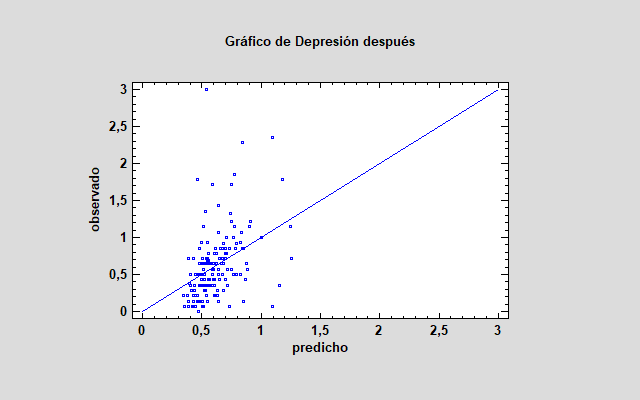


Figura 27. Gráfico de regresión lineal múltiple.

Tabla 7. Tabla de análisis de regresión lineal múltiple.

