UNIVERSIDAD COMPLUTENSE MADRID

Estadística Aplicada – FDI
Ingeniería del Software
6 de mayo de 2022

MEMORIA

Informe del estudio estadístico sobre las horas y calidad del sueño

AUTORES

Trabajo realizado por
· Beatriz Espinar Aragón
· Steven Mallqui Aguilar

LOS HÁBITOS **DE SUEÑO**

CONTENIDO

1. Introducción

Hoy en día, nuestra sociedad se enfrenta a severos problemas derivados de la falta de sueño, con todos los peligros que ello supone para la salud del ser humano. Por supuesto, esta problemática no es ninguna novedad, ya que los horarios laborales, el ritmo de vida y otros factores que influyen en los hábitos de sueño llevan presentes en la sociedad de las últimas décadas. Sin embargo, la revolución de la tecnología, entre otros, ha propiciado conductas que perjudican considerablemente la calidad de sueño. Por consiguiente, los malos hábitos de sueño se han convertido en una de las grandes preocupaciones con las que debemos lidiar actualmente, por las consecuencias que estos provocan en la salud.

Expertos en los campos de medicina y psicología han analizado en numerosas ocasiones los efectos que provocan en nuestro cuerpo y mente dormir menos de lo recomendado, o no de manera suficientemente profunda para permitir al cerebro que descanse lo necesario. La falta de sueño influye no sólo en el rendimiento y la concentración, sino en el estado anímico y, a largo plazo, puede aumentar las probabilidades de padecer ciertas enfermedades, tanto de carácter físico como psicológico.

Por todo ello, y debido al impacto que generan los hábitos de sueño de forma generalizada en la población, se ha decidido llevar a cabo este estudio de naturaleza estadística, con el fin de recopilar datos que ilustren el conflicto al que nos enfrentamos, y del cual poca gente es consciente. El objetivo principal de este informe es, por tanto, registrar, explicar e interpretar los resultados obtenidos a partir del estudio realizado.

El análisis de los datos se ha podido llevar a cabo mediante el lenguaje de programación *R*, empleando el entorno de desarrollo *RStudio*. *R* es un software orientado al análisis estadístico, que integra todas las herramientas necesarias para calcular medidas estadísticas, distribuciones y todo tipo de gráficos.

En primer lugar, la variable principal de estudio, que se tratará más adelante, han sido las horas de sueño efectivo, es decir, el tiempo que realmente está dormido un individuo, incluyendo todas las fases del sueño. Se ha considerado que esta variable es la más representativa para llegar a conclusiones de peso que cumplan con el objetivo de este estudio, dado que es el principal factor que determina de manera directa la calidad del sueño.

En particular, el objetivo que se pretende alcanzar es demostrar de manera suficientemente fiable que existe una gran parte de la población, de hecho, una mayoría, que duerme menos horas de las necesarias, para concienciar de la importancia del descanso respaldando las recomendaciones con datos y hechos objetivos. En este estudio, por tanto, se han realizado los cálculos necesarios para lograr el propósito principal, acompañándolos de otros cálculos de utilidad que permiten al lector tener una idea más completa de la realidad de los datos.

En este documento se explicará en primer lugar el fichero de los datos de los que parte nuestro estudio, con las variables que se han tenido en cuenta, y el por qué de su elección. A continuación, se detallarán todos los análisis, correspondientes a la estadística descriptiva e inferencial, que conforman la base del proyecto. Se incluirán gráficos, explicaciones e interpretaciones que den un valor útil a los resultados obtenidos. Por último, tras presentar las conclusiones finales del estudio, se proporciona una guía de ayuda para utilizar la herramienta de programación y poder así ejecutar el código desarrollado para los cálculos.

2. Fichero de datos

En esta sección se describe el fichero de datos a partir del cual se ha realizado el estudio, junto a una breve explicación de las variables que entran en juego para situar en contexto a los apartados posteriores.

El fichero de datos fue extraído de *Kaggle*, una plataforma dirigida a la comunidad de científicos de datos y profesionales del aprendizaje automático. Estos datos fueron recogidos por *Sleep Cycle App*, una aplicación móvil orientada a monitorizar y registrar los hábitos de sueño del usuario. Así, a partir del *feedback* obtenido de los usuarios, la aplicación recopiló los datos correspondientes a los años 2019-2022, generando el *dataset* que se ha empleado en el proyecto.

La variable principal de estudio ha sido, como se ha mencionado previamente, las horas de sueño efectivo. Esta variable recoge el tiempo que el individuo ha dormido un día concreto. Originalmente estaba expresada en segundos, pero ha sido recalculada en horas por ser esta unidad más representativa para el caso. En el fichero original, traducido a *dataframe*, ocupa la columna número 13, y tiene el identificador "*Time asleep (seconds)*", pero ha sido renombrada con la abreviatura "**HS**" para facilitar el manejo de la variable.

Además de la variable en torno a la cual gira todo el estudio estadístico, se han utilizado otras variables secundarias para servir como apoyo y como medio de comparación para, de este modo, realizar un análisis más exhaustivo de los datos. Así pues, las variables tomadas han sido:

- ✓ Horas en la cama ➜ Recoge el tiempo desde que el individuo se acuesta hasta que se levanta, independientemente de si está dormido o no. Al igual que ocurre con las horas de sueño, estaba inicialmente expresada en segundos y ha sido recalculada en horas para ilustrar mejor el significado de la variable. En el *dataframe* original, ocupa la columna 12 y tiene el identificador "*Time in bed (seconds)*", pero ha sido renombrada con la abreviatura "**HC**", por la misma razón. Esta variable ha sido empleada para compararla con la variable principal, realizando una regresión para comprobar que a mayor tiempo acostado, mayor será el número de horas de sueño.
- ✓ Tiempo antes de dormir ➜ Recoge el tiempo que pasa desde que el individuo se acuesta hasta que se queda dormido. Estaba expresada en segundos y se ha pasado a minutos, por ser lo más intuitivo dado el valor de los datos, que oscilaban entre 0 y 100 minutos. En el *dataframe* ocupa la columna 14 y tiene el identificador "*Time before sleep (seconds)*", pero se ha renombrado del mismo modo con la abreviatura "**TA**". Esta variable resultó interesante desde el principio, por ver si mostraba algún tipo de relación de dependencia tanto con el número de horas de sueño como, especialmente, con la calidad del sueño, que se comenta a continuación.
- ✓ Calidad del sueño ➜ Esta variable se ha calculado de manera artificial, y no hace referencia a la que aparece en el fichero de datos, ya que no se proporciona información acerca de la obtención de los porcentajes. Para referirnos a ella se utiliza la abreviatura "**CS**", y para calcularla basta con dividir las horas de sueño entre las horas en la cama, del siguiente modo: $CS = \frac{HS}{HC} * 100$. Esta variable muestra, por tanto, la proporción en que el individuo realmente duerme con respecto al tiempo que pasa en la cama, y se utiliza para analizar dependencias con las demás variables a través de los modelos de regresión.

El fichero cuenta con otras 18 variables que no se han considerado de interés para este estudio, aunque podrían ser de utilidad para profundizar en los factores que influyen en la calidad del sueño.

Entre otras, registra el momento en que el individuo se acuesta y en el que se levanta (año, mes, día y hora), y el estado de ánimo con el que se levanta.

Una vez escogidas las variables a analizar, podemos comenzar con el estudio. En primer lugar, el fichero de datos tenía un formato Excel (.csv) que ha tenido que traducirse para obtener el *dataframe* y poder trabajar con él. Una vez hecho esto, es inmediato visualizar los datos en forma de tabla, como se adjunta en la imagen 1.

A continuación, para poder analizar HC, HS, TA y CS se obtuvieron las correspondientes variables por separado, tanto en formato *dataframe*, como *tibble*, como *vector*, puesto que todos se emplean en las distintas funciones. Además, se guardan las variables agrupadas para poder realizar posteriormente las regresiones, y el tamaño de la muestra.

	Start	End	Sleep Quality	Regularity	Mood	Heart rate (bpm)	Steps	Alarm mode	Air Pressure (Pa)	City	Movements per hour	Time in bed (seconds)	Time asleep (seconds)	Time before sleep (seconds)	Window start	Window stop	Did snore	Snore time	Weather temperature (°C)	Weather type	Notes
1	2019-05-12 23:26:13	2019-05-13 06:11:03	60%	0%	NA	0	8350	Normal	NA		35.0	24289.2	22993.8	161.9	2019-05-13 06:00:00	2019-05-13 06:00:00	TRUE	92	0.0	No weather	
2	2019-05-13 22:10:31	2019-05-14 06:10:42	73%	0%	NA	0	4746	Normal	NA		78.6	28810.2	25160.9	192.1	2019-05-14 05:50:00	2019-05-14 05:50:00	TRUE	0	0.0	No weather	
3	2019-05-14 21:43:00	2019-05-15 06:10:41	86%	96%	NA	0	4007	Normal	NA		60.5	30461.5	28430.8	203.1	2019-05-15 05:50:00	2019-05-15 05:50:00	TRUE	74	0.0	No weather	
4	2019-05-15 23:11:51	2019-05-16 06:13:59	77%	92%	NA	0	6578	Normal	NA		45.2	25327.6	23132.5	168.9	2019-05-16 05:50:00	2019-05-16 05:50:00	TRUE	0	0.0	No weather	
5	2019-05-16 23:12:13	2019-05-17 06:20:32	78%	94%	NA	0	4913	Normal	NA		44.6	25698.4	22614.6	171.3	2019-05-17 05:50:00	2019-05-17 05:50:00	TRUE	188	0.0	No weather	
6	2019-05-19 01:25:12	2019-05-19 08:43:11	72%	80%	NA	0	4020	No alarm	NA		58.0	26278.2	20759.8	175.2	NA	NA	TRUE	0	0.0	No weather	
7	2019-05-20 22:41:13	2019-05-21 06:22:58	73%	58%	NA	0	5133	Normal	NA		64.0	27705.2	24565.2	184.7	2019-05-21 05:50:00	2019-05-21 05:50:00	TRUE	0	0.0	No weather	
8	2019-05-21 22:39:27	2019-05-22 06:00:55	78%	77%	NA	0	4927	Normal	NA		51.0	26488.8	22780.4	176.6	2019-05-22 05:50:00	2019-05-22 05:50:00	TRUE	279	0.0	No weather	
9	2019-05-22 22:36:59	2019-05-23 06:03:59	84%	98%	NA	0	6637	Normal	NA		43.5	26819.9	25925.9	178.8	2019-05-23 05:50:00	2019-05-23 05:50:00	TRUE	0	0.0	No weather	
10	2019-05-23 23:15:23	2019-05-24 06:33:47	88%	95%	NA	0	4298	Normal	NA		31.6	26304.2	23761.4	175.4	2019-05-24 06:35:00	2019-05-24 06:35:00	TRUE	0	0.0	No weather	
11	2019-05-26 00:21:00	2019-05-26 08:19:28	71%	83%	NA	0	4741	No alarm	NA		69.0	28708.4	26603.1	191.4	NA	NA	TRUE	516	0.0	No weather	
12	2019-05-26 22:49:24	2019-05-27 06:09:41	73%	73%	NA	0	6110	Normal	NA		56.5	26417.5	22190.7	176.1	2019-05-27 05:50:00	2019-05-27 05:50:00	TRUE	617	0.0	No weather	
13	2019-05-27 22:31:45	2019-05-28 05:43:07	74%	82%	NA	0	3999	Normal	NA		52.4	25882.0	23466.4	172.5	2019-05-28 05:35:00	2019-05-28 05:35:00	TRUE	0	0.0	No weather	
14	2019-05-28 22:57:37	2019-05-29 06:01:20	75%	86%	NA	0	5811	Normal	NA		49.1	25422.7	22711.0	169.5	2019-05-29 05:50:00	2019-05-29 05:50:00	TRUE	0	0.0	No weather	
15	2019-05-29 22:24:24	2019-05-30 06:10:41	92%	94%	NA	0	5256	Normal	NA		34.9	27976.8	26111.6	186.5	2019-05-30 05:50:00	2019-05-30 05:50:00	TRUE	243	0.0	No weather	
16	2019-05-30 21:58:16	2019-05-31 06:12:03	73%	95%	NA	0	5286	Normal	NA		68.2	29627.1	26664.3	1185.1	2019-05-31 05:50:00	2019-05-31 05:50:00	TRUE	239	0.0	No weather	
17	2019-05-31 23:28:30	2019-06-01 06:39:58	85%	91%	NA	0	4188	Normal	NA	Nelson	35.0	25888.3	23299.5	172.6	2019-06-01 06:30:00	2019-06-01 06:30:00	TRUE	0	0.0	No weather	
18	2019-06-01 22:27:26	2019-06-02 07:02:34	97%	89%	NA	0	11111	No alarm	NA		40.3	30908.4	28538.8	206.1	NA	NA	TRUE	0	0.0	No weather	
19	2019-06-02 22:34:04	2019-06-03 06:11:11	75%	88%	NA	0	11651	Normal	NA	Nelson	56.3	27427.0	24501.4	182.8	2019-06-03 05:50:00	2019-06-03 05:50:00	TRUE	0	11.6	Sunny	
20	2019-06-03 22:45:51	2019-06-04 06:11:28	84%	93%	NA	0	7521	Normal	NA		39.7	26737.4	24865.8	178.2	2019-06-04 05:50:00	2019-06-04 05:50:00	TRUE	0	0.0	No weather	

Ilustración 1. Dataframe original

En concreto, la traducción a *dataframe* se realiza en el paso 3 del script en R y la declaración de variables en el paso 4, puesto que el paso 1 es importar las librerías necesarias (e instalar los paquetes si es necesario) y en el paso 2 se codifican todas las funciones que se utilizan en el script.

Tras finalizar estos 4 primeros pasos, podemos empezar con el análisis de las variables basado en la estadística descriptiva.

3. Estadística descriptiva

La estadística descriptiva tiene como objetivo, como su propio nombre indica, describir un conjunto de datos de manera cuantitativa, calculando así parámetros básicos que proporcionen información útil y rápida del comportamiento de una muestra, y permitiendo representarla gráficamente.

Para nuestro estudio sobre los hábitos del sueño, se han analizado en primera instancia las variables HC, HS y TA de manera independiente, con el objetivo de entender cómo se comportan cada una de ellas por separado, y contar con más información en el momento de compararlas mediante una regresión.

Así pues, se exponen a continuación los resultados del análisis descriptivo de las horas que pasa el individuo en la cama, es decir, HC.

3.1. Horas en la cama (HC)

TABLA DE FRECUENCIAS

En la imagen 2 podemos ver la tabla de frecuencias de HC. Para calcularla, el primer paso ha sido dividir los valores HC en intervalos, para lo cual se ha empleado la regla de "Sturges", que ha generado 13 intervalos de igual tamaño, 1 hora, que se ajustan bien al tamaño de la muestra (n = 921). Se han tomado los intervalos cerrados por la izquierda porque se ha comprobado que existían valores de HC iguales a 0 (y no existía ningún valor que alcanzase el 13).

	HC	f.a.	f.a.a.	f.r.	f.r.a.
	<chr>	<int>	<int>	<dbl>	<dbl>
1	[0, 1)	4	4	0.00434	0.00434
2	[1, 2)	2	6	0.00217	0.00651
3	[2, 3)	2	8	0.00217	0.00869
4	[3, 4)	0	8	0	0.00869
5	[4, 5)	2	10	0.00217	0.0109
6	[5, 6)	18	28	0.0195	0.0304
7	[6, 7)	172	200	0.187	0.217
8	[7, 8)	414	614	0.450	0.667
9	[8, 9)	232	846	0.252	0.919
10	[9, 10)	64	910	0.0695	0.988
11	[10, 11)	9	919	0.00977	0.998
12	[11, 12)	1	920	0.00109	0.999
13	[12, 13)	1	921	0.00109	1

Ilustración 2. Tabla frecuencias HC

Sólo echando un vistazo rápido a la tabla podemos sacar un par de conclusiones. Parece evidente que la gran mayoría de los datos se agrupan en torno a los intervalos [6,7), [7,8) y [8,9), y que en el resto de intervalos los valores son prácticamente puntuales o, incluso, atípicos.

Podemos comprobarlo con las frecuencias acumuladas, tanto las absolutas como las relativas, que acumulan muy pocos datos hasta que, de repente, tienen una subida bastante notable.

MEDIDAS DE POSICIÓN

Una vez que hemos obtenido la tabla de frecuencias, el siguiente paso es calcular las medidas de posición, con el objetivo de resumir los datos con un solo valor y para visualizar cómo están distribuidos. Seguidamente se presentan los resultados:

- $Media = \bar{x} \approx 7.64$
- $Mediana \approx 7.62$
- $Intervalo\ modal = [7,8)$
- $Percentil\ 25\% = Q_{0.25} \approx 7.08$
- $Percentil\ 75\% = Q_{0.75} \approx 8.20$

Como vemos, la media, la mediana y el intervalo modal están en torno al mismo valor, por lo que parece que la media es bastante representativa de lo que ocurre en la muestra. Además, por el valor

de los percentiles, se puede deducir que la muestra tiene una leve asimetría a la izquierda, aunque lo comprobaremos más adelante.

MEDIDAS DE DISPERSIÓN

Dado que las medidas de posición no son suficientes para saber la disposición real de los datos, calculamos entonces las medidas de dispersión, que nos darán un poco más de información sobre la muestra. Se exponen, por tanto, los resultados:

- $Varianza = \sigma_n^2 = 1.183681$
- $Cuasivarianza = s_n^2 = 1.184968$
- $Desviación\ típica = \sigma_n = 1.087971$
- $Cuasidesviación\ típica = s_n = 1.088562$
- $Coeficiente\ de\ variación = cv \approx 0.14$

Así pues, podemos concluir que las horas en la cama no varían apenas con respecto a la media, como podemos comprobar con la varianza y con el coeficiente de variación, que es menor al 30% (0.3) y, de hecho, es bastante cercano al 0. Es decir, la media es representativa del conjunto de datos, tal como habíamos concluido previamente y, por consiguiente, podemos decir que el conjunto de datos es homogéneo.

MEDIDAS DE FORMA

Por último, se han calculado las medidas de forma para comprobar que la muestra es simétrica con respecto a la media, y para conocer su nivel de apuntamiento. El coeficiente de asimetría que se ha obtenido ha sido, redondeando, de -1.359. Es negativo, por lo que es ligeramente asimétrica a la izquierda, tal y como se dedujo en apartados anteriores. A pesar de esto, es bastante próxima a 0, por lo que podría considerarse una muestra prácticamente simétrica. Esto tiene sentido, puesto que existen pocos individuos que permanezcan muchas horas en la cama, y muy pocas. Es decir, las frecuencias en los extremos son muy reducidas, tal y como se vio previamente en la tabla de frecuencias.

Por otro lado, contamos con el coeficiente de Kurtosis para analizar el grado de concentración que presentan los valores alrededor de la media. En este caso, se ha obtenido un valor aproximado de 13.37, que es positivo, por lo que nos encontramos ante lo que se conoce como distribución leptocúrtica. Esto es, presenta un elevado grado de concentración en torno a los valores centrales de la variable, en particular, un grado mayor al de una distribución normal.

GRÁFICOS

Una vez calculados todos los parámetros básicos de la estadística descriptiva, se presentan a continuación los gráficos realizados para poder tener los resultados obtenidos de manera visual.

En la imagen 3 puede verse el histograma correspondiente a esta variable y, superpuesto, el polígono de frecuencias (absolutas). Ahora se puede visualizar sin problema cómo la muestra es simétrica con respecto a la media, y como los datos están bastante centralizados y concentrados, con un bajo nivel de dispersión. Esto también puede observarse en la imagen 4, que representa el polígono de frecuencias relativas acumuladas de la variable. Efectivamente, todos los datos se acumulan en torno a los valores (6,8) de la variable.

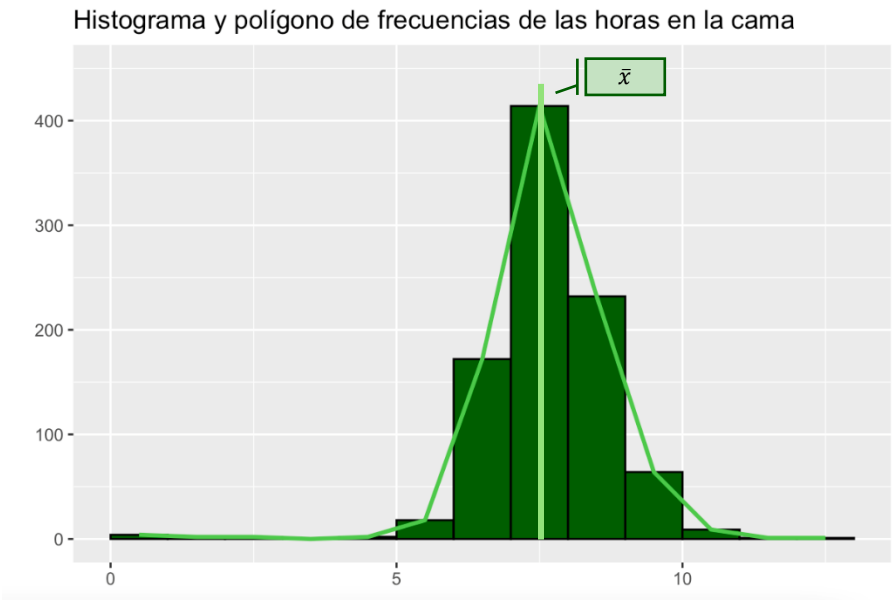


Ilustración 3. Histograma HC

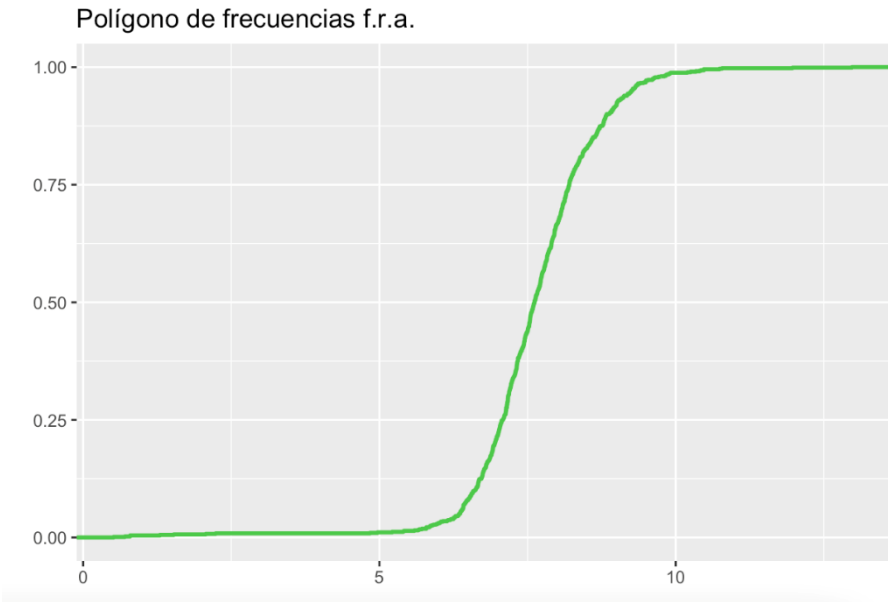


Ilustración 4. Polígono FRA HC

3.2. Horas de sueño (HS)

TABLA DE FRECUENCIAS

En la imagen 5 podemos ver la tabla de frecuencias de HS. Para calcularla, el primer paso ha sido dividir los valores HS en intervalos, para lo cual se ha aplicado el mismo criterio que en HC, ya que los valores de la variable oscilan dentro del mismo rango [0,13).

	HS	f.a.	f.a.a.	f.r.	f.r.a.
	<chr>	<int>	<int>	<dbl>	<dbl>
1	[0, 1)	6	6	0.00651	0.00651
2	[1, 2)	1	7	0.00109	0.00760
3	[2, 3)	1	8	0.00109	0.00869
4	[3, 4)	1	9	0.00109	0.00977
5	[4, 5)	32	41	0.0347	0.0445
6	[5, 6)	187	228	0.203	0.248
7	[6, 7)	374	602	0.406	0.654
8	[7, 8)	240	842	0.261	0.914
9	[8, 9)	66	908	0.0717	0.986
10	[9, 10)	11	919	0.0119	0.998
11	[10, 11)	1	920	0.00109	0.999
12	[11, 12)	0	920	0	0.999
13	[12, 13)	1	921	0.00109	1

Ilustración 5. Tabla frecuencias HS

Para esta tabla ocurre lo mismo que con la anterior. Los datos se agrupan en torno a los intervalos centrales, [5,6), [6,7) y [7,8), y el resto de intervalos presentan una frecuencia muy reducida. Eso sí, parece que estos "intervalos modales" son menores que para las horas en la cama, lo cual indica ya una posible relación entre ambas variables. Parece que los individuos duermen menos horas de las que están en la cama, lo cual tiene sentido, evidentemente. En el resto de columnas el comportamiento es muy similar a las de HC, por lo que no queda información relevante que comentar sobre la tabla.

MEDIDAS DE POSICIÓN

Siguiendo los mismos pasos que con HC, se prosigue con el cálculo de las medidas de posición para resumir el contenido de los datos del fichero, listándose a continuación:

- $Media = \bar{x} \approx 6.642$
- $Mediana \approx 6.644$
- $Intervalo\ modal = [6,7)$
- $Percentil\ 25\% = Q_{0.25} \approx 6.01$
- $Percentil\ 75\% = Q_{0.75} \approx 7.29$

Efectivamente, el intervalo modal es, en esta ocasión, el [6,7), en lugar del [7,8). Además, se puede apreciar una cercanía incluso mayor entre la media y la mediana, por lo que podemos deducir que esta media es muy representativa de la realidad de la muestra. Ya vemos a partir de esta muestra que la media de horas de sueño es inferior a la que recomiendan los médicos, algo que nos va acercando cada vez más al objetivo de nuestro estudio.

MEDIDAS DE DISPERSIÓN

En cuanto a las medidas de dispersión, los resultados han sido los que siguen:

- $Varianza = \sigma_n^2 = 1.275764$
- $Cuasivarianza = s_n^2 = 1.27715$
- $Desviación\ típica = \sigma_n = 1.129497$
- $Cuasidesviación\ típica = s_n = 1.130111$
- $Coeficiente\ de\ variación = cv \approx 0.17$

En este caso, observamos que los datos de las horas de sueño varían más que para las horas en la cama, aunque no con una gran diferencia ($cv(HC) = 0.14$, $cv(HS) = 0.17$). Por ello, de nuevo concluimos que las medidas de posición representan con fidelidad los datos, y que estos tienen un carácter homogéneo.

MEDIDAS DE FORMA

Del mismo modo que con la variable anterior, se han calculado las medidas de forma. En esta ocasión, el coeficiente de asimetría es, aproximadamente, de -0.92. Es decir, nos encontramos ante una muestra incluso más simétrica que la anterior. Por otro lado, el coeficiente de Kurtosis es de 9.82, menor por tanto que el de HC. No obstante, sigue siendo positivo, por lo que el grado de concentración de los datos es elevado con respecto a una distribución normal.

GRÁFICOS

Por último, se presentan los gráficos correspondientes a las horas de sueño efectivo, que se tendrán en cuenta en el resto de las secciones, dado que es la variable principal de nuestro estudio.

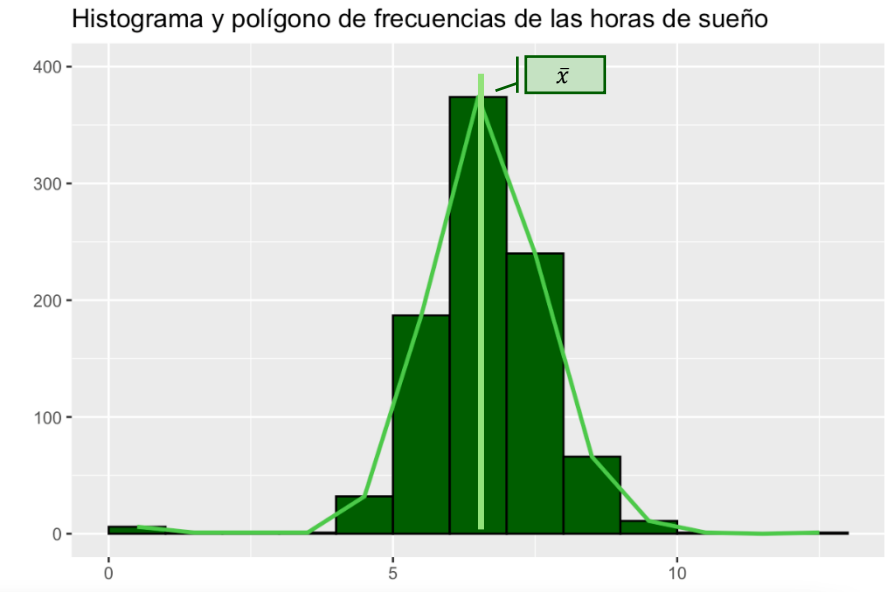


Ilustración 6. Histograma HS

En la imagen 6 se muestra el histograma con el polígono de frecuencias (absolutas), en el que se observa que, efectivamente, la muestra parece ser más simétrica con respecto a su media. Además, dado que el nivel de apuntamiento es menor, se puede apreciar cómo el gráfico se asemeja más al de una distribución normal.

Del mismo modo, en la imagen 7 se adjunta el polígono de frecuencias relativas acumuladas, muy similar al de HC.

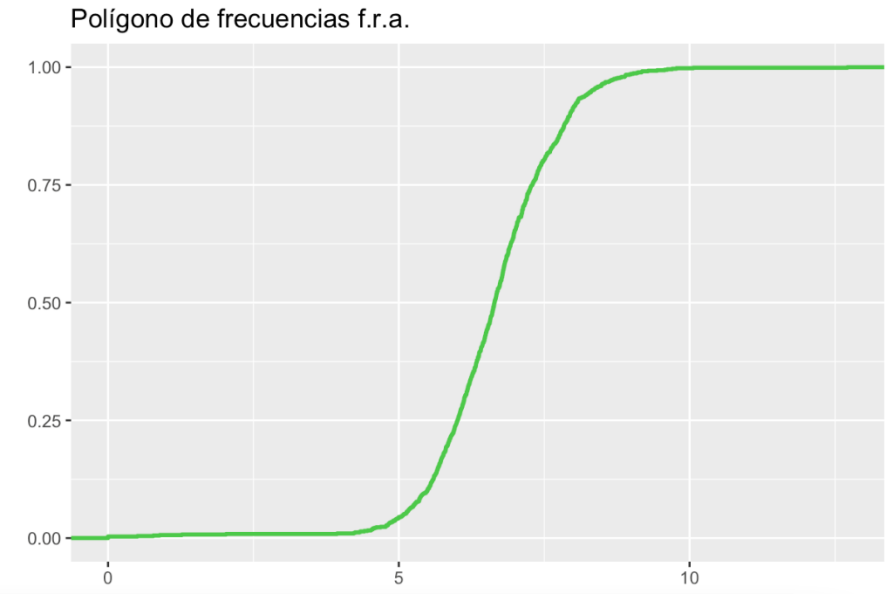


Ilustración 7. Polígono FRA HS

3.3. Tiempo antes de dormir (TA)

TABLA DE FRECUENCIAS

Para esta variable se calculará todo de manera análoga. La tabla de frecuencias del tiempo antes de dormir se puede ver en la imagen 8 y, en este caso, los intervalos son ligeramente más complejos que en los casos anteriores. Se aplicó en una primera instancia la regla de "Sturges", que dividió los datos en 10 intervalos, de un tamaño de 10 minutos cada uno. Sin embargo, se hicieron pequeños ajustes

para contar con unos intervalos más representativos de la muestra. Así, se subdividió el primer intervalo, que tenía una frecuencia demasiado elevada, y se agruparon los últimos 3 intervalos en uno solo, debido a su frecuencia excesivamente reducida.

	TA	f.a.	f.a.a.	f.r.	f.r.a.
	<chr>	<int>	<int>	<dbl>	<dbl>
1	[0, 5)	213	213	0.231	0.231
2	[5, 10)	333	546	0.362	0.593
3	[10, 20)	129	675	0.140	0.733
4	[20, 30)	78	753	0.0847	0.818
5	[30, 40)	54	807	0.0586	0.876
6	[40, 50)	39	846	0.0423	0.919
7	[50, 60)	31	877	0.0337	0.952
8	[60, 70)	33	910	0.0358	0.988
9	[70, 100)	11	921	0.0119	1

Ilustración 8. Tabla frecuencias TA

Esta variable, de acuerdo con los valores de la tabla, parece seguir una distribución distinta a las dos anteriores, ya que tiene un "centro de gravedad" en los primeros intervalos, y a partir de él las frecuencias decrecen rápidamente conforme avanzan los intervalos. Esto tiene su sentido, ya que, intuitivamente, no demasiados individuos tardan más de media hora en quedarse dormidos, a no ser que cuenten con problemas de insomnio, u otras circunstancias puntuales. Respecto a las frecuencias acumuladas, podemos apreciar un incremento rápido al principio, como consecuencia de lo que acaba de mencionarse, el cual quedará reflejado en el gráfico correspondiente, como veremos más adelante.

MEDIDAS DE POSICIÓN

Veamos ahora qué ocurre con las medidas de posición de la variable TA:

- $Media = \bar{x} \approx 16.69$
- $Mediana = 9.395$
- $Intervalo\ modal = [5,10)$
- $Percentil\ 25\% = Q_{0.25} \approx 7.50$
- $Percentil\ 75\% = Q_{0.75} = 21.45$

En este caso, la media del tiempo en quedarse dormido está en torno a los 15 minutos, tal y como se ha investigado en otras fuentes. Eso sí, no parece excesivamente representativa la media teniendo en cuenta lo que difiere con la mediana, y que ni siquiera está incluida en el intervalo modal.

MEDIDAS DE DISPERSIÓN

Las medidas de dispersión obtenidas para esta variable han sido:

- $Varianza = \sigma_n^2 = 304.0884$
- $Cuasivarianza = s_n^2 = 304.4189$
- $Desviación\ típica = \sigma_n = 17.43813$
- $Cuasidesviación\ típica = s_n = 17.4476$
- $Coeficiente\ de\ variación = cv \approx 1.04$

Parece que se cumple lo que acabamos de comentar. Los datos varían mucho más para los minutos antes de quedarse dormido, y se puede ver reflejado en el coeficiente de variación, o en la misma variación. Concluimos entonces que los datos en este caso tienden a ser más heterogéneos, aunque puede deberse a la existencia de valores extremos (cercanos a 100) que hayan provocado este desplazamiento de la media.

MEDIDAS DE FORMA

Veamos, por último, en qué medida es simétrica la muestra de TA, y cómo están concentrados los datos. El coeficiente de asimetría obtenido es de 1.72, por lo que, al ser positivo, presenta una cierta asimetría a la derecha. Junto a esto, se ha comprobado que el coeficiente de Kurtosis es 5.16, lo que indica una concentración de los datos más parecida a la de una distribución normal, pero que sigue considerándose como leptocúrtica.

GRÁFICOS

Una vez hecho esto, se muestra el histograma de TA, en la imagen 9, con su correspondiente polígono de frecuencias, y el polígono de frecuencias relativas acumuladas, en la imagen 10.

Tal y como se ha comentado, este gráfico es radicalmente distinto al de las dos variables anteriores. Presenta una asimetría a la derecha, y los datos no se encuentran agrupados en torno a la media. Nótese la diferencia de amplitud en los intervalos que no tienen la misma longitud, y que por tanto aparecen representados como barras de una anchura mayor.

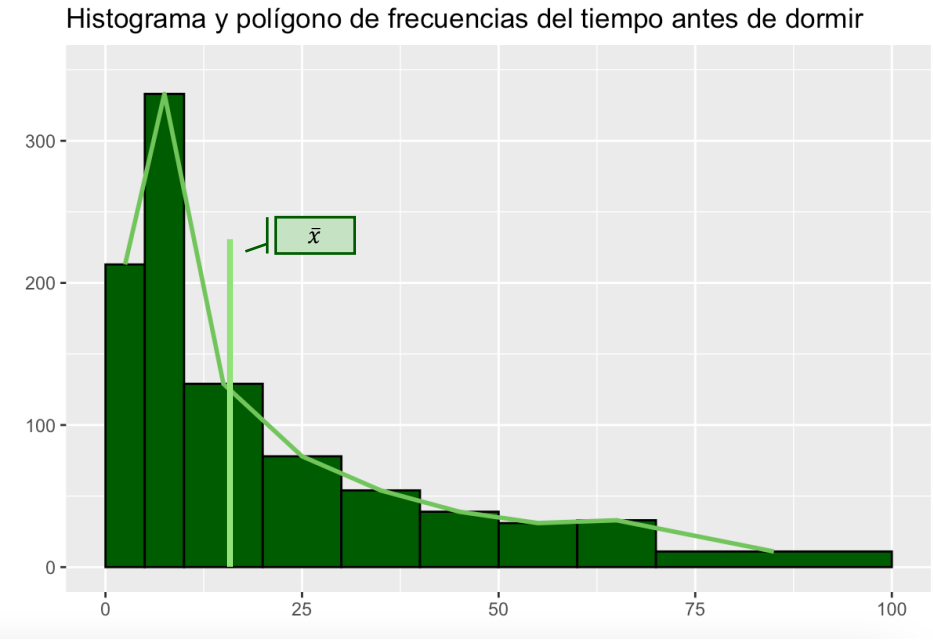
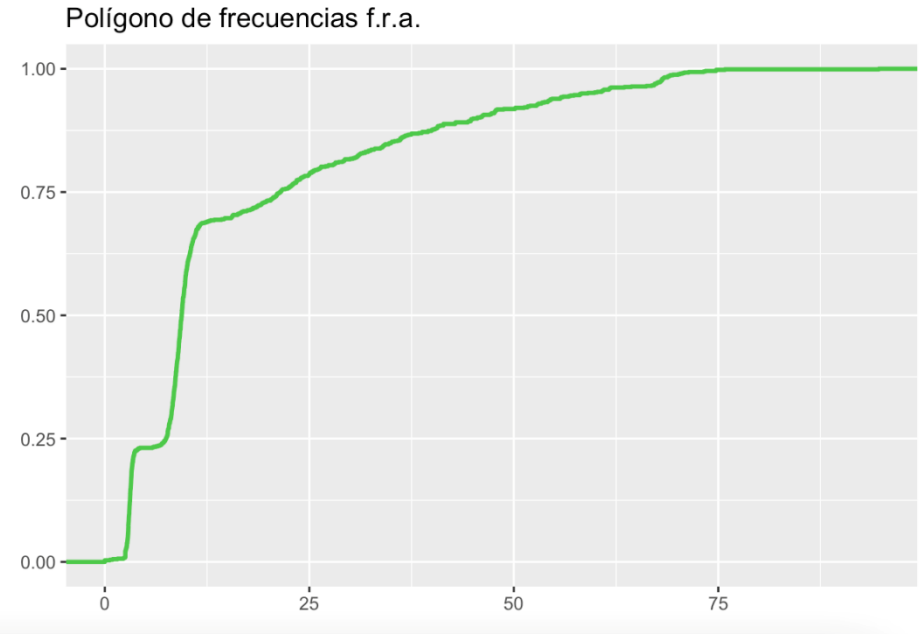


Ilustración 9. Histograma TA



En cuanto al polígono de frecuencias, se puede comprobar la subida casi vertical mencionada anteriormente que acumula la mayor parte de los datos cuando apenas ha llegado a la media.

Ilustración 10. Polígono FRA TA

3.4. Regresión HC y HS

Ahora que se han analizado las distintas variables por separado, se realizan a continuación los análisis estadísticos correspondientes a 2 variables. En particular, comenzaremos con las horas en la cama y las horas de sueño.

Antes de empezar con los cálculos, es importante mencionar que se prevé una relación de dependencia entre estas variables, aplicando el sentido común. Cuantas más horas pase un individuo acostado, se entiende que habrá dormido más horas. Dicho esto, intentemos probar esta teoría.

TABLA DE FRECUENCIAS

En la imagen 11 se puede ver la tabla de frecuencias absolutas para ambas variables, con sus correspondientes marginales. Comprobamos además que la suma total es igual al tamaño de la muestra, como debe ser. Del mismo modo, en la imagen 12, se presenta la tabla de frecuencias relativas con sus marginales.

	vec.HS												
vec.HC	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	[7,8)	[8,9)	[9,10)	[10,11)	[11,12)	[12,13)	Sum
[0,1)	4	0	0	0	0	0	0	0	0	0	0	0	0	4
[1,2)	1	1	0	0	0	0	0	0	0	0	0	0	0	2
[2,3)	1	0	1	0	0	0	0	0	0	0	0	0	0	2
[3,4)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[4,5)	0	0	0	0	2	0	0	0	0	0	0	0	0	2
[5,6)	0	0	0	1	9	8	0	0	0	0	0	0	0	18
[6,7)	0	0	0	0	19	107	46	0	0	0	0	0	0	172
[7,8)	0	0	0	0	2	68	271	73	0	0	0	0	0	414
[8,9)	0	0	0	0	0	4	56	147	25	0	0	0	0	232
[9,10)	0	0	0	0	0	0	1	20	35	8	0	0	0	64
[10,11)	0	0	0	0	0	0	0	0	6	2	1	0	0	9
[11,12)	0	0	0	0	0	0	0	0	0	1	0	0	0	1
[12,13)	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Sum	6	1	1	1	32	187	374	240	66	11	1	0	1	921

Ilustración 11. Tabla frecuencias absolutas HC y HS

	vec.HS												
vec.HC	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	[7,8)	[8,9)	[9,10)	[10,11)	[11,12)	[12,13)	Sum
[0,1)	0.004343105	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.004343105
[1,2)	0.001085776	0.001085776	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.002171553
[2,3)	0.001085776	0.000000000	0.001085776	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.002171553
[3,4)	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
[4,5)	0.000000000	0.000000000	0.000000000	0.000000000	0.002171553	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.002171553
[5,6)	0.000000000	0.000000000	0.000000000	0.001085776	0.009771987	0.008686211	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.019543974
[6,7)	0.000000000	0.000000000	0.000000000	0.000000000	0.020629750	0.116178067	0.049945711	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.186753529
[7,8)	0.000000000	0.000000000	0.000000000	0.000000000	0.002171553	0.073832790	0.294245385	0.079261672	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.449511401
[8,9)	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.004343105	0.060803474	0.159609121	0.027144408	0.000000000	0.000000000	0.000000000	0.000000000	0.251900109
[9,10)	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.001085776	0.021715527	0.038002172	0.008686211	0.000000000	0.000000000	0.000000000	0.069489685
[10,11)	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.006514658	0.002171553	0.001085776	0.000000000	0.000000000	0.009771987
[11,12)	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.001085776	0.000000000	0.000000000	0.000000000	0.001085776
[12,13)	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.001085776	0.001085776
Sum	0.006514658	0.001085776	0.001085776	0.001085776	0.034744843	0.203040174	0.406080347	0.260586319	0.071661238	0.011943540	0.001085776	0.000000000	0.001085776	1.000000000

Ilustración 12. Tabla frecuencias relativas HC y HS

GRÁFICO DE DISPERSIÓN

Antes de calcular la regresión, se ha generado el gráfico de dispersión, o nube de puntos, que dará una primera impresión de la posible relación entre las dos variables, y puede verse en la imagen 13.

Observando el gráfico vemos que no sería necesario siquiera calcular el coeficiente de correlación lineal para saber que este será muy próximo a 1, ya que los datos muestran una clara dependencia lineal (directa). Así pues, calculamos a continuación la regresión, obteniendo los parámetros a y b que conforman por tanto la siguiente función, que queda representada además en la imagen 14:

$$y = 1.86645 + 0.86926x$$

De hecho, no sólo existe una dependencia entre las variables, sino que la relación es muy cercana a la función $y = x$. Así pues, si quisiéramos predecir el número de horas que duerme una persona que está acostado 10 horas, basta con calcular y(10), como se ha hecho en el script. Esto da como

resultado 10.55, lo cual lógicamente carece de sentido, puesto que es imposible dormir más horas que las que se está acostado. Sin embargo, se ha de tener en cuenta que el modelo de regresión lineal no es perfecto, sino que es el que más se acerca a la realidad, puesto que su finalidad es reducir el error cuadrático medio.

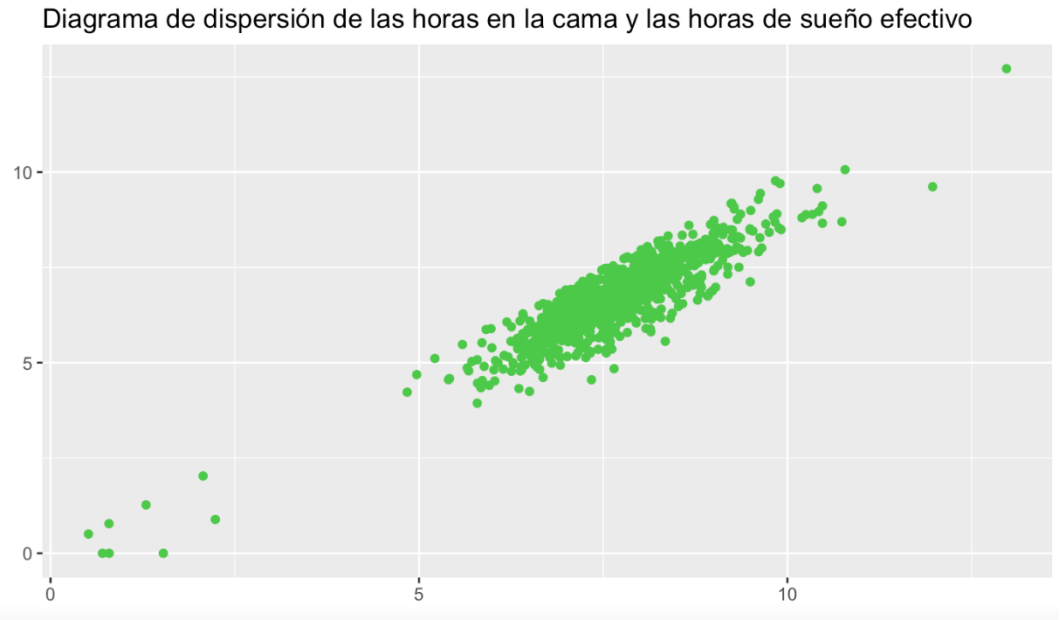


Ilustración 13. Diagrama de dispersión HC y HS

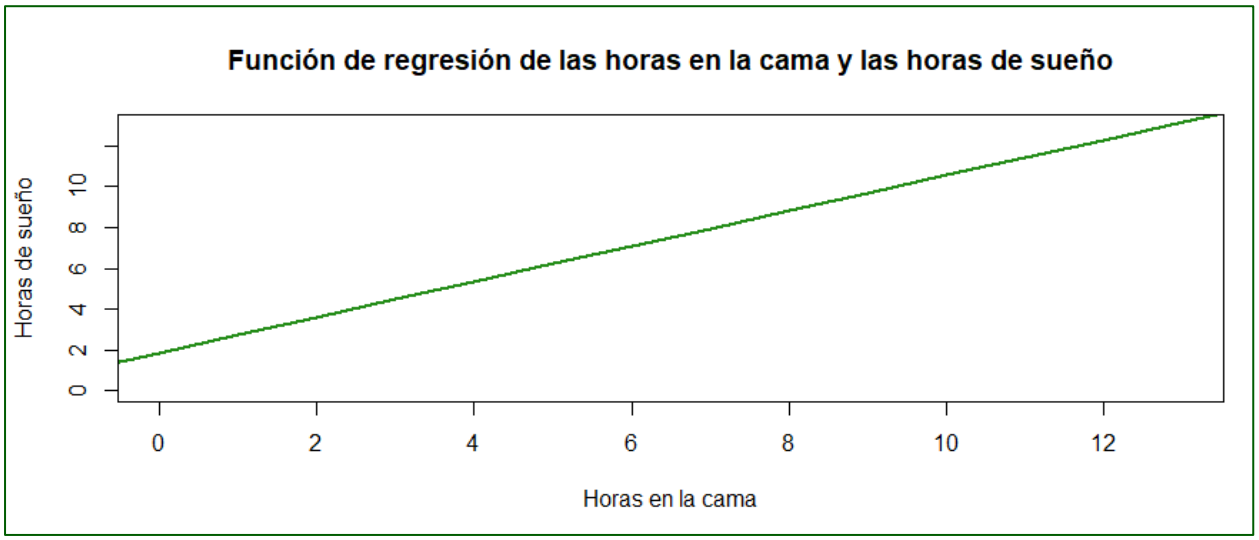


Ilustración 14. Función de regresión HC y HS

OTROS CÁLCULOS

Ahora que ya se ha comprobado con el gráfico y la regresión que las horas en la cama y las horas de sueño están intrínsecamente relacionadas, se han calculado los coeficientes que relacionan ambas variables, obteniendo los siguientes resultados:

- $Covarianza = \sigma_{HC,HS} \approx 1.11$
- $Coeficiente\ de\ correlación\ lineal = \rho_{HC,HS} \approx 0.90$
- $Coeficiente\ de\ determinación = R^2 \approx 0.81$

Tal y como habíamos previsto, el coeficiente de correlación lineal es bastante cercano a 1, por lo que se puede concluir que HC y HS tienen una relación directamente proporcional.

3.5. Regresión HS y TA

Puede ser de utilidad comprobar si existe relación entre el tiempo que tarda un individuo en dormirse y las horas que duerme, para poder sacar conclusiones acerca de las personas que tienen dificultades para conciliar el sueño rápidamente. Por ello, se realizan en este apartado los cálculos necesarios para comprobar si tienen relación. A priori, se piensa que sí podría existir, y que un individuo que se duerma rápido tendrá más probabilidades de dormir un mayor número de horas, pero hemos de ver lo que dicen los datos.

TABLA DE FRECUENCIAS

Al igual que en el apartado anterior, se han generado la tabla de frecuencias absolutas para ambas variables, en la imagen 15, y la tabla de frecuencias relativas, en la imagen 16. Ambas, por supuesto, con sus respectivas marginales.

	vec.TA								
vec.HS	[0,5)	[5,10)	[10,20)	[20,30)	[30,40)	[40,50)	[50,60)	[60,70)	[70,100)	Sum
[0,1)	5	0	0	1	0	0	0	0	0	6
[1,2)	1	0	0	0	0	0	0	0	0	1
[2,3)	1	0	0	0	0	0	0	0	0	1
[3,4)	0	0	0	0	0	0	1	0	0	1
[4,5)	1	15	0	5	3	3	2	3	0	32
[5,6)	16	83	8	18	18	12	12	16	4	187
[6,7)	92	172	27	28	23	14	8	8	2	374
[7,8)	73	62	57	17	9	8	5	5	4	240
[8,9)	20	1	31	8	1	1	2	1	1	66
[9,10)	4	0	4	1	0	1	1	0	0	11
[10,11)	0	0	1	0	0	0	0	0	0	1
[11,12)	0	0	0	0	0	0	0	0	0	0
[12,13)	0	0	1	0	0	0	0	0	0	1
Sum	213	333	129	78	54	39	31	33	11	921

Ilustración 15. Tabla frecuencias absolutas HS y TA

	vec.TA									
vec.HS	[0,5)	[5,10)	[10,20)	[20,30)	[30,40)	[40,50)	[50,60)	[60,70)	[70,100)	Sum
[0,1)	0.005428882	0.000000000	0.000000000	0.001085776	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.006514658
[1,2)	0.001085776	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.001085776
[2,3)	0.001085776	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.001085776
[3,4)	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.001085776	0.000000000	0.000000000	0.001085776
[4,5)	0.001085776	0.016286645	0.000000000	0.005428882	0.003257329	0.003257329	0.002171553	0.003257329	0.000000000	0.034744843
[5,6)	0.017372421	0.090119435	0.008686211	0.019543974	0.019543974	0.013029316	0.013029316	0.017372421	0.004343105	0.203040174
[6,7)	0.099891422	0.186753529	0.029315961	0.030401737	0.024972856	0.015200869	0.008686211	0.008686211	0.002171553	0.406080347
[7,8)	0.079261672	0.067318132	0.061889251	0.018458198	0.009771987	0.008686211	0.005428882	0.005428882	0.004343105	0.260586319
[8,9)	0.021715527	0.001085776	0.033659066	0.008686211	0.001085776	0.001085776	0.002171553	0.001085776	0.001085776	0.071661238
[9,10)	0.004343105	0.000000000	0.004343105	0.001085776	0.000000000	0.001085776	0.001085776	0.000000000	0.000000000	0.011943540
[10,11)	0.000000000	0.000000000	0.001085776	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.001085776
[11,12)	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
[12,13)	0.000000000	0.000000000	0.001085776	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.001085776
Sum	0.231270358	0.361563518	0.140065147	0.084690554	0.058631922	0.042345277	0.033659066	0.035830619	0.011943540	1.000000000

Ilustración 16. Tabla frecuencias relativas HS y TA

GRÁFICO DE DISPERSIÓN

Obtenemos entonces el gráfico de dispersión, presentado en la imagen 17. De la misma manera que en el caso anterior, no es necesario calcular el coeficiente de correlación lineal, pero esta vez porque es evidente que no existe una relación entre estas dos variables, a diferencia de lo que se planteó en un primer momento.

Se ha calculado, en cualquier caso, la función de regresión, tal y como aparece en la imagen 18. Se ha calculado también las horas de sueño que se pueden predecir para un individuo que tarde 15 minutos en dormirse, pero este resultado no es de gran utilidad dado el poco ajuste del modelo de regresión con la realidad. Aún así, puede consultarse en el script de R. La función de regresión es la siguiente:

$$y = 6.747164 - 0.006275x$$

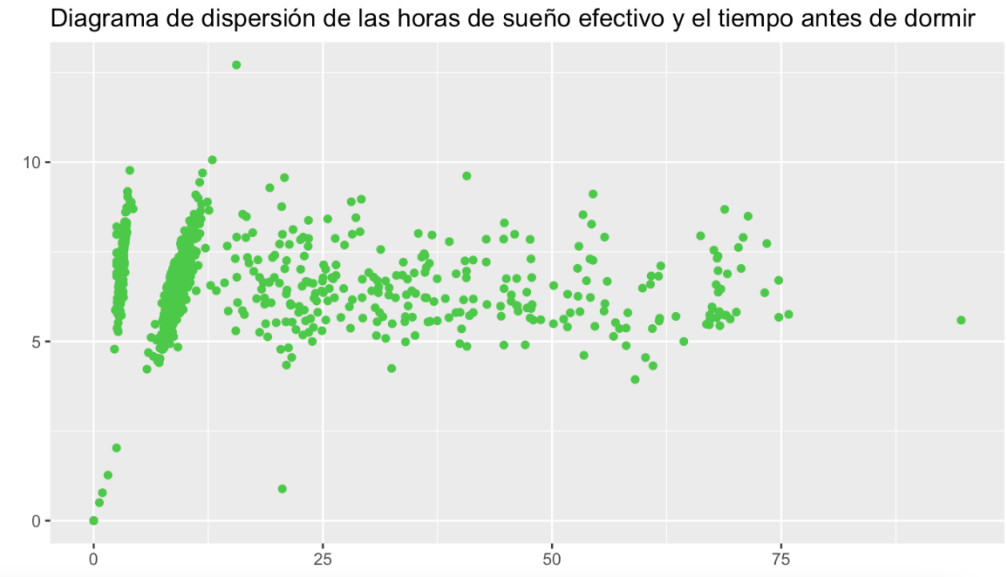


Ilustración 17. Diagrama de dispersión HS y TA

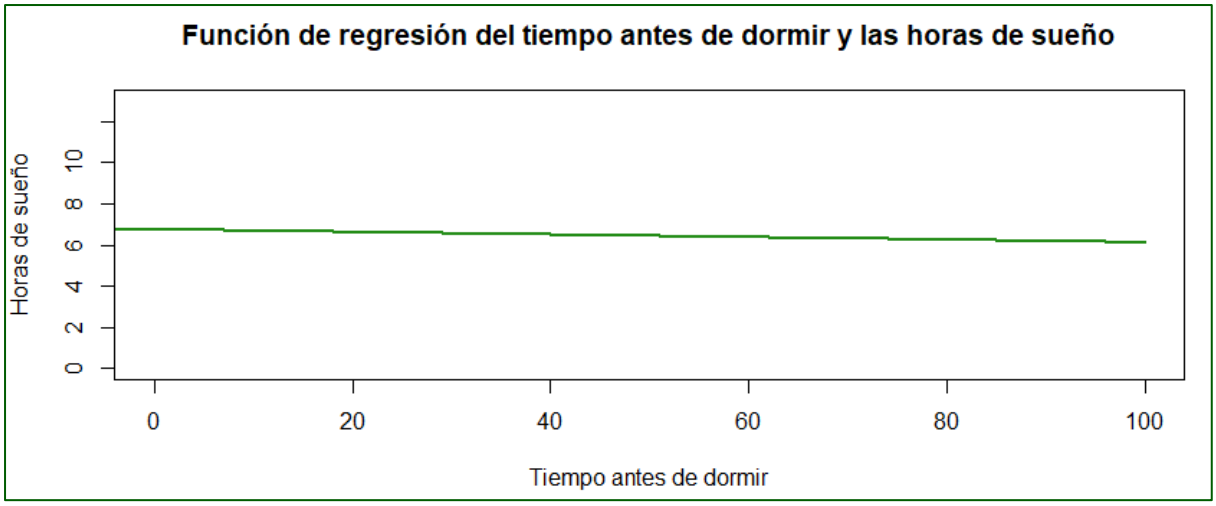


Ilustración 18. Función de regresión HS y TA

OTROS CÁLCULOS

Calculamos pues los coeficientes que relacionan HS y TA:

- $Covarianza = \sigma_{HS,TA} \approx -1.91$
- $Coeficiente\ de\ correlación\ lineal = \rho_{HS,TA} \approx -0.097$
- $Coeficiente\ de\ determinación = R^2 \approx 0.009$

Efectivamente, el coeficiente de correlación lineal es muy próximo a 0 y, evidentemente, el coeficiente de determinación lo es aún más. Así pues, hemos de descartar nuestra hipótesis inicial y concluir que el tiempo antes de dormir y las horas de sueño son variables independientes.

3.6. Regresión TA y CS

En este último apartado de estadística descriptiva comprobaremos si existe una dependencia entre el tiempo antes de dormir y la calidad del sueño. Recordemos antes de nada que la calidad del sueño es un ratio calculado a partir de las horas de sueño en relación con las horas en la cama. Si bien es cierto que, en un primer momento, se presuponía una cierta conexión entre estas variables, tras el resultado obtenido en el apartado anterior han aparecido muchas dudas. Aún así, comprobémoslo.

TABLA DE FRECUENCIAS

Calculamos, en primer lugar, la tabla de frecuencias absolutas de ambas variables (imagen 19) y la de frecuencias relativas (imagen 20). Antes de continuar se ha de mencionar que la variable de CS, una vez calculada, se ha dividido en intervalos mediante la regla de "Sturges", salvo una excepción: se han agrupado los primeros 5 intervalos en uno, puesto que su frecuencia era demasiado reducida.

	vec.CS						
vec.TA	[0,50)	[50,60)	[60,70)	[70,80)	[80,90)	[90,100)	Sum
[0,5)	3	0	0	6	69	135	213
[5,10)	0	0	1	41	196	95	333
[10,20)	0	0	1	9	71	48	129
[20,30)	1	0	0	15	46	16	78
[30,40)	0	0	2	15	34	3	54
[40,50)	0	0	0	7	28	4	39
[50,60)	0	0	2	8	21	0	31
[60,70)	0	0	2	7	24	0	33
[70,100)	0	0	0	2	9	0	11
Sum	4	0	8	110	498	301	921

Ilustración 19. Tabla frecuencias absolutas TA y CS

	vec.CS						
vec.TA	[0,50)	[50,60)	[60,70)	[70,80)	[80,90)	[90,100)	Sum
[0,5)	0.003257329	0.000000000	0.000000000	0.006514658	0.074918567	0.146579805	0.231270358
[5,10)	0.000000000	0.000000000	0.001085776	0.044516830	0.212812161	0.103148751	0.361563518
[10,20)	0.000000000	0.000000000	0.001085776	0.009771987	0.077090119	0.052117264	0.140065147
[20,30)	0.001085776	0.000000000	0.000000000	0.016286645	0.049945711	0.017372421	0.084690554
[30,40)	0.000000000	0.000000000	0.002171553	0.016286645	0.036916395	0.003257329	0.058631922
[40,50)	0.000000000	0.000000000	0.000000000	0.007600434	0.030401737	0.004343105	0.042345277
[50,60)	0.000000000	0.000000000	0.002171553	0.008686211	0.022801303	0.000000000	0.033659066
[60,70)	0.000000000	0.000000000	0.002171553	0.007600434	0.026058632	0.000000000	0.035830619
[70,100)	0.000000000	0.000000000	0.000000000	0.002171553	0.009771987	0.000000000	0.011943540
Sum	0.004343105	0.000000000	0.008686211	0.119435396	0.540716612	0.326818675	1.000000000

Ilustración 20. Tabla frecuencias relativas TA y CS

GRÁFICO DE DISPERSIÓN

El gráfico de dispersión, expuesto en la imagen 21, informa efectivamente de que no existe ningún tipo de dependencia entre las variables en cuestión. Por ello, a pesar de lo que la intuición pueda llevar a pensar, hemos de concluir que el tiempo que un individuo tarda en quedarse dormido no influye positiva ni negativamente en su calidad de sueño.

Se incluye también en la imagen 22 la función de regresión, que a simple vista se puede comprobar que no va a corresponderse con la realidad, basándonos en la nube de puntos. De igual manera que en el apartado anterior, se ha intentado predecir la calidad del sueño si un individuo tarda 15 minutos (la media, aproximadamente) en dormirse, pero el resultado es muy poco fiable.

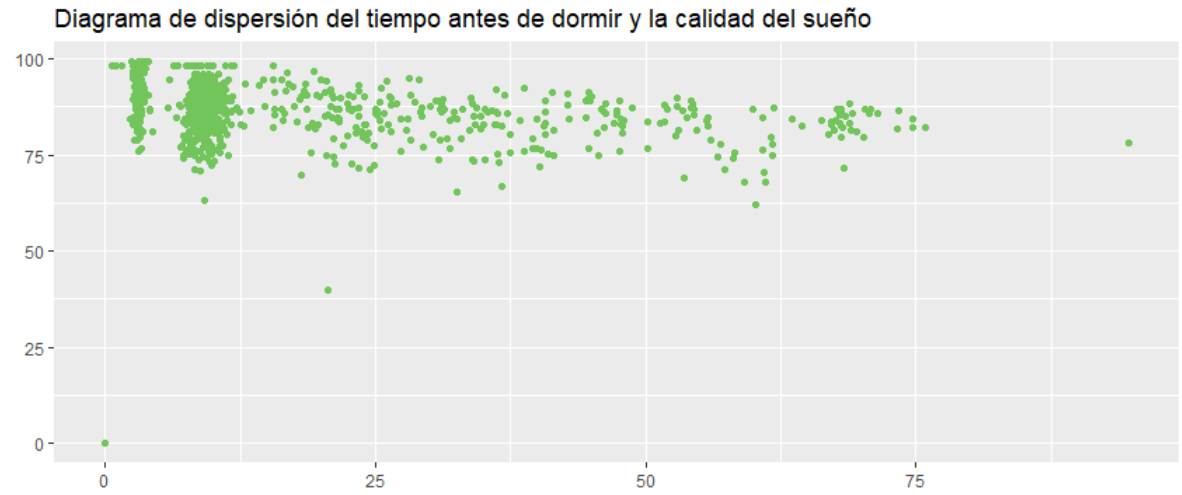


Ilustración 21. Diagrama dispersión TA y CS

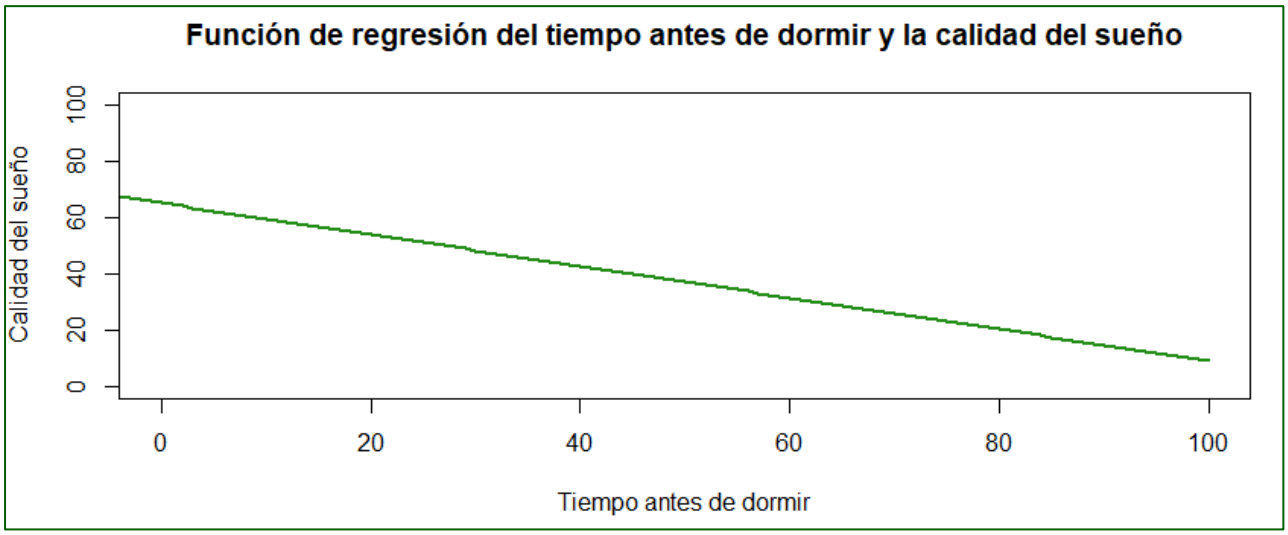


Ilustración 22. Función regresión TA y CS

OTROS CÁLCULOS

Por último, los coeficientes que relacionan estas dos variables han sido los que siguen:

- $Covarianza = \sigma_{TA,CS} \approx -38.13$
- $Coeficiente\ de\ correlación\ lineal = \rho_{TA,CS} \approx -0.26$
- $Coeficiente\ de\ determinación = R^2 \approx 0.07$

Confirmamos, por tanto, que no existe relación entre ellas, ya que, aunque el coeficiente de correlación lineal sea más alejado del 0 que en la regresión anterior, sigue siendo muy cercano a 0.

4. Probabilidad

En esta sección se proponen distintos supuestos en los cuales se han calculado determinadas probabilidades que pueden ser de utilidad para ilustrar el comportamiento de las variables de estudio, especialmente de las horas de sueño.

Se han contemplado varios casos básicos en los que basta aplicar la Regla de Laplace, otros que requieren de la intersección de sucesos, y dos últimos que calculan una probabilidad condicionada. En algunos de estos casos se compara las horas de sueño con '7'. Esto es debido a que en numerosos estudios científicos se concluye que 7 horas es el tiempo de sueño mínimo recomendado, si bien es cierto que para edades tempranas se recomienda un número incluso mayor. Así, se utiliza el mínimo recomendable común para todas las edades para probar uno de los objetivos de este estudio.

CASOS BÁSICOS

- En numerosas ocasiones, a la hora de acostarse, hay gente que tarda muy poco en conciliar el sueño. Por ello, decidimos comprobar la probabilidad de que un individuo tarde menos de 10 minutos en dormirse: $P(TA < 10) \approx 0.59 \rightarrow 59\%$
- En contraposición, pensamos también en aquellas que les cuesta más llegar a las primeras fases del sueño. Así, calculamos la probabilidad de que un individuo tarde más de 15 minutos en dormirse: $P(TA > 15) \approx 0.30 \rightarrow 30\%$
- Además, decidimos obtener el porcentaje de gente que sigue la cantidad de horas de sueño recomendado por la OMS. Es decir, calculamos la probabilidad de que un individuo duerma 7 horas o más: $P(HS \geq 7) \approx 0.78 \rightarrow 78$

INTERSECCIÓN DE SUCESOS

- En lo referente a este tipo de probabilidad nos resultó interesante comprobar la probabilidad de que un individuo duerma menos de 7 horas y tenga una calidad del sueño superior al 60%, lo cual afirmaría en cierta medida las recomendaciones médicas: $P(HS > 7 \cap CS > 60) \approx 0.65 \rightarrow 65\%$. Podría esperarse que fuera menor, pero no es tampoco demasiado elevada.
- También aparece la posible relación entre retrasar mucho la hora de dormir con la calidad del sueño, si bien es cierto que la teoría ha sido desmentida tras realizar el estudio de regresión. Aún con ello, calculamos la probabilidad de que un individuo tarde más de 45 minutos en dormirse y su calidad del sueño sea superior al 70%: $P(TA > 45 \cap CS > 70) \approx 0.098 \rightarrow 9{,}8\%$. Este resultado confirma, en efecto, que es muy poco probable tener una buena calidad de sueño tardando mucho en conciliarlo.

PROBABILIDAD CONDICIONADA

- Por último, se ha considerado relevante averiguar la probabilidad de que un individuo tenga un sueño efectivo 7 horas o más sabiendo que está en la cama 9 horas o más. Dado que el coeficiente de correlación es muy cercano a 1, comprobamos que esta probabilidad viene determinada por la dependencia de las dos variables: $P(HS \geq 7 | HC \geq 9) \approx 0.987 \rightarrow 98{,}7\%$.

- Calculamos finalmente la probabilidad de que un individuo duerma más de 6 horas sabiendo que ha tardado más de media hora en dormirse. Estas variables no cuentan con una dependencia entre sí, como ya se ha comprobado. $P(HS > 6|TA > 30) \approx 0.56 \rightarrow 56\%$. Por el alto porcentaje del resultado confirmamos que dentro de nuestro *dataset* hay bastantes individuos que pueden llegar a dormir las horas recomendadas aun habiendo tenido una peor conciliación del sueño.

5. Modelos de probabilidad

El siguiente paso para continuar con el estudio de las horas de sueño efectivo es calcular sus correspondientes funciones de densidad y distribución, comparando su distribución con una normal, para poder posteriormente calcular distintas probabilidades que se detallarán más adelante.

Así pues, la función de densidad queda representada en la imagen 23. A partir de ella podemos deducir que la variable sigue una distribución normal, dada la característica forma de Campana de Gauss. Sin embargo, al realizar el diagrama de cuantiles, encontramos ciertos valores que se alejan de la recta, tal y como se puede ver en la imagen 24.

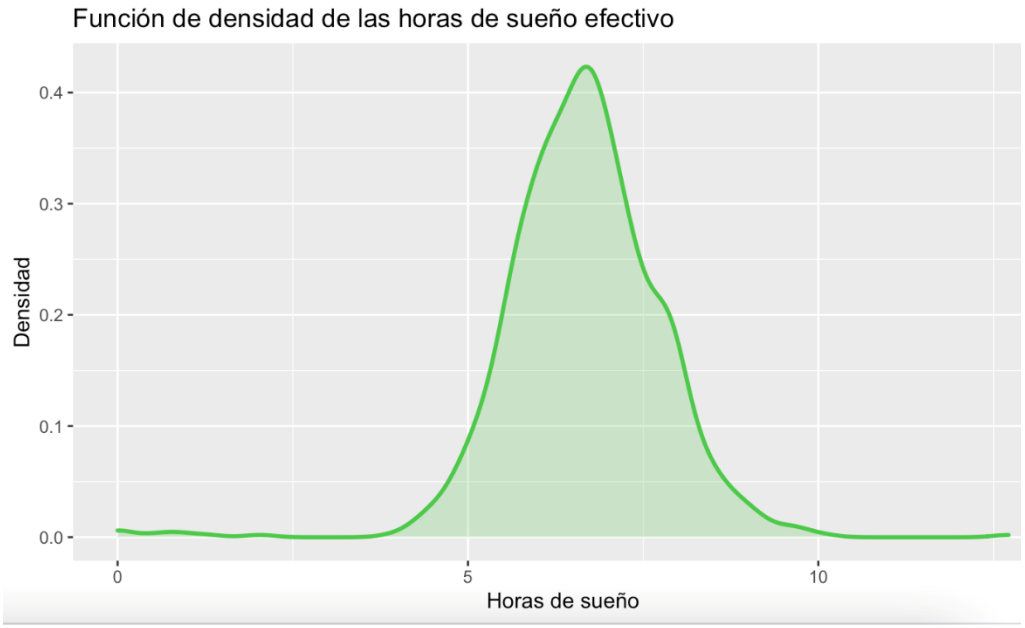


Ilustración 23. Función densidad HS

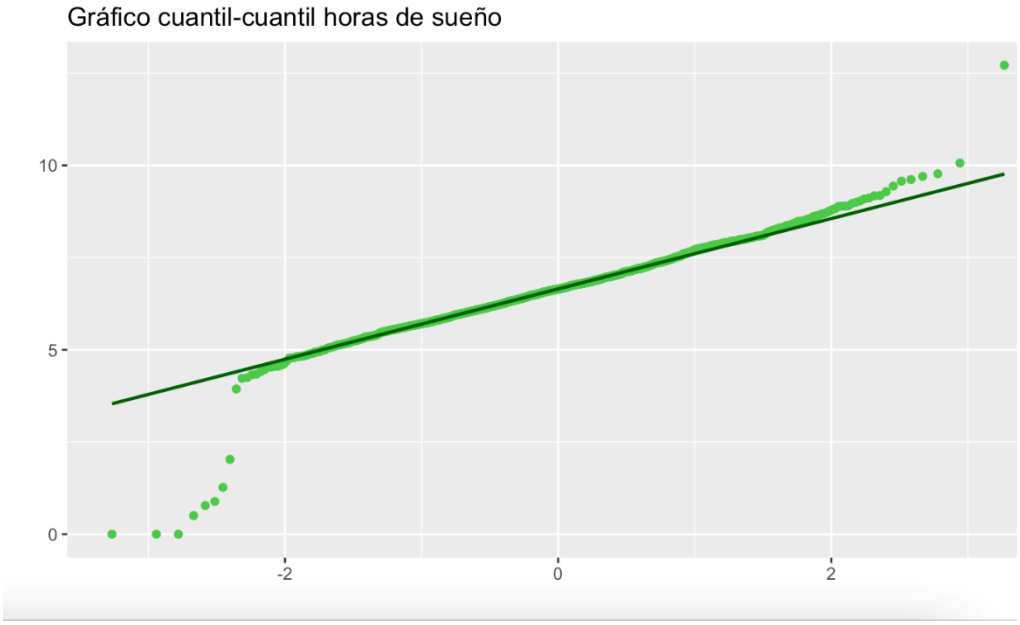


Ilustración 24. Diagrama cuantiles HS

Por ello, se eliminan los valores atípicos que alteran la distribución de HS. Una vez hecho esto, es inmediato ver en la imagen 25 el gráfico mucho más ajustado a la recta de la distribución normal.

De este modo, al contar ahora con una distribución normal, es posible entonces calcular la función de distribución, que permite calcular las probabilidades necesarias, y que puede verse representada en la imagen 26.

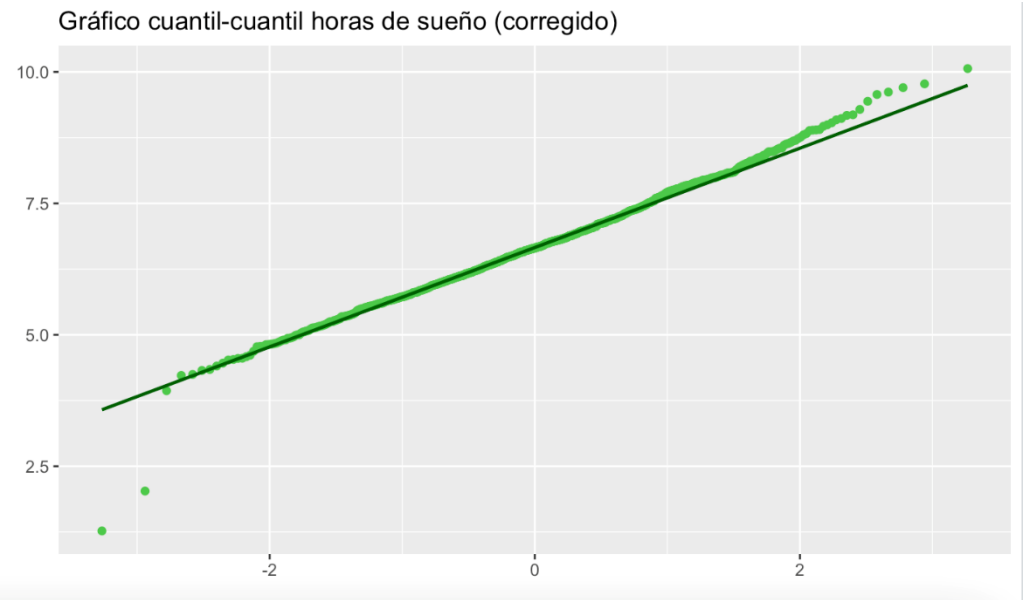


Ilustración 25. Diagrama cuantiles HS (ajustado)

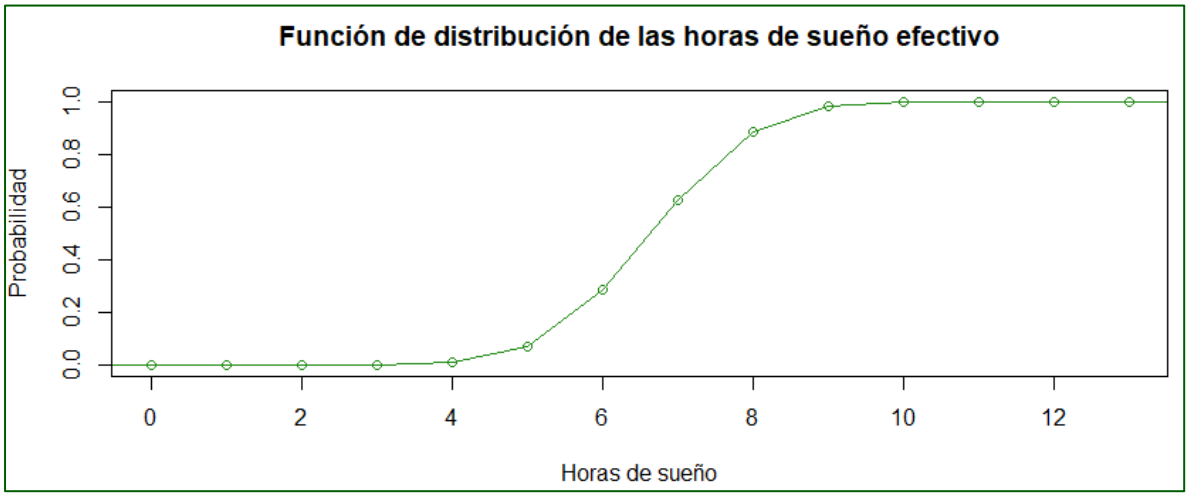


Ilustración 26. Función de distribución HS

Ahora se calcularán, por tanto, probabilidades relevantes para el estudio, empleando la función de distribución.

- $P(HS > 6) \approx 0.72 \rightarrow 72\%$
- $P(6 < HS \leq 8) \approx 0.60 \rightarrow 60\%$
- $P(HS \leq 4) \approx 0.0097 \rightarrow 0{,}97\%$
- $P(HS > 10) \approx 0.0015 \rightarrow 0{,}15\%$
- $\boldsymbol{P(HS > 7) \approx 0.38 \rightarrow 38\%}$

6. Contrastes de hipótesis

Para finalizar el estudio sobre las horas de sueño efectivo se han llevado a cabo dos contrastes de hipótesis para alcanzar el objetivo de este proyecto: demostrar que la población duerme, en media, menos de lo mínimo recomendado.

PRIMER CONTRASTE

Así pues, en el primer contraste, el más relevante, se trata de probar, con una confianza del 95%, que la media poblacional de la variable HS es inferior a 7 que, como se ha explicado anteriormente, es el número mínimo de horas comunes a todas las franjas de edades que es recomendable dormir.

Resumiendo brevemente el contraste, las hipótesis planteadas han sido:

$$\left.\begin{array}{l} H_0: \mu \geq 7 \\ H_1: \mu < 7 \end{array}\right\}$$

Partiendo de una σ desconocida, y aplicando por tanto una T de Student para calcular el estadístico, se ha comprobado en el script que $t \in R.C.$, por lo que hemos de rechazar H_0. Así pues, tal como se quería comprobar, existe suficiente evidencia estadística para concluir que la media de HS es menor que 7. Esto es, podemos afirmar con un 95% de confianza que, en media, se duerme menos de las 7 horas recomendadas para mantener un buen nivel de salud.

SEGUNDO CONTRASTE

En el segundo contraste de hipótesis se comprueba, tal y como se explica en el script de R, que el número de horas de sueño varía más que el número de horas en la cama. El resultado de este contraste, de nuevo, confirma nuestra suposición inicial, de forma que se puede asegurar con un 90% de confianza que $\sigma_{HS} > \sigma_{HC}$.

7. Conclusiones

Debido a problemas con la elección de los ficheros de datos, y cambios de proyecto a última hora, no ha habido tiempo suficiente para analizar con la profundidad requerida los apartados 5, 6 y 7.

Sin embargo, es relevante mencionar que el resultado principal del estudio concuerda con las hipótesis iniciales, que afirmaban que se duerme menos de lo necesario, o lo recomendado. Además, los resultados más destacados acerca de las horas de sueño efectivo son la poca variación de los datos y su homogeneidad, junto a la simetría que presentan y el alto nivel de concentración en torno al valor de la media, 6.64 horas.

8. Bibliografía

A continuación, se adjuntan las fuentes principales de información de las que se ha hecho uso para la elaboración de este estudio.

· Obtención del fichero de datos:
https://www.kaggle.com/datasets/danagerous/sleep-data

· Para el tratamiento de dataframes:
https://swcarpentry.github.io/r-novice-gapminder-es/05-data-structures-part2/

· Para la manipulación de los datos iniciales:
https://www.institutomora.edu.mx/testU/SitePages/martinpaladino/manipulacion_de_datos_con_r_dplyr_y_tidyr.html

· Para las tablas de frecuencia, las medidas de estadística descriptiva, el paquete *ggplot2* y otras cuestiones básicas de R:
https://www.odiolaestadistica.com/estadistica-r/

· Guía de estilo del lenguaje R:
https://www.datanalytics.com/2014/01/27/guia-de-estilo-de-r-de-google/

· Paquetes que proporciona R:
https://www.rstudio.com/products/rpackages/

· Información básica del paquete *ggplot2*:
https://ggplot2.tidyverse.org/

· Manual de R (fuente más utilizada):
https://fhernanb.github.io/Manual-de-R/

· Manual de R en inglés especializado en gráficos:
https://r-graphics.org/

· Web principal del paquete *ggplot2*, donde se han investigado todos los gráficos realizados:
https://r-charts.com/es/ggplot2/

· Para la información médica (y de otros tipos) acerca del sueño:
- https://www.mayoclinic.org/es-es/healthy-lifestyle/adult-health/expert-answers/how-many-hours-of-sleep-are-enough/faq-20057898
- https://www.normodorm.es/insomnio-consecuencias-de-dormir-menos-de-8h.html
- https://www.elmundo.es/ciencia-y-salud/salud/2022/04/11/6253f0a3e4d4d875478b45a0.html
- https://www.cdc.gov/media/releases/2016/p0215-enough-sleep.html
- https://www.elconfidencial.com/tecnologia/ciencia/2019-01-02/horas-sueno-tecnologia-movil_1734030/
- https://www.mundodeportivo.com/elotromundo/bienestar/20170918/431390429481/falta-sueno-truco-dormir-bien-consejos.html

· Otras fuentes:
- https://stackoverflow.com/
- https://rpubs.com/JoanClaverol/488759

- https://r-coder.com/
- https://estadistica-dma.ulpgc.es/cursoR4ULPGC/10-distribProbabilidad.html
- https://torres.epv.uniovi.es/centon/qqplots.html
- https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/normality-test/interpret-the-results/key-results/
- https://aprender-uib.github.io/AprendeR2/
- https://www.ine.es/prodyser/microdatos.htm#:~:text=Los%20ficheros%20de%20microdatos%20contienen,valores%20que%20toma%20cada%20variable
- https://www.ine.es/
- https://cvmdp.ucm.es/moodle/pluginfile.php/2247388/mod_resource/content/1/ManualIntroduccionR.pdf
- https://economipedia.com/definiciones/estadistica-descriptiva.html
- https://es.wikipedia.org/wiki/Regla_de_Sturges
- https://es.wikipedia.org/wiki/Coeficiente_de_variaci%C3%B3n
- https://pablopenalver.com/medidas-de-forma/#:~:text=Las%20medidas%20de%20forma%20son,un%20tipo%20particular%20de%20distribuci%C3%B3n.