Análisis Estadístico Descriptivo: Fumadores, Correlación Entre Edad y Cigarros Fumados al Día, Diferencias de Consumo Entre Hombres y Mujeres

Armando Delfín, Eduardo Longaart, y Sonia Eveligret

Universidad Central De Venezuela

Departamento de Estadística y Probabilidad

Computación I

Jesús Ochoa

12/7/2024

Nota del Autor

Correspondence concerning this article should be addressed to Armando Delfín

Resumen

En este trabajo de investigación se realiza un análisis descriptivo sobre una base de datos que contiene un número de observaciones plasmadas en un registro con 3900 hombres y mujeres fumadores y ex-fumadores, haciendo uso de gráficos, y de estadísticas descriptivas para obtener información con la cual posteriormente hacer un estudio que nos permita conocer si existe una relacion entre la edad de las personas, su sexo y el numero de cigarros que consumen por dia. Lo que sea desea tener como resultado de este trabajo de investigación, es una estimación básica de lo anterior mencionado partiendo desde todos lo registros que contengan información de las personas que son fumadores regulares, y de forma práctica, a través del uso de graficos y medidas descriptivas, encontrar caracteristicas asociadas a quellos que consuman mayores cantidades de cigarros diariamente.

Palabras Claves: Fumadores, Características, Descriptivas

Análisis Estadístico Descriptivo: Fumadores, Correlación Entre Edad y Cigarros Fumados al Día, Diferencias de Consumo Entre Hombres y Mujeres

¿De que manera se han manejado los datos suministrados?

El analisis exploratorio de datos definido por Jonh W. Tukey, el cual es el tratamiento estadístico al que se someten las muestras recogidas durante un proceso de investigación en cualquier campo científico. Antes de manejar la base de datos, conviene analizar los datos que se utilizaran. Esto permite observar las caracteristicas fundamentales de los mismos y comprender la estructura del conjuto de los datos, identificando la variable que se tiene como objetivo y explorando las posibles técnicas de modelado. (Tukey, J.W, 1977)

En este trabajo de investigacion se extrajo toda la informacion necesaria de la base de datos de la siguiente manera:

Lo primero que se llevó a cabo fue observar como se comportaban los datos y si dentro de la base de datos habian registros con valores nulos o valores atípicos. En el caso de la columna que contiene a la variable, cigarros consumidos al día, se encontraban 14 registros con valores nulos, y más de 40 datos atípicos, por encima de 35 cigarros fumados diariamente, valor que fue definido como el límite superior de la distribución por medio de gráficos de caja y bigote.

Una vez explorada la base de datos lo siguiente a realizar fue la depuración de esta, empleando el método de remplazar dichos valores nulos por la media aritmética de la variable en estudio, y para los valores atípicos por encima del límite superior, 35, contarlos como registros de personas que fuman exactamente 35 cigarros al día.

Luego de hacer este proceso de depuración, se estudiaron a las variables sexo, y edad, para determinar que tanto abarcaban esta dos variables dentro de la base de datos, ordenándolas de forma ascendente para hacer mas fácil el manejo de dichas variables dentro de la base de datos.

Al momento de finalizar todo el proceso de filtrado, depuración y ordenamiento de los registros, el siguiente paso fue realizar gráficos que facilitaran las interpretaciones de los datos y asi poder transmitir de forma clara y concisa, todos los resultados obtenidos de la investigación.

Análisis Descriptivo

A continuacion, en las siguientes tablas se muestran estadísticas descriptivas para las variables en estudio, Toda la informacion respecto a los datos a manejar fue extraída de https://www.kaggle.com/datasets/jacepranter/smoker-health-data, de la cual se estudia solo a los 1932 fumadores activos de entre los 3900 registros que incluyen ex-fumadores.

- En promedio, los fumadores consumen 18 cigarros diariamente, con una desviación estándar de 9 cigarros. Menos del 25% de ellos fuma más de 20 cigarros (un paquete) por día. La distribución es asimétrica positiva, y leptocúrtica.
- En promedio, los fumadores tienen 48 años de edad, con una desviación estándar de 8 años. El 75% de ellos tiene 53 años o menos. La distribución es asimétrica positiva , y leptocúrtica.
- De todas las personas en observación, 1105 son de sexo masculino, y 827 de sexo femenino. Representando el 57% y 43% respectivamente.

Tabla 1

Estadística Descriptiva Sobre las Variables Cuantitativas

Variable	Promedio	D. Estandar	Q1	Q2	Q3	C.S	Kurtosis
C.D	17.9	9.43	10	20	20	0.07	2.30
EDAD	47.7	7.97	41	46	53	0.51	2.40

Tabla 2

Frecuencias Relativas de los Fumadores Respecto a su Sexo

Sexo	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Relativa Porcentual
M	1105	0.57	57%

Sexo	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Relativa Porcentual
F	827	0.43	43%
TOTAL	1932	1	1

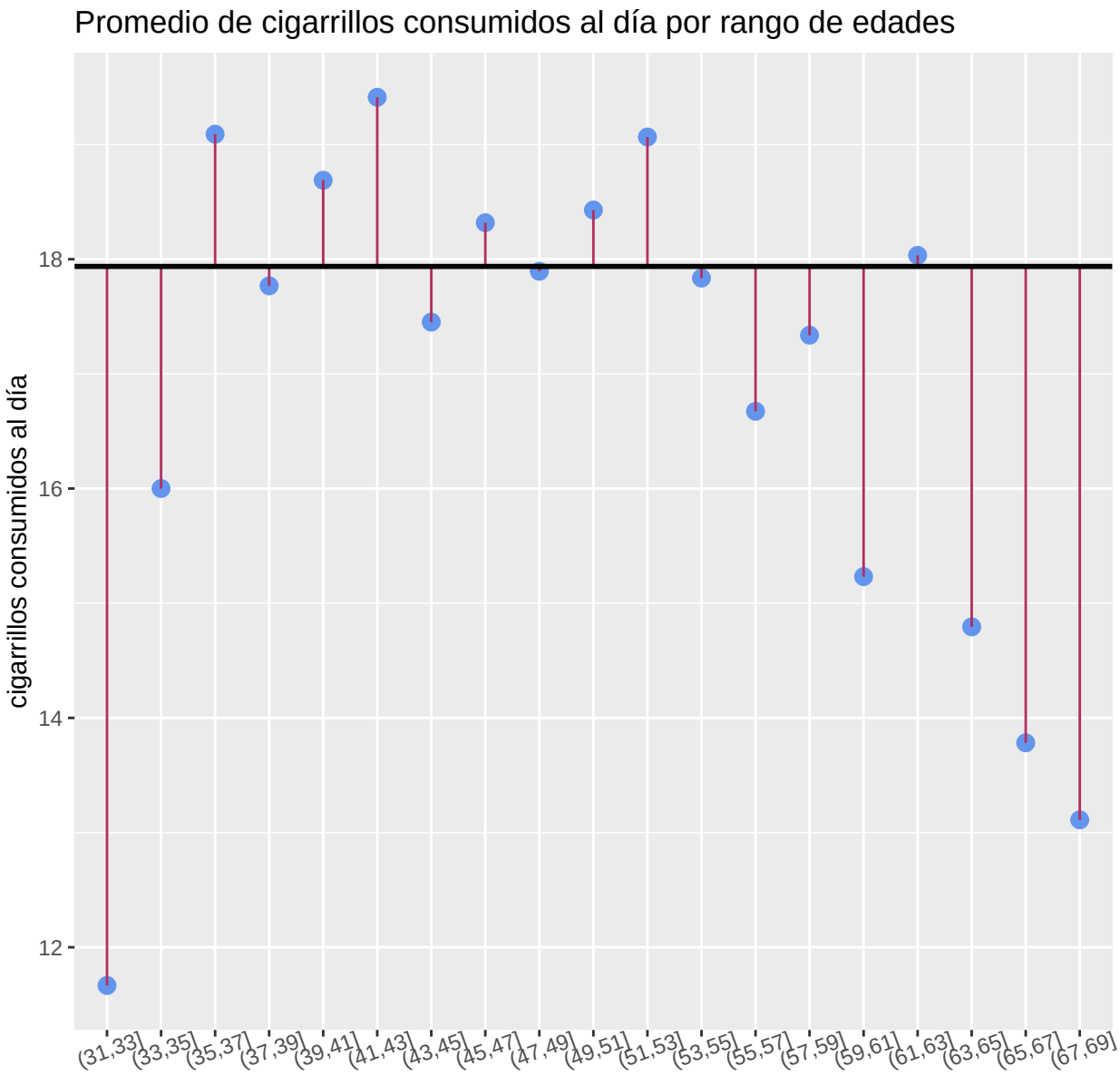
Análisis Bivariante: ¿Existe Relacion Entre la Edad de los Fumadores y la Cantidad de Cigarros que Consumen al día?

A traves del análisis descriptivo bivariante (coeficiente de correlacion de Pearson) y el uso de graficos se determino que existe correlacion inversa entre el número de cigarros que fuma una persona y su edad

Covarianza	Índice de Correlación
-5.30	-0.07

La covarianza entre las dos variables, edad, y cigarros consumidos al día, tienen un valor de covarianza negativo, lo que significa que tienen una relación negativa o inversa. A su vez, el índice de correlación, el cual nos indica la fuerza de la relación negativa antes descubierta es de -0.07. Teniendo en cuenta que una relación inversa puede llegar a un máximo de -1, ésta correlación parece ser muy tenue.

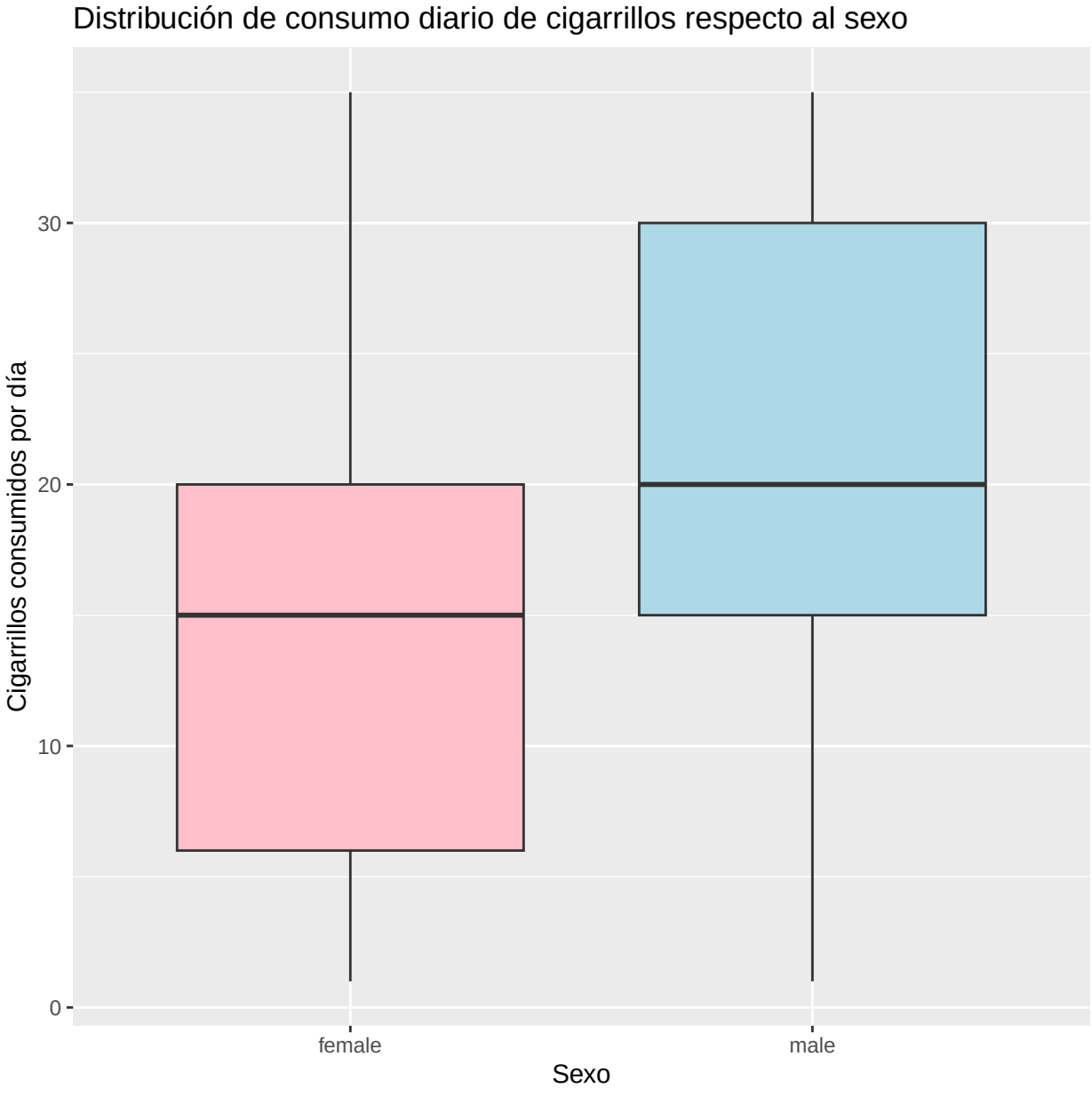
Para observar estas dos variables gráficamente, nos referimos a la siguiente figura



En este gráfico se muestra la correlación lineal, viendo como se distancian las medias aritméticas de cada intervalo de edad respecto a la media general extraída en la primera tabla, 17.9 cigarros de media. Observamos que las medias más altas se encuentran en la parte izuierda del gráfico, es decir que le pertenecen en los intervalos de edades que comprenden a las personas más relativamente jóvenes en estudio. A partir de los 55 años, se puede apreciar un declive lineal en cuanto a los promedios de cigarros fumados al día, lo que se corresponde con las medidas númericas descriptivas de covariaza y correlación negativa halladas previamente.

Cantidad de cigarros que consumen las personas dependiendo de su sexo

De manera visual indicaremos que sexo dentro de la base datos consumen una mayor cantidad de cigarros por día, a través del análisis descriptivo y el gráfico que nos indica que generalmente el sexo más propenso a consumir una mayor cantidad de cigarros al dia son los hombres en observación, comparado a la distribución conformada por las mujeres.



Este gráfico emplea el poderoso ‘five point summary’ (Tukey, 1977) a través de modelos de cajas y bigotes. Podemos observar , que el tercer cuartil de la distribución de las mujeres

fumadoras es igual a la mediana de los hombres; 20 cigarros al día. De esto podemos derivar que menos del 25% de las mujeres fumadoras en observación consumen cantidades mayores a las de un paquete estándar de 20 cigarros, contrastado con que, por el otro lado, el 50% de los hombres observados fuman 20 cigarros o más. Fumar más de un paquete por día perece ser mucho más común entre los hombres fumadores obseravdos.

Conclusión

Dentro de este análisis estadistico descriptivo se han obtenido como resultados:

- Una correlacion lineal inversa en cuanto a la edad de los fumadores y la cantidad cde cigarros que consumen por dia. Esto quiere decir que a medida que la edad de los fumadores observados se incrementaba, en promedio, podíamos ver un ligero decrecimiento en la cantidad de cigarros consumidos al día. Puede ser visto tambien desde la perspectiva opuesta, a medida que la edad decrecía, la media de cigarros consumidos al día tendía a incrementarse.

- Una disparidad considerable entre el consumo de cigarros por día entre hombres y mujeres fumadores, con una mayor proporción de hombres, el 50%, fumando más de un paquete (20 cigarrillos) por día, cantidades que a su vez, solo son correspondidas por menos de una cuarta parte, 25% , de las mujeres fumadoras observadas .

Así, se ha podido dar respuesta a las preguntas planteadas en este trabajo, pudiendo resumir en que, en promedio, de entre 1932 fumadores observados: los hombres fuman más que las mujeres, y los de una edad determinada suelen fumar más que las personas mayores a ellos.

Bibliografía

Tukey, J.W (1977) *Exploratory Data Analysis.*

Reading, MA: Addison-Wesley

Kaggle.com *Smoker´s Health Data*

Smoker’s Health Data (kaggle.com)