Facultad de Estudios Superiores Acatlan
Universidad Nacional Autonoma de Mexico

Lic. Matemáticas Aplicadas y Computación

Proyecto Semestral
Estadística I

Gomez Gonzalez Astrid Yoatziry
Mendoza Villar Jose Ricardo
Navarro Ramos Karen Lizbeth

Profra: Martinez Elizabeth Gomez

1 feb 2021

Índice

Introducción

1. Contexto del problema

Hacer un análisis de los productos más vendidos para una empresa es una práctica bastante útil, ya que así se pueden tomar decisiones en el futuro para maximizar las ventas de dicha empresa, ya sea modificando los precios, aumentando la visibilidad de ciertos productos, crear promociones atractivas, disminución o aumento de inventario dependiendo del producto, entre otras acciones.

Para este proyecto con ayuda de web scraping recolectamos la información de los 100 libros más vendidos dentro de la plataforma Amazon.com, desde el año 1995 hasta 2020, con la que creamos un dataset. Realizando un análisis de estos datos podemos observar el formato preferido por los consumidores, así como el rango de precios entre los que oscilan y la cantidad de reviews para posteriormente enfocarse en las ventas con mayor impacto y así maximizar las ventas de los años siguientes.

2. Datos y Variables

Con un total de observaciones en nuestro conjunto de datos de 2583 (posterior a eliminar los datos nulos*). Definimos las siguientes variables para su análisis:

$X_1 : Type$

Formato del libro

(Audio CD, Boarbook, Cards, Hardcover, Hardcover-spiral, Imitation Leather, Mass Market Paperback, Novelty Book, Pamphlet, Paperback, Plastic-comb, Spiral-bond).

$X_2 : Price$

Precio del libro en dólares.

$X_3 : Reviews$

Total de reseñas en Amazon

*Consultar código en R (Apéndice 1.1)

Tema 1: Introducción a la Estadística (Análisis Exploratorio de Datos)

1. Análisis exploratorio

Para el análisis exploratorio de nuestras variables de interés distinguimos su tipo y escala individualmente obteniendo la siguiente tabla:

Tabla 1.1: Escalas de medición

Variable	X_1	X_2	X_3
	Formato de libro	Precio del libro	Total de reviews en amazon
Tipo	Cualitativa	Cuantitativa	Cuantitativa
Escala	Nominal	Razón	Razón

Asimismo, obtuvimos algunos estadísticos de las variables de interés resaltando los datos en la tabla, recordemos que X_1 es cualitativa por lo cual no es posible la obtención de todos ellos.

Tabla 1.2: Estadísticos

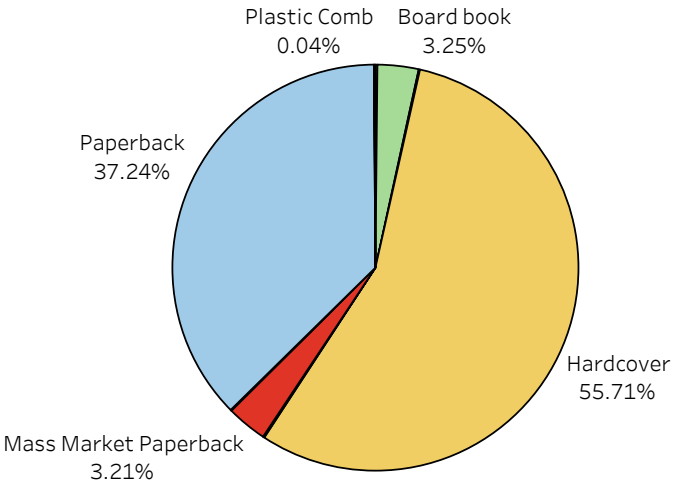
Estadístico	X_1	X_2 (USD)	X_3
Mínimo	-	0.01	1
q1	-	1.605	715
q2 (Mediana)	-	11.95	3369
Media	-	12.357	8211
q3	-	17.235	10600
Máximo	-	379.99	129991
Desv Estándar	-	14.82503	12616.37
Moda	Hardcover	0.94	3851

*datos obtenidos con Rstudio (Apéndice 1.2)

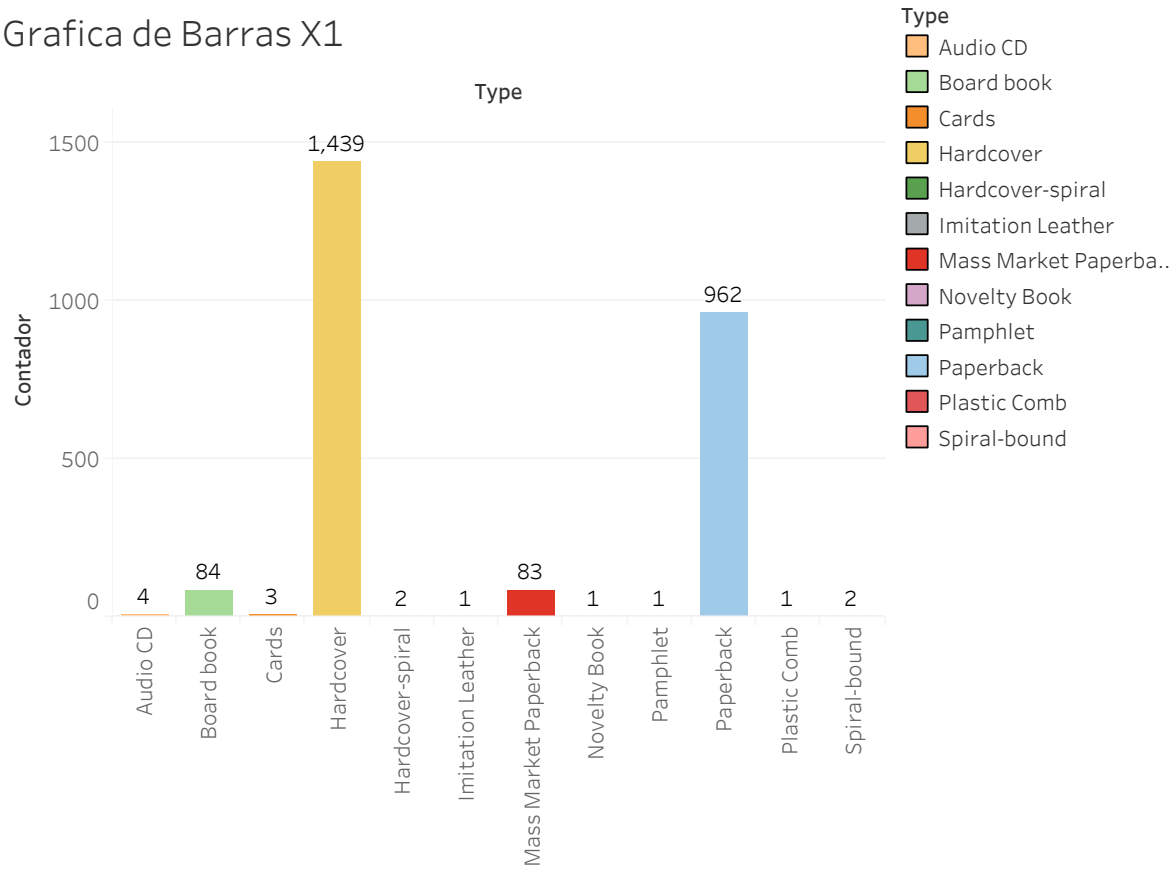
Utilizando dichos datos, construimos las siguientes gráficas para enriquecer nuestro análisis de una manera visual y más atractiva, así como identificar características de nuestras variables que no se pueden apreciar a simple vista con los estadísticos.

X1: Type

Grafica Circular X1

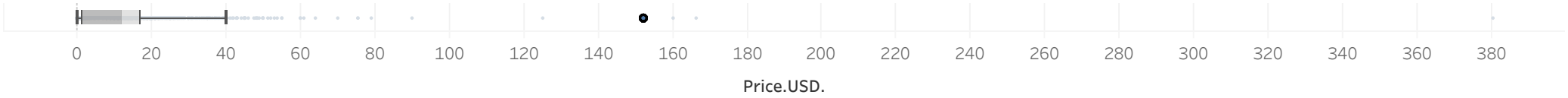


Grafica de Barras X1

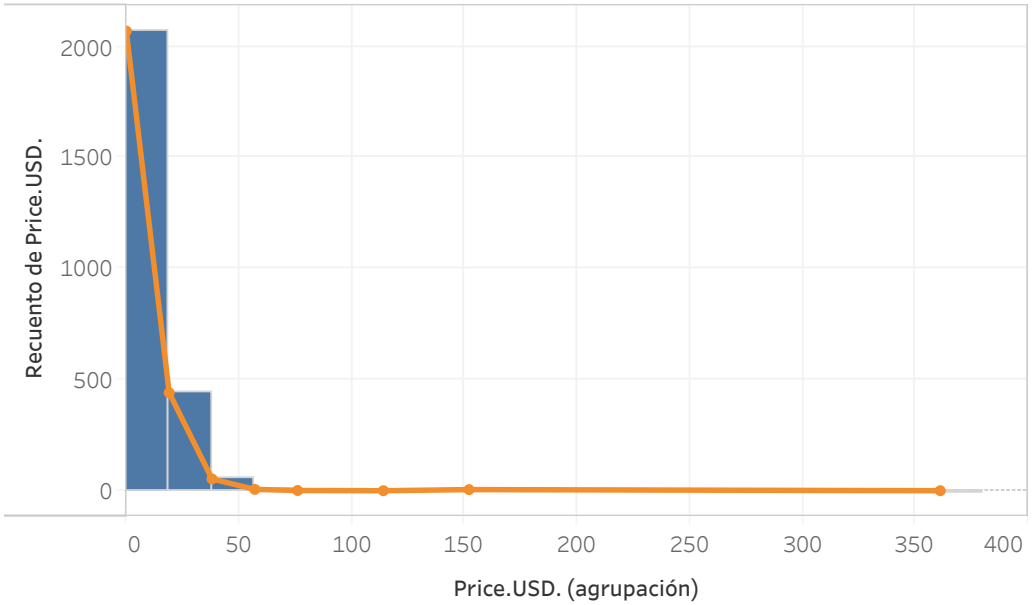


X2: Price

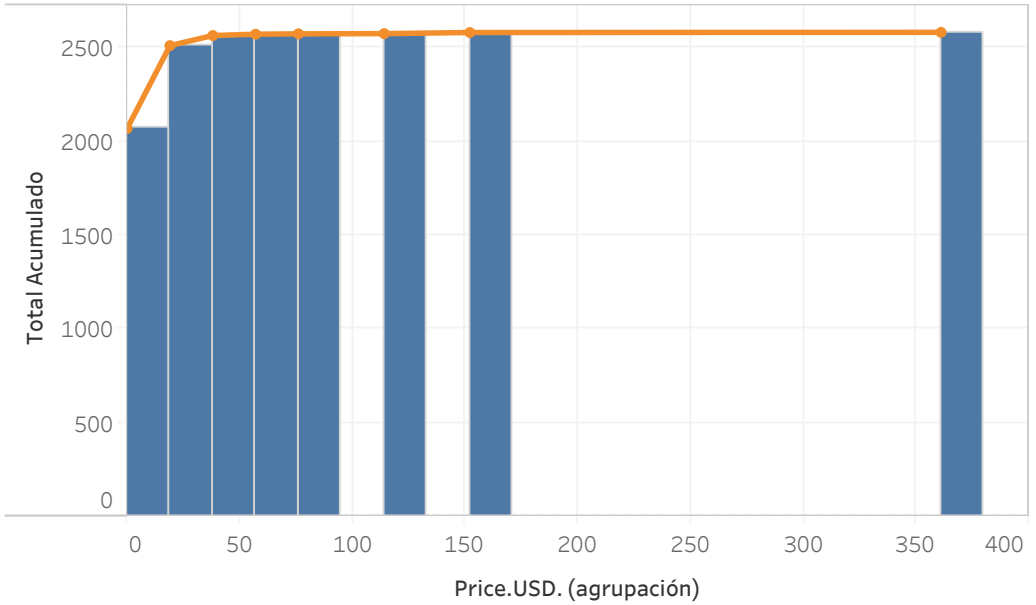
BoxPlot X2



Histograma de Frecuencias X2

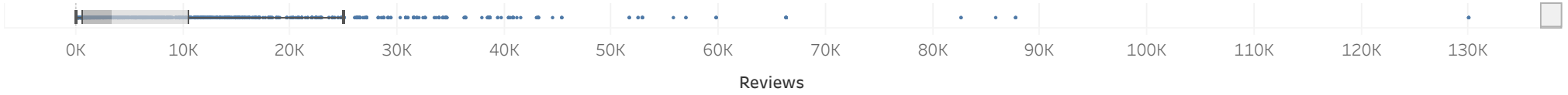


Ojiva X2

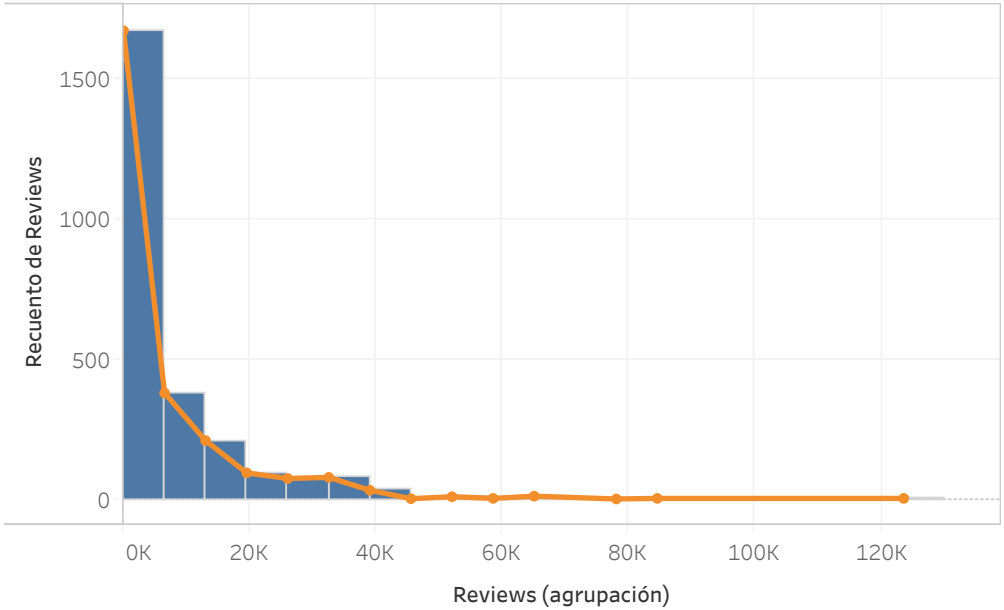


X3: Reviews

BoxPlot X3



Histograma de Frecuencias X3



Ojiva X3

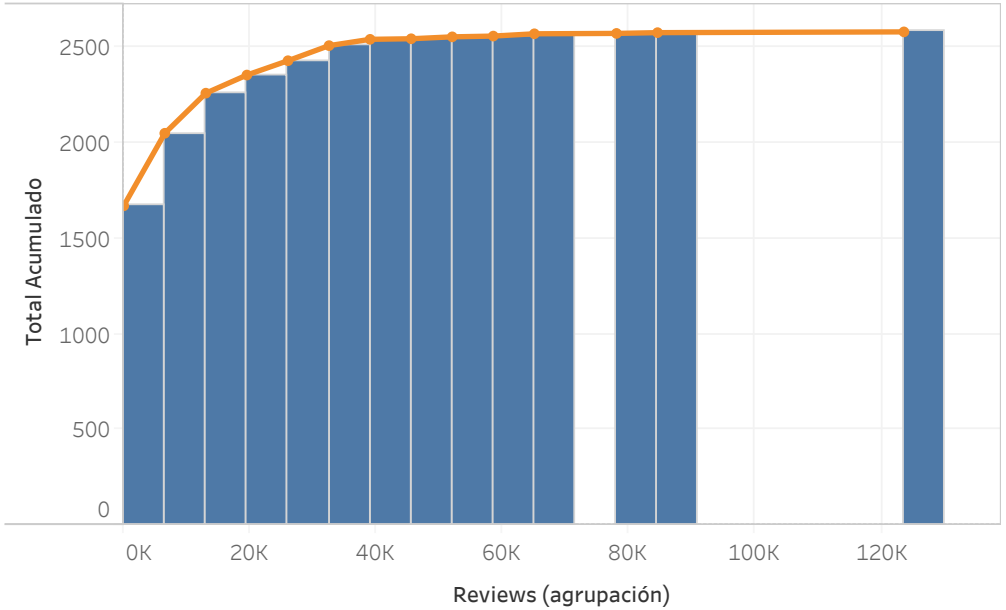


Tabla 1.3: Datos extras

	X_1	X_2	X_3
Curtosis	-	170.1053	21.92606
Coeficiente de Asimetría	-	8.649496	3.415323
Clases (datos agrupados)	-	20	20
Anchura (datos agrupados)	-	19	6500

*datos obtenidos de R (Apéndice 1.3)

Complementando con una tercera tabla de estadísticos y datos extras que nos serán útiles para el análisis de las variables de interés, podemos dar como resultado:

Respecto a $X_1(Type)$ podemos observar como la categoría que predomina en nuestros datos es Hardcover, la cual abarca un 55.71% de nuestros datos, seguido de Paperback con 37.24% y BoardBook con 3.25%. Al sumar estas 3 categorías englobamos el 96.2% de nuestras observaciones totales, por esta razón, las 9 categorías restantes, al tener un porcentaje muy pequeño del total, es difícil observarlas en las gráficas, sin embargo en la siguiente tabla de resumen podemos apreciar el número total de observaciones que caen dentro de cada categoría.

Tabla 1.4: Tabla de resumen X_1

Type	Frecuencias absolutas		Type	Frecuencias absolutas
Audio CD	4		Mass Market Paperback	83
Boardbook	84		Novelty Book	1
Cards	3		Pamphlet	1
Hardcover	1439		Paperback	962
Hardcover-spiral	2		Plastic-comb	1
Imitation Leather	1		Spiral-bond	2

*Datos obtenidos de Tableau

En cuanto a $X_2\,(Price)$, para realizar los diagramas de Histograma y Ojiva correspondientes, clasificamos a nuestra variable en intervalos con anchura 19, tomando como marca de clase el mínimo de esta misma para un manejo más sencillo en Tableau, de esta forma obtenemos un total de 20 clases.

Sabemos que tiene una asimetría a la derecha siendo muy evidente en nuestro BoxPlot e histograma de frecuencias, además podemos comprobarlo al obtener su coeficiente de asimetría, $8.649496 > 0$ y con:

$$\begin{array}{ccccc} Moda & < & Mediana & < & Media \\ 0.94 & < & 11.95 & < & 12.36 \end{array}$$

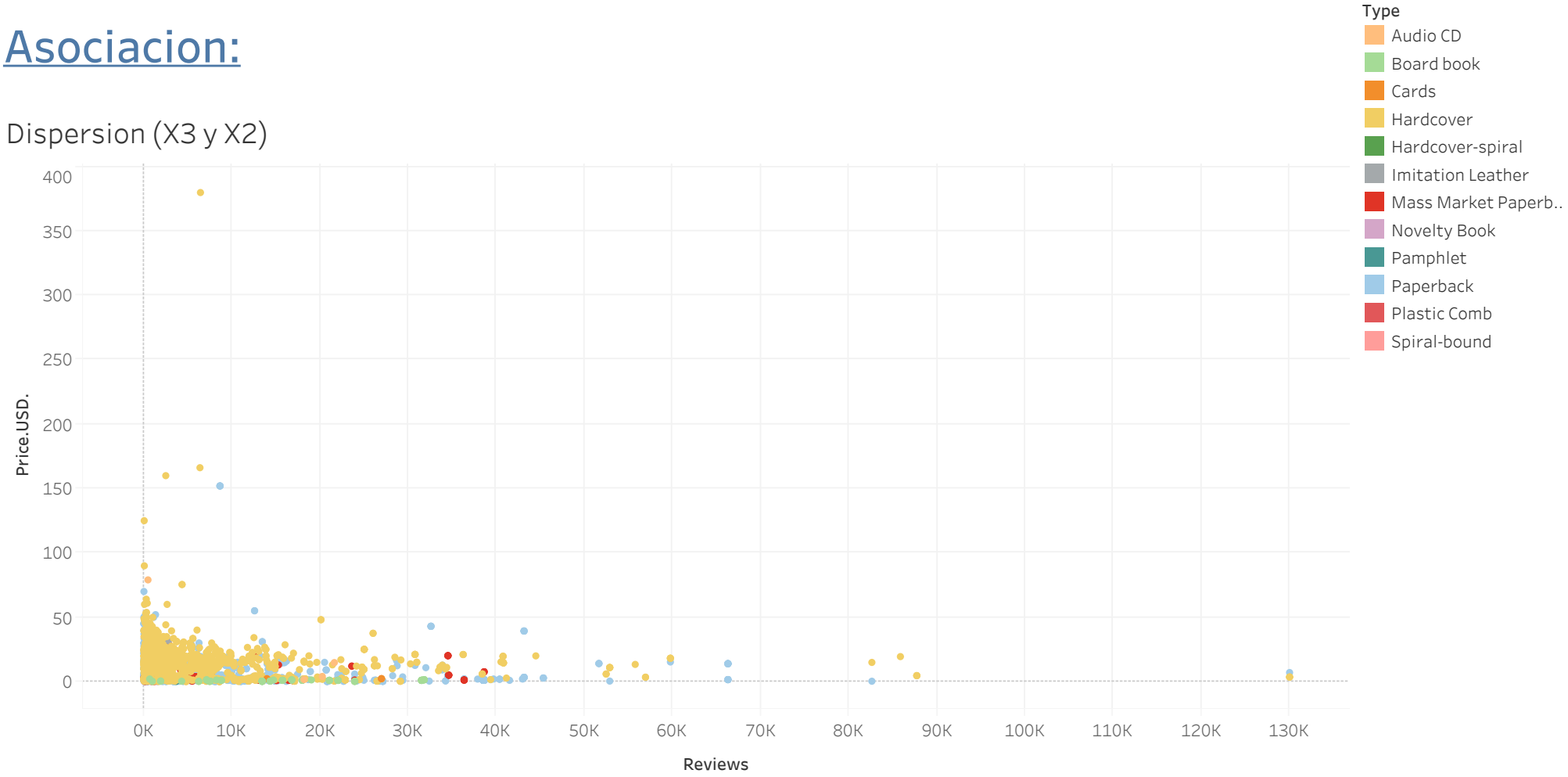
Asimismo podemos resumir que es Leptocúrtica al observar el histograma, o al obtener su curtosis, que al ser igual a 170.1053 es mayor que 3. Resaltamos outliers muy marcados a la derecha en el precio, esto significa que hay ciertos libros que excedan del precio promedio drásticamente. Además, con la ojiva observamos que el 50% de los datos se acumulan en nuestra primera clase, la cual hace referencia al intervalo de 0 a 14 USD.

En relación con $X_3\,(Reviews)$ utilizamos una anchura de 6500 para obtener 20 clases. Observamos en la ojiva como el 50% se acumula desde nuestra primera categoría que abarca de 0 a 6500 reviews. Del mismo modo, en el histograma observamos que es Leptocúrtica y lo corroboramos con su curtosis mayor a 3 de 21.92606.
Sabemos que es asimétrica a la derecha por su coeficiente de asimetría (3.415323) mayor a cero, corroborando al observar el BoxPlot y el histograma. Adicionalmente, observamos datos atípicos (Outliers) dándonos a entender que existen libros con un total de reviews muy por encima al promedio de esta variable (8211), por ejemplo existe un libro con 129991 reviews.

Posterior a dicho análisis, podemos comenzar a pensar en relación respecto precio y reviews (X_2, X_3), ya que al ser un libro más barato, tendrá mayor acceso a distintas personas, y así es más probable que una cantidad mayor de personas escriban una reseña sobre él. Además, pensamos en una relación de precio y tipo (X_2, X_1) ya que dependiendo de los materiales utilizados en la fabricación del libro, así como el tamaño del producto, el precio puede variar.

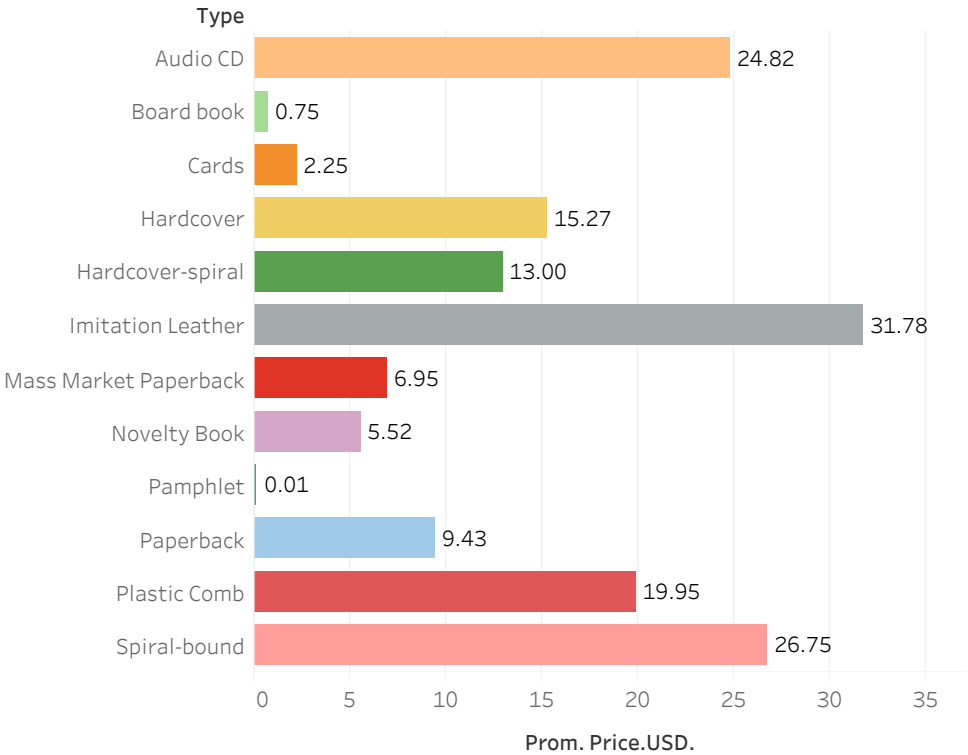
Asociacion:

Dispersion (X3 y X2)

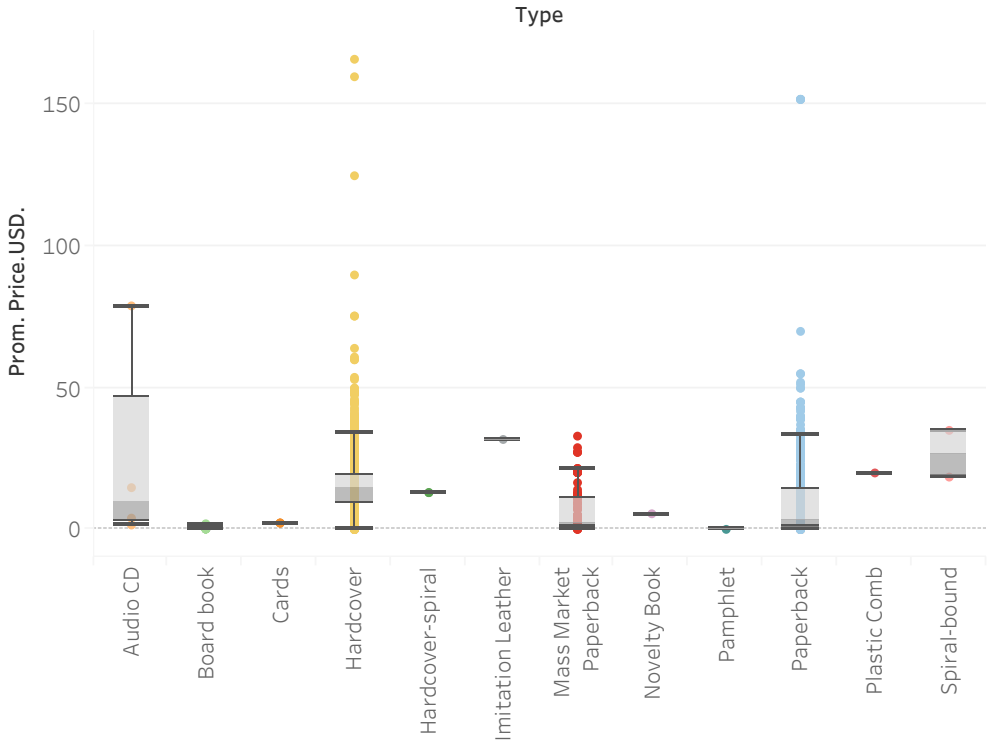


Comparacion:

Grafica de Barras (X1 y X2)



BoxPlot (X1 y X3)



2. Asociación:

Con las ideas planteadas anteriormente de la relación entre variables, utilizando las dos cualitativas (X_2, X_3) podemos considerar una relación entre precio y total de reviews, retomando la idea del punto anterior, sabemos que al ser un libro más barato, es más probable que mayor cantidad de personas lo adquieran y por lo tanto de obtener reviews.
Aunque es preciso recordar que al comprar un libro no estás obligado a escribir una reseña sobre él, por lo que este número no refleja la cantidad total de libros vendidos. Sin embargo, también sabemos que gracias a las políticas de Amazon, no es posible escribir una reseña sin haber comprado el producto en la página con anterioridad.

Para verificar si hay relación entre ambas variables, calculamos su covarianza (-30638.82) y además la correlación de Pearson (-0.1638107) para evitar problemas con la diferencia de unidades entre las variables*. Con este último podemos observar que no existe asociación lineal, ya que la magnitud de r se encuentra muy alejada de nuestro número de referencia 1.

$$|r| = |-0.1638107| < 1$$

Dicha relación podemos observarla en la gráfica de dispersión donde no se esboza una línea recta entre la nube de datos. Este resultado se puede deber a los grandes outliers que existen entre nuestros datos, ya que desequilibran tanto cada variable y esto repercute posteriormente.

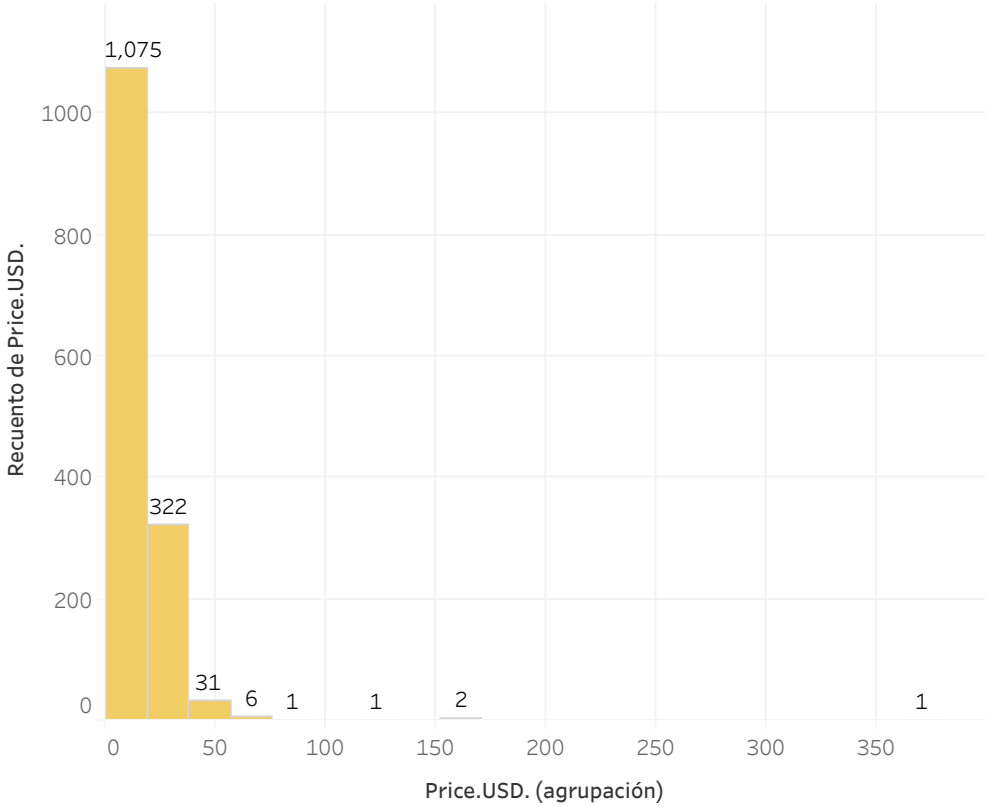
3. Comparación:

De manera similar podemos pensar en una relación entre entre precio y tipo de libro (X_2, X_1), ya que mientras el formato es de mejor calidad y tamaño, el precio promedio del libro aumentará. Un ejemplo de este comportamiento lo podemos observar en la gráfica de barras de los precios promedios respecto al tipo de libro, donde “Imitation Leather” al utilizar materiales más caros en comparación a los demás (Imitación de Piel) su precio promedio es el más alto. Por otro lado tenemos “Cards” (Cartas) que al tener un tamaño más pequeño que la mayoría, su precio promedio es de los más bajos.
Un factor importante también es la cantidad de libros que tenemos en dicha categoría, esto se ve reflejado en nuestros diagramas BoxPlot donde entre menos libros existen en dicha categoría, su caja será más compacta, esto quiere decir que sus medidas de tendencia se acercan a un valor similar. Como ejemplo tomaremos “Plastic comb” que al solo tener un libro de esta categoría, los bigotes, cuartiles superiores e inferiores, junto con la mediana son iguales a 20.
Para un análisis más profundo de nuestras categorías realizamos histogramas con las 4 más dominantes, para observar su comportamiento respecto al precio en cada categoría, indicando en cada eje horizontal el precio en USD por clases y en el vertical la cantidad de libros pertenecientes a dicha categoría en ese rango de precios. En las 4 gráficas observamos que están sesgadas a la derecha a pesar de que el precio varía mucho entre el cambio de categoría.

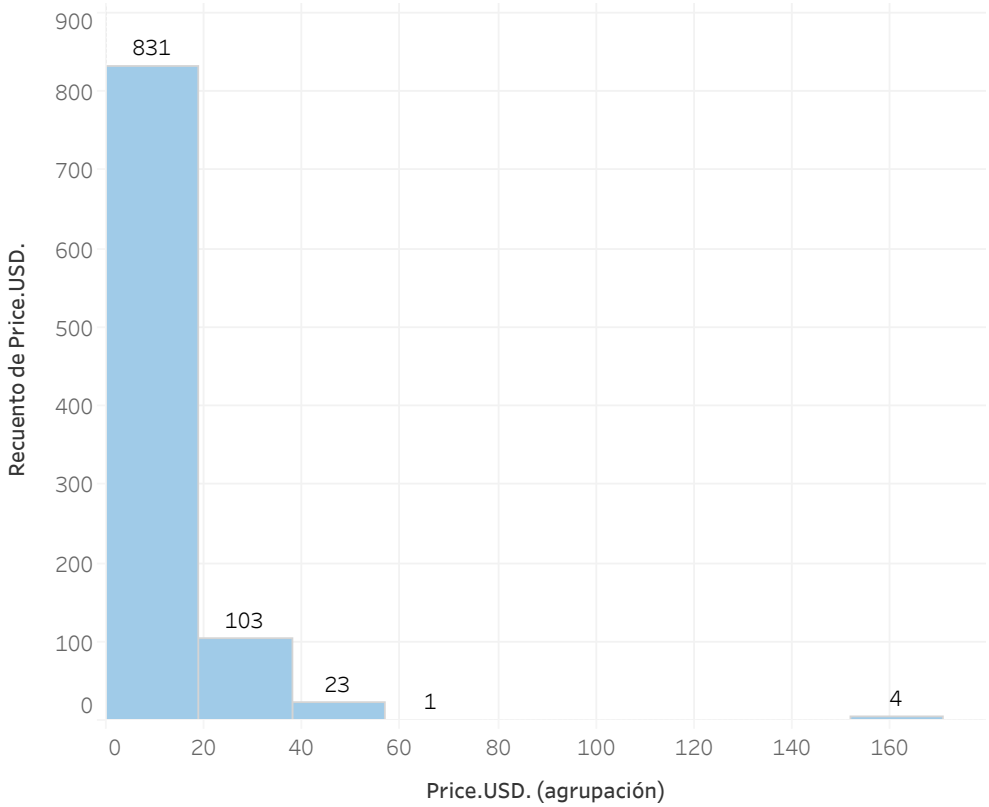
*Datos obtenidos con Rstudio (Apéndice 1.4)

Comparacion (continuacion):

Hardcover

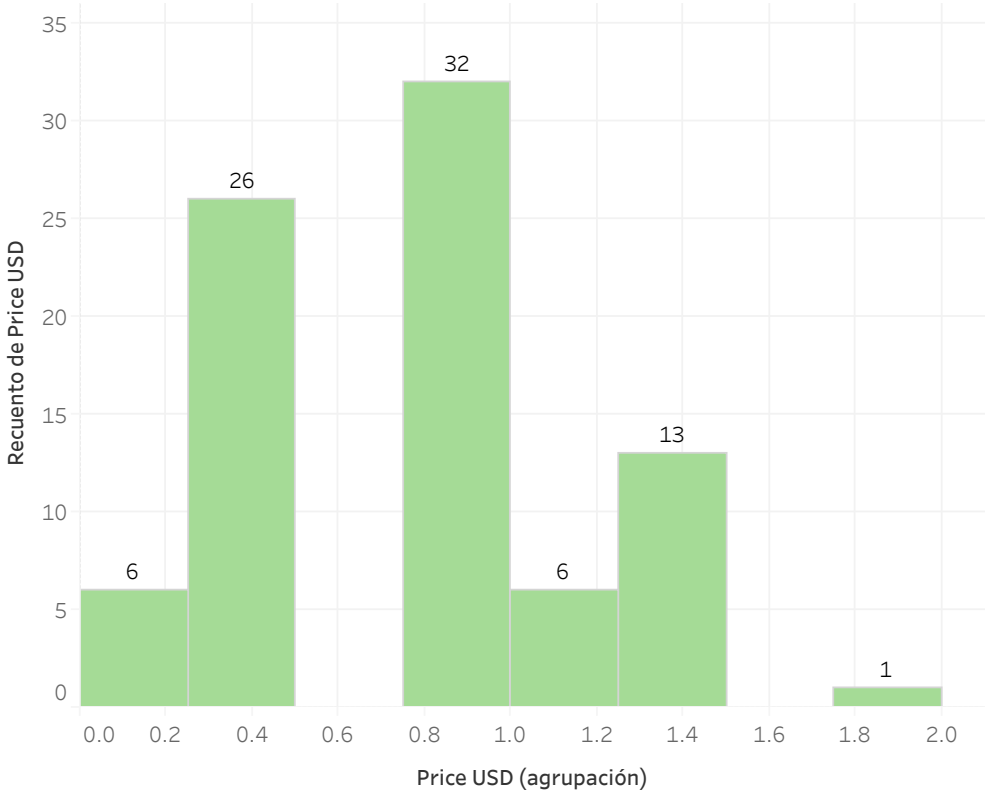


Paperback

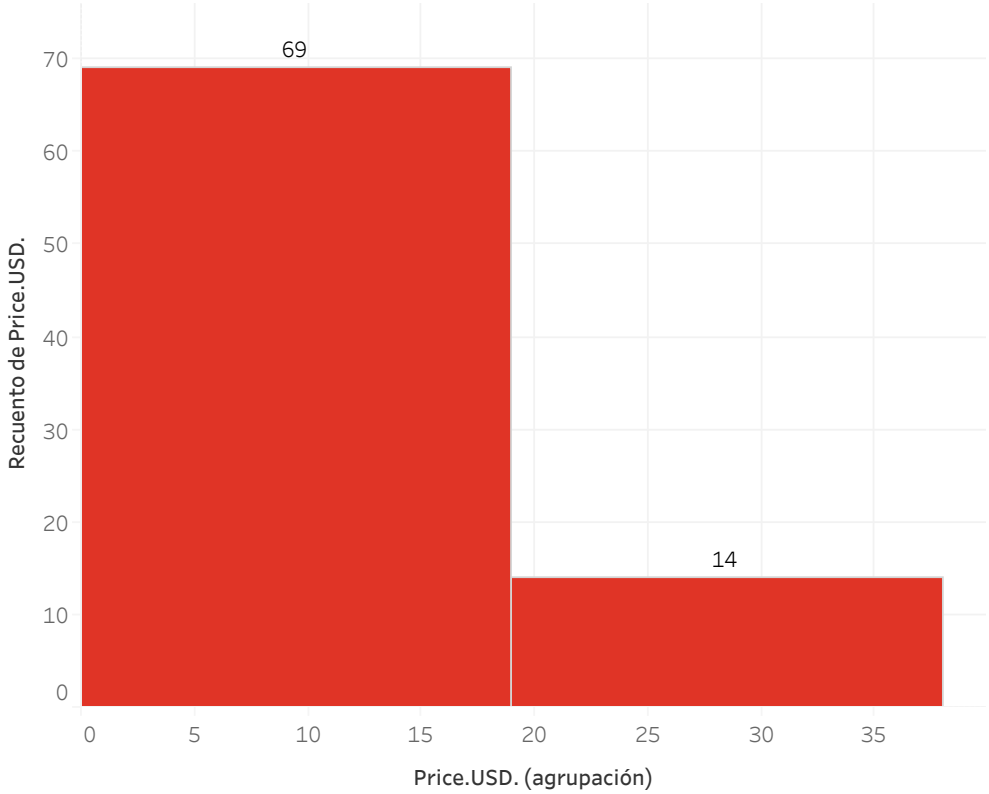


Comparacion (continuacion):

Boardbook



MassMarket



Tema 2: Muestreo
Tema 4: Distribuciones Muestrales

1.Muestreo

Realizando una muestra, mediante la técnica Muestreo Aleatorio Simple (MAS) con respecto a nuestra variable cuantitativa X_2 que hace referencia al precio, obtuvimos una muestra de tamaño 6.

$$\Omega = \{u_1, u_2, u_3, u_4, u_5, u_6\}$$
$$\Omega = \{11.9,\ 0.25,\ 5.43,\ 4.86,\ 13.9,\ 21.0\}$$

*datos obtenidos mediante la función sample del lenguaje R (Apéndice 2.1)

2.Distribuciones Muestrales

Mediante la "población" Ω obtenida con anterioridad, obtenemos los siguiente:

-Media (μ)

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

donde $N = 6$

$$\mu = \tfrac{1}{6}(11.9 + 0.25 + 5.43 + 4.86 + 13.9 + 21.0)$$
$$\mu = \tfrac{1}{6}(57.34)$$
$$\mu = 9.556666667$$

-Varianza (σ^2)

$$\sigma^2 = \frac{1}{N}(\sum_{i=1}^{N} x_i^2 - N\mu^2)$$

donde $N = 6$

$$\mu^2 = (9.556666667)^2 = 91.32987778$$
$$\sigma^2 = \tfrac{1}{6}\left[(11.9^2 + 0.25^2 + 5.43^2 + 4.86^2 + 13.9^2 + 21.0^2) - (6)(91.32987778)\right]$$
$$\sigma^2 = \tfrac{1}{6}\left[(141.61 + 0.0625\ + 29.4849 + 23.6196 + 193.21 + 441) - 547.9792667\right]$$
$$\sigma^2 = \tfrac{1}{6}\left[(828.987) - 547.9792667\right]$$
$$\sigma^2 = \tfrac{1}{6}\,(281.0077333)$$
$$\sigma^2 = 46.83462222$$

Basándonos en la “población” Ω las muestras posibles de tamaño n=3 (sin reemplazo) es la siguiente:

$\Omega = \{11.9,\ 0.25,\ 5.43,\ 4.86,\ 13.9,\ 21.0\}$

$C_6^3 = 20$

Tabla 2.1: Muestras posibles de tamaño 3

#	Muestra
1	(0.25, 4.86, 5.43)
2	(0.25, 4.86, 11.9)
3	(0.25, 4.86, 13.9)
4	(0.25, 4.86, 21,0)
5	(0.25, 5.43, 11.9)
6	(0.25, 5.43, 13.9)
7	(0.25, 5.43, 21.0)
8	(0.25, 11.9, 13.9)
9	(0.25, 11.9, 21.0)
10	(0.25, 13.9, 21.0)
11	(4.86, 5.43, 11.9)
12	(4.86, 5.43, 13.9)
13	(4.86, 5.43, 21.0)
14	(4.86, 11.9, 13.9)
15	(4.86, 11.9, 21.0)
16	(4.86, 13.9, 21.0)
17	(5.43, 11.9, 13.9)
18	(5.43, 11.9, 21.0)
19	(5.43, 13.9, 21.0)
20	(11.9, 13.9, 21.0)

*datos obtenidos del lenguaje R (Apéndice 2.2)

A continuación se mostrará la distribución de muestreo para la media ($\overline{X}$) y para la varianza muestral (S^2).
Nota: Suponiendo sin orden.

Tabla 2.2: Promedio y varianza muestral

#	Muestra	$\overline{X}$	S^2
1	(0.25, 4.86, 5.43)	3.51333	8.068233
2	(0.25, 4.86, 11.9)	5.67	34.4227
3	(0.25, 4.86, 13.9)	6.336667	48.21603
4	(0.25, 4.86, 21.0)	8.703333	118.719
5	(0.25, 5.43, 11.9)	5.86	34.0693
6	(0.25, 5.43, 13.9)	6.526667	47.48263
7	(0.25, 5.43, 21.0)	8.893333	116.6366
8	(0.25, 11.9, 13.9)	8.683333	54.34083
9	(0.25, 11.9, 21.0)	11.05	108.1825
10	(0.25, 13.9, 21.0)	11.716667	111.2158
11	(4.86, 5.43, 11.9)	7.396667	15.29123
12	(4.86, 5.43, 13.9)	8.063333	25.63123
13	(4.86, 5.43, 21.0)	10.43	83.8749
14	(4.86, 11.9, 13.9)	10.22	22.5472
15	(4.86, 11.9, 21.0)	12.586667	65.47853
16	(4.86, 13.9, 21.0)	13.253333	65.43853
17	(5.43, 11.9, 13.9)	10.41	19.6003
18	(5.43, 11.9, 21.0)	12.776667	61.18263
19	(5.43, 13.9, 21.0)	13.443333	60.76263
20	(11.9, 13.9, 21.0)	15.6	22.87

*datos obtenidos para $\overline{X}$ en R (Apéndice 2.3)
*datos obtenidos para S^2 en R (Apéndice 2.4)

Tabla 2.3: Distribución de muestreo para la media ($\overline{X}$)

$\overline{X}$	$f(\overline{X})$
3.51333	$\frac{1}{20}$
5.67	$\frac{1}{20}$
6.336667	$\frac{1}{20}$
8.703333	$\frac{1}{20}$
5.860000	$\frac{1}{20}$
6.526667	$\frac{1}{20}$
8.893333	$\frac{1}{20}$
8.683333	$\frac{1}{20}$
11.05	$\frac{1}{20}$
11.716667	$\frac{1}{20}$
7.396667	$\frac{1}{20}$
8.063333	$\frac{1}{20}$
10.43	$\frac{1}{20}$
10.22	$\frac{1}{20}$
12.586667	$\frac{1}{20}$
13.253333	$\frac{1}{20}$
10.41	$\frac{1}{20}$
12.776667	$\frac{1}{20}$
13.443333	$\frac{1}{20}$
15.6	$\frac{1}{20}$

-Valor esperado de la media ($\overline{X}$) :

$E[\overline{X}] = \sum \overline{X} f(\overline{X}) = 9.556665$ *datos obtenidos en R (Apéndice 2.5)

Comparando el resultado con el de la media (μ)=9.556666, podemos observar que son equivalentes.

Tabla 2.4: Distribución de muestreo para la varianza (S^2)

S^2	$f(S^2)$

8.068233	$\frac{1}{20}$
34.4227	$\frac{1}{20}$
48.21603	$\frac{1}{20}$
118.719	$\frac{1}{20}$
34.0693	$\frac{1}{20}$
47.48263	$\frac{1}{20}$
116.6366	$\frac{1}{20}$
54.34083	$\frac{1}{20}$
108.1825	$\frac{1}{20}$
111.2158	$\frac{1}{20}$
15.29123	$\frac{1}{20}$
25.63123	$\frac{1}{20}$
83.8749	$\frac{1}{20}$
22.5472	$\frac{1}{20}$
65.47853	$\frac{1}{20}$
65.43853	$\frac{1}{20}$
19.6003	$\frac{1}{20}$
61.18263	$\frac{1}{20}$
60.76263	$\frac{1}{20}$
22.87	$\frac{1}{20}$

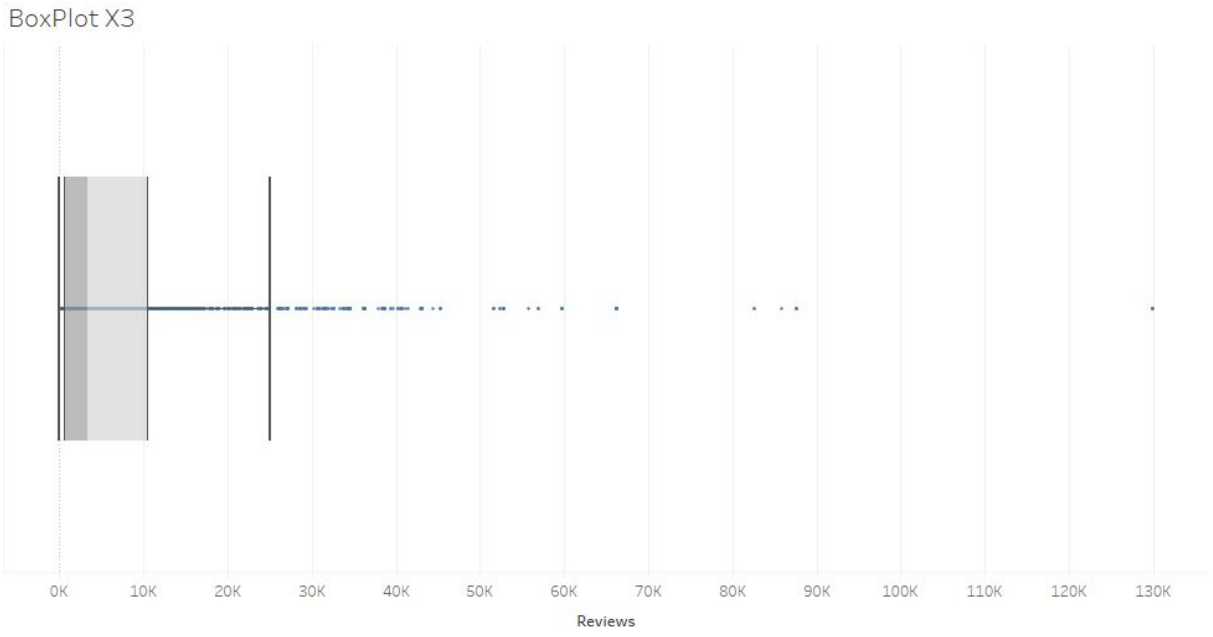
-Valor esperado de la varianza (S^2).

$$E[S^2] = \sum S^2 f(S^2) = 56.2015402$$ *datos obtenidos en R (Apéndice 2.5)

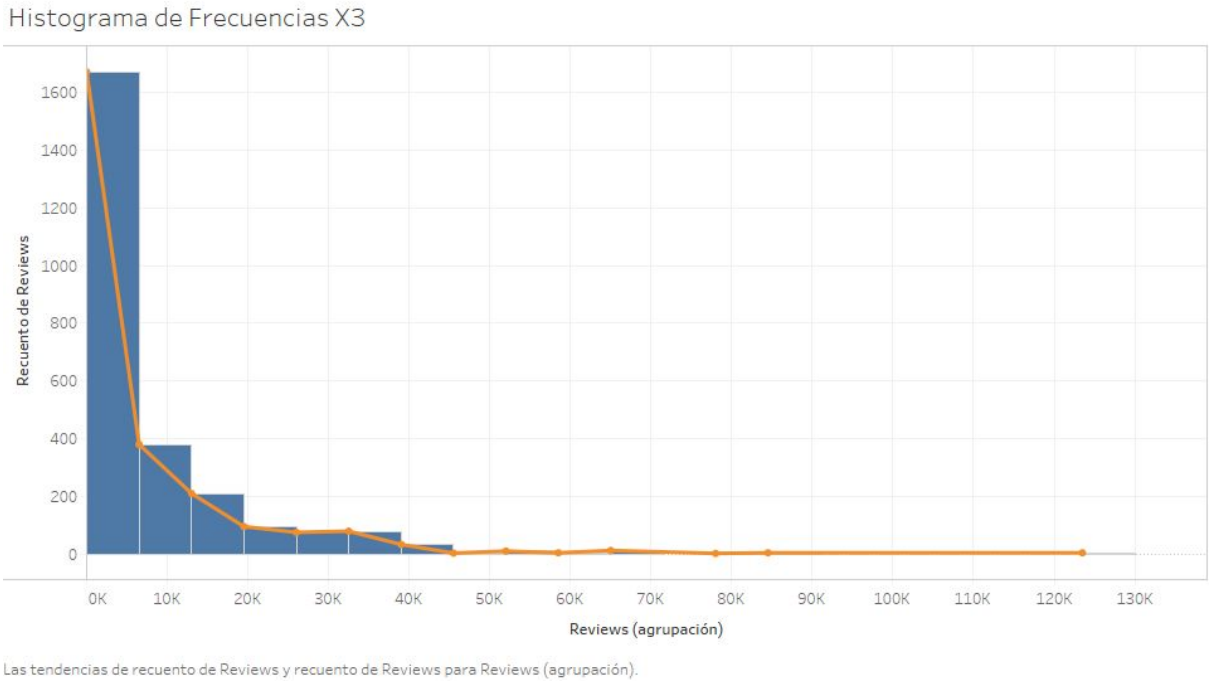
Comparando el resultado con el de la varianza $\sigma^2 = 46.83462222$, podemos observar que son distintos.

Teorema Central del Límite

Considerando a nuestra variable cuantitativa X3 (Reviews) se graficó su histograma de frecuencias, acompañado de su polígono de frecuencias. Al igual se requirió de la herramienta del diagrama de caja y brazos.. A continuación se muestran dichas gráficas:



Reviews.



Las tendencias de recuento de Reviews y recuento de Reviews para Reviews (agrupación).

De acuerdo a las gráficas no se puede asegurar que la distribución de la variable X3 tenga una distribución normal ya que no tiene presenta la forma de normal.

A continuación se realizaron varias muestras aleatorias con reemplazo y de tamaño fijo n ($n \geq 25$), donde n=50

*datos obtenidos en R (Apéndice 2.6)

A cada muestra obtenida se le estimó la media.

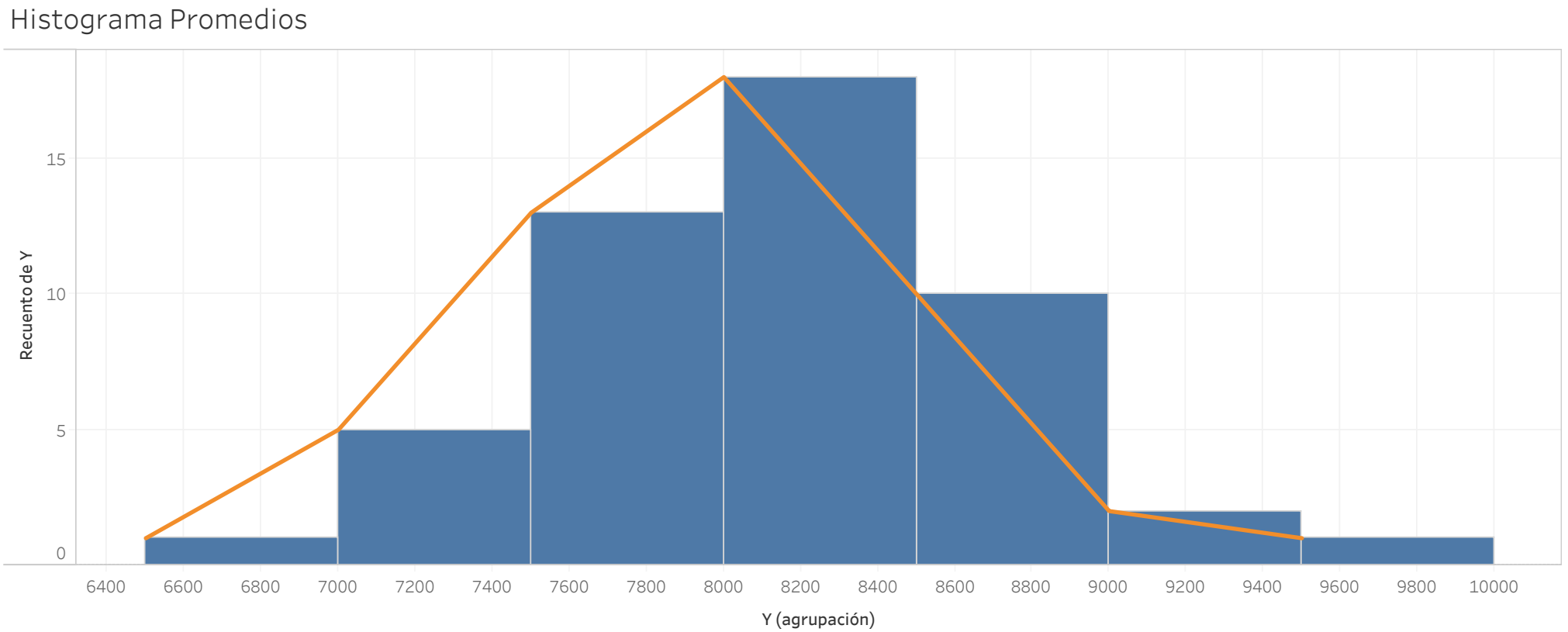
*datos obtenidos en R (Apéndice 2.7)

Por consiguiente se obtuvo un conjunto de promedios muestrales:

Y= {7257.09, 8156.717, 7714.38, 8337.537, 8512.363, 8628.187, 7937.79, 8668.937, 7765.267, 7580.407, 9310.927, 8144.65, 8285.603, 7757.467, 8962.79, 8274.477, 7703.5, 8019.933, 8312.303, 8154.74, 7369.487, 6618.873, 8354.73, 8083.973, 7337.92, 7623.683, 8110.483, 8718.87, 8593.58, 8534.423, 7692.563, 7749.97, 9262.94, 8189.417, 8432.273, 7656.66, 8393.997, 8495.17, 7366.54, 7502.177, 8113.523, 8304.193, 8710.467, 7259.447, 8855.343, 8890.033, 9874.653, 7928.48, 7754.01, 8182.06}

Considerando el conjunto obtenido anteriormente (Y) se graficó su histograma, mediante el cual podemos observar que tiene forma normal, por lo que se puede validar el Teorema Central del Límite.

Medias Muestrales:



Tema 5: Estimación puntual y por intervalo

Después de haber explorado los datos con un análisis descriptivo de los datos realizaremos un análisis inferencial que nos permitirá entre muchas cosas a tomar decisiones, como primer realización exploraremos la estimación puntual describiendo la proporción muestral de X_1 para los diferentes tipos de libros físicos que se venden en la plataforma de Amazon.

1. Estimación puntual

Sea $\widehat{p}_k$ la $k-ésima$ proporción muestral para la variable aleatoria X_1, con: $k \in \{1 = Audio\ CD, 2 = Boardbook, ..., 12 : Spiral-bond\}$

Tabla 5.1: $k-ésima$ proporción muestral

k	$\widehat{p}_k$
Audio CD	0.001548587
Boardbook	0.03252033
Cards	0.00116144
Hardcover	0.5571041
Hardcover-spiral	0.0007742935
Imitation Leather	0.0003871467
Mass Market Paperback	0.03213318
Novelty Book	0.0003871467
Pamphlet	0.0003871467
Paperback	0.3724352
Plastic-comb	0.0003871467
Spiral-bond	0.0007742935

*Apéndice 5.1

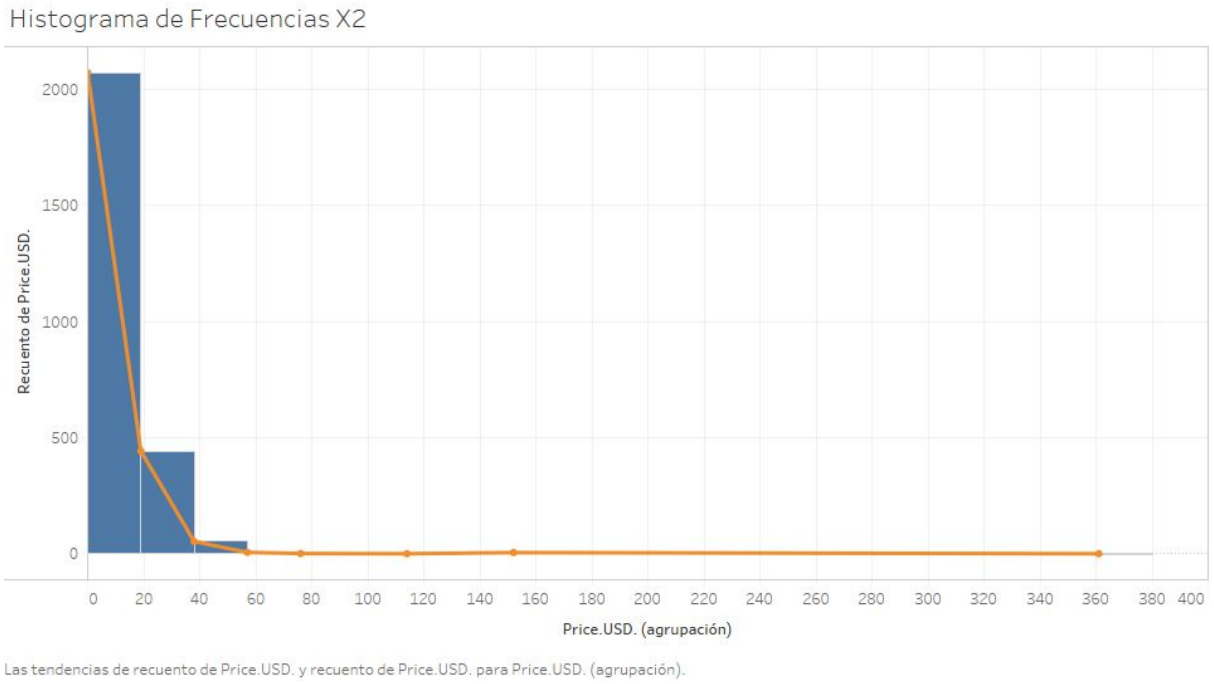
Para nuestras dos variables cuantitativas X_2 y X_3 calcularemos el coeficiente de variación muestral lo que nos permitirá conocer la dispersión relativa de nuestros datos:

$$C.V_{X_2} = \frac{s_{X_2}^2}{\overline{X_2}}(100\%) = \frac{14.82503*(100\%)}{12.35684} = 119.9742\%$$

$$C.V_{X_3} = \frac{s_{X_3}^2}{\overline{X_3}}(100\%) = \frac{12616.37*(100\%)}{8211.067} = 153.6508\%$$ Apéndice 5.2

Cómo $C.V_{X_3} > C.V_{X_2}$ esto significa que las observaciones de la variable aleatoria X_3 tiene una mayor variabilidad que las de X_2, o bien, la variable X_2 cuenta con más homogeneidad.

Como siguiente punto de interés, dado el histograma de frecuencias de la variable X_2 y con base en su perfil, daremos una función de masa de densidad que aproxime la distribución del precio de los libros vendidos en la plataforma:



Debido a que es asimétrica a la derecha y tiene un valor grande cuando $x_2 -> 0$, podemos suponer que se puede distribuir como una exponencial o como una Gamma; para calcular estos parámetros haremos uso del método de máxima verosimilitud para estimar valores para λ si esta es una exponencial o de α y λ si se distribuye como una función Gamma.

Estimador para una función de distribución Exponencial:

Sea $Y_1, Y_2, ..., Y_n$ una muestra aleatoria de $Y \sim exp(\theta)$

$$L(\theta) = \prod_{i=i}^{n} \theta e^{-\theta y} = \theta^n e^{-\theta \sum_{i=1}^{n} y_i}$$

$$ln(L(\theta)) = nln(\theta) - \theta \sum_{i=1}^{n} y_i$$

$$\frac{d}{d\theta} ln(L(\theta)) = \frac{n}{\theta} - \sum_{i=1}^{n} y_i = 0$$

$$\Rightarrow \widehat{\theta}_{MLE} = \frac{1}{\bar{y}}$$

Como $\overline{X_2} = 0.08092683$, entonces $X_2 \sim exp(\lambda = 0.08092683)$

Estimadores para una función de distribución Gamma:

Calcular los estimadores de máxima verosimilitud para α y λ de la distribución Gamma resulta bastante complicado cuando se intenta derivar la función de máxima verosimilitud

Sea $Y_1, Y_2, ..., Y_n$ una muestra aleatoria de $Y \sim Gamma(\alpha, \lambda)$

$$L(\theta) = \prod_{i=i}^{n} \Gamma(\alpha)^{-1} \lambda^{\alpha} (y_i)^{\alpha-1} e^{-\lambda y_i} = \Gamma(\alpha)^{-n} \lambda^{n\alpha} e^{-\lambda \sum_{i=1}^{n} y_i} \prod_{i=1}^{n} (y_i)^{\alpha-1}$$

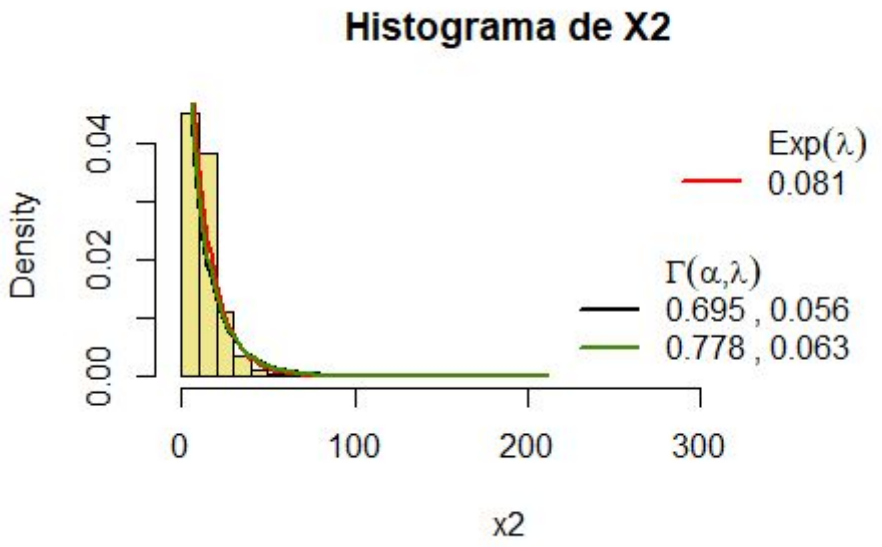
$$ln(L(\theta)) = - n * ln(\Gamma(\alpha)) + n\alpha ln(\lambda) + (\alpha - 1) \sum_{i=1}^{n} ln(y_i) - \lambda \sum_{i=1}^{n} y_i$$

La dificultad radica en derivar $ln(\Gamma(\alpha))$, por lo que resulta conveniente estimar estos valores con ayuda de métodos numéricos, para los valores iniciales de α y λ usaremos como estimadores los que se obtienen por el método de momentos, esto es:

$$\widehat{\alpha}_{MOM} = \frac{\overline{Y}^2}{s_y^2}; \ \widehat{\lambda}_{MOM} = \frac{\overline{Y}}{s_y^2}$$

Después con ayuda de la función en R llamada nlm(...) maximizamos los valores para $\widehat{\alpha}$ y $\widehat{\lambda}$. Apéndice 5.3

Después de haber obtenido los valores de los estimadores y graficarlos sobre el histograma de frecuencias con una cantidad de 30 clases para propósitos más descriptivos obtenemos la siguiente gráfica:



Comparando las distribuciones exponencial y las gammas, podemos ver que las 3 dan una aproximación bastante “similar”, la distribución Gamma de color negro es la que se obtuvo con los valores iniciales para α y λ aproximados con el métodos de momentos y la distribución Gamma de color verde es la que se obtuvo después de haber aplicado el método de maxima verosimilitud, pero ¿Qué distribución elegir? Debido a que la distribución exponencial se utiliza en mayor medida para modelar el

tiempo de funcionamiento o de espera de un evento, y que además la exponencial es un caso de la distribución Gamma, elegimos la Gamma como una “mejor” distribución para modelar el precio de los libros más vendidos, de forma inmediata surge el cuestionamiento ¿Qué valores para los parámetros estimados conviene usar, los obtenidos por el método de momentos o por máxima verosimilitud? En la naturaleza de ambos métodos, el método de momentos resulta más fácil de obtener pero con la posible consecuencia de obtener estimadores sesgados o más sesgados que con la función de máxima verosimilitud, por lo que consideraremos los valores para $\widehat{\alpha}(0.7781659)$ y $\widehat{\lambda}(0.062974)$ obtenidos con la aproximación numérica del método de máxima verosimilitud.

2. Estimación por intervalos

Después de haber realizado inferencia de manera puntual, el siguiente objetivo es realizar intervalos de confianza que a diferencia de un estimador puntual, nos proporcionan un rango de valores posibles para los valores con una confianza determinada (normalmente con un valor elevado), resulta conveniente expresar los parámetros estimados en intervalos puesto que a pesar de que el estimador puntual sea una buena aproximación para el valor poblacional, este valor no significa que coincida con el verdadero. El primer paso es calcular los valores de la media y la varianza poblacional para nuestras variables X_2 y X_3

$$\mu_{X_2} = \frac{1}{2583}(31917.72) = 12.35684; \qquad \sigma_{X_2}{}^2 = \frac{1}{2583}(961877.7 - 2583(12.35684)^2) = 219.6963$$
$$\mu_{X_3} = \frac{1}{2583}(21209185) = 8211.067; \qquad \sigma_{X_3}{}^2 = \frac{1}{2583}(585134198015 - 2583(8211.067)^2) = 159111176$$

Resolveremos el siguiente problema planteado: Para la variable cualitativa X_1 estime un intervalo de confianza del 90% para la proporción p_k. Considere las k categorías que describen a la variable. ¿El valor verdadero p_k para cada variable se encuentra dentro del intervalo o está fuera?

Sea Z la cantidad pivotal:

$$Z = \frac{\widehat{p}_k - p}{\sqrt{\frac{pq}{n}}} \quad \Rightarrow \; p \in (\widehat{p}_k - Z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}_k(1-\widehat{p}_k)}{n}}, \widehat{p}_k + Z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}_k(1-\widehat{p}_k)}{n}})$$

Con ayuda de R se estimaron todos los intervalos de confianza para las proporciones muestrales, en la siguiente tabla se muestran los intervalos obtenidos con la respuesta a la pregunta:

Tabla 5.2: Intervalos de confianza para la $k-\acute{e}sima$ proporción muestral

k	Intervalo de confianza para p	¿p está en el intervalo?
Audio CD	(0.000275974,0.002821200)	Sí
Boardbook	(0.02677964,0.03826101)	Sí
Cards	(5.911145e-05,2.263769e-03)	Sí
Hardcover	(0.5410279 0.5731804)	Sí
Hardcover-spiral	(-0.0001259286,0.0016745155)	Sí
Imitation Leather	(-0.0002495297,0.0010238231)	Sí
Mass Market Paperback	(0.02642563,0.03784073)	Sí
Novelty Book	(-0.0002495297,0.0010238231)	Sí
Pamphlet	(-0.0002495297,0.0010238231)	Sí
Paperback	(0.3567886,0.3880817)	Sí
Plastic-comb	(-0.0002495297,0.0010238231)	Sí
Spiral-bond	(-0.0001259286 ,0.0016745155)	Sí

*Apéndice 5.4

Finalmente el último tema a explorar en este sección, se enuncia en un problema de intervalos de confianza para la diferencia de medias, se deben considerar las las subpoblaciones generadas en el desarrollo del tema 1, el problema dice:

> Estime los intervalos de confianza del 85% para todas las posibles comparaciones de medias. Interprete cada intervalo en el contexto de su problema. Escriba los supuestos necesarios para sus estimaciones. Si conoce las diferencias de medias poblacionales, considérelas para sus interpretaciones.

Supuestos: sean Y_i y Y_j con $i \neq j$ dos poblaciones con una muestra aleatoria para cada una, ya que sea desea conocer un intervalo de confianza para la diferencia de medias y la varianza poblacional es desconocida la cantidad pivotal será de:

$$T = \frac{(\overline{Y}_i - \overline{Y}_j) - (\mu_i - \mu_j)}{s_p\sqrt{\frac{1}{n_i} + \frac{1}{m_j}}} \Rightarrow (\mu_i - \mu_j) \in (\overline{Y}_i - \overline{Y}_j) \pm t_{\frac{\alpha}{2}, n+m-2} s_p \sqrt{\frac{1}{n_i} + \frac{1}{m_j}}$$

Con ayuda de un script de R, se obtuvo la siguiente información:

Tabla 5.2: Intervalos de confianza para la diferencia de medias con una confianza del 85%:

i,j	Intervalo de confianza para $\mu_i - \mu_j$
Hardcover ,Paperback	(4.956782,6.724532)
Hardcover, Board book	(12.11359,16.90956)
Hardcover, Mass Market Paperback	(5.876786,10.744815)
Hardcover, Cards	(0.3262618,25.7052184)
Hardcover, Spiral-bound	(-27.023811,4.065291)
Hardcover, Hardcover-spiral	(-13.27565,17.80713)
Hardcover, Audio CD	(-20.609354,1.500835)
Paperback, Board book	(6.480256,10.861577)
Paperback, Mass Market Paperback	(0.2309989,4.7092879)
Paperback, Cards	(-4.417972,18.768139)
Paperback, Spiral-bound	(-31.523666,-3.116167)
Paperback, Hardcover-spiral	(-17.77348,10.62365)
Paperback, Audio CD	(-25.542216,-5.247617)
Board book, Mass Market Paperback	(-7.556998,-4.844548)
Board book, Cards	(-1.859580,-1.132087)
Board book, Spiral-bound	(-27.38784.-24.59383)
Board book, Hardcover-spiral	(-12.69144,-11.80023)
Board book, Audio CD	(-29.15318,-18.97849)
Mass Market Paperback, Cards	(-2.536306,11.946186)
Mass Market Paperback,Spiral-bound	(-28.76046,-10.81966)
Mass Market Paperback, Hardcover-spiral	(-14.915968,2.825847)
Mass Market Paperback, Audio CD	(-25.95285, -9.777264)
Cards, Spiral-bound	(-36.33517,-12.65483)

Cards, Hardcover-spiral	(-10.75,-10.75)
Cards, Audio CD	(-59.33855,14.19855)
Spiral-bound, Hardcover-spiral	(-5.092337,32.582337)
Spiral-bound, Audio CD	(-47.6713,51.5213)
Hardcover-spiral, Audio CD	(-60.5949,36.9549)

*Apéndice 5.6

Debido a que eran bastantes categorías (12) la cantidad de combinaciones posibles es una lista bastante extensa, para este análisis se consideraron las muestras con longitud mayor a 1, puesto que al calcular la varianza, esta divertiría con un valor de muestra de 1.

¿Puede concluir que las varianzas poblacionales son iguales? Justifique mediante intervalos de confianza del 97%. Escriba los supuestos necesarios para sus estimaciones.

Los supuestos son los mismos que se realizaron en el inciso anterior. Con el cambio de parámetro para α para la construcción de los intervalos, se obtuvo la siguiente tabla:

Tabla 5.3: Intervalos de confianza para la diferencia de medias con una confianza del 97%:

i,j	Intervalo de confianza para $\mu_i - \mu_j$	¿$\mu_i = \mu_j$?
Hardcover ,Paperback	(4.507851,7.173463)	$\mu_i > \mu_j$
Hardcover, Board book	(10.89505,18.12810)	$\mu_i > \mu_j$
Hardcover, Mass Market Paperback	(4.639936,11.981664)	$\mu_i > \mu_j$
Hardcover, Cards	(-6.122385,32.153865)	$\mu_i = \mu_j$
Hardcover, Spiral-bound	(-34.92338,11.96486)	$\mu_i = \mu_j$
Hardcover, Hardcover-spiral	(-21.17361,25.70509)	$\mu_i = \mu_j$
Hardcover, Audio CD	(-26.227421,7.118901)	$\mu_i = \mu_j$
Paperback, Board book	(5.366413,11.975420)	$\mu_i > \mu_j$
Paperback, Mass Market Paperback	(-0.9074981,5.8477849)	$\mu_i = \mu_j$
Paperback, Cards	(-10.31340,24.66357)	$\mu_i = \mu_j$

Paperback, Spiral-bound	(-38.746730,4.106897)	$\mu_i = \mu_j$
Paperback, Hardcover-spiral	(-24.99391,17.84407)	$\mu_i = \mu_j$
Paperback, Audio CD	(-30.70242311,-0.08741057)	$\mu_i < \mu_j$
Board book, Mass Market Paperback	(-8.253513,-4.148033)	$\mu_i < \mu_j$
Board book, Cards	(-2.0484992,-0.9431674)	$\mu_i < \mu_j$
Board book, Spiral-bound	(-28.11360,-23.86806)	$\mu_i < \mu_j$
Board book, Hardcover-spiral	(-12.92294,-11.56873)	$\mu_i < \mu_j$
Board book, Audio CD	(-31.79469,-16.33698)	$\mu_i < \mu_j$
Mass Market Paperback, Cards	(-6.298246,15.708125)	$\mu_i = \mu_j$
Mass Market Paperback, Spiral-bound	(-33.422041,-6.158079)	$\mu_i < \mu_j$
Mass Market Paperback, Hardcover-spiral	(-19.525847,7.435726)	$\mu_i = \mu_j$
Mass Market Paperback, Audio CD	(-30.153431,-5.576689)	$\mu_i < \mu_j$
Cards, Spiral-bound	(-48.4670345,-0.5229655)	$\mu_i < \mu_j$
Cards, Hardcover-spiral	(-10.75,-10.75)	$\mu_i < \mu_j$
Cards, Audio CD	(-87.54222,42.40222)	$\mu_i = \mu_j$
Spiral-bound, Hardcover-spiral	(-32.83614,60.32614)	$\mu_i = \mu_j$
Spiral-bound, Audio CD	(-90.05056,93.90056)	$\mu_i = \mu_j$
Hardcover-spiral, Audio CD	(-102.27229,78.63229)	$\mu_i = \mu_j$

*Apéndice 5.6 con valor para alpha de 0.03 en la función i.c(alpha)

Se concluye que algunas medias son iguales debido a que el intervalo contiene al número cero, esta condición sólo se cumple cuando $\mu_i = \mu_j$, por lo que se concluye esta igualdad para el intervalo, si todos los elementos del intervalo son positivos, se debe a que $\mu_i > \mu_j$ ya que la diferencia de $\mu_i - \mu_j > 0$, de forma análoga, si todos los elementos son negativos, esto se deba a que

Tema 6: Pruebas de hipótesis paramétricas

1. Contraste de Hipótesis bilateral

De una muestra aleatoria de 600 libros vendidos dentro de la plataforma Amazon.com, reveló un precio promedio de 11.95 USD. Suponiendo que el precio de los libros tiene una distribución normal, con media $\mu = 12.35684$ y varianza $\sigma^2 = 219.6963$, por lo que ¿esto parece indicar que el precio medio de los libros no sea de 12.35684? Se considera con un nivel de significancia del 2.5%.

*Promedio de muestra aleatoria de 600 libros vendidos (Apéndice B.6.1)

-Hipótesis nula H_0: La media de los libros es de 12.35684

-Hipótesis alternativa H_a: La media de los libros no es de 12.35684

-Hipótesis estadísticas:

$H_0: \mu = 12.35684$

$H_a: \mu \neq 12.35684$

Supuestos:

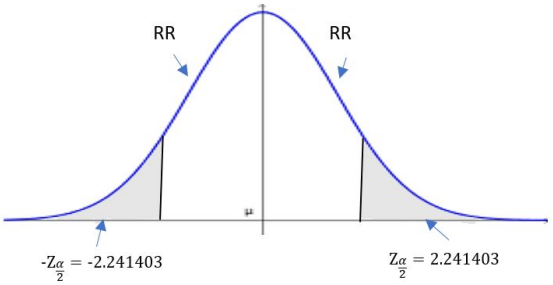
X_2 = Precio de libros

$X_2 \sim N(\mu, \sigma^2 = 219.6963)$

Estadístico de prueba:

$$Z = \frac{\overline{X} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)_0} \sim^{H_0} N(0,1)$$

$$Z = \frac{11.95 - 12.35684}{\left(\frac{14.82215571}{\sqrt{600}}\right)} = -0.672338374$$



Región Crítica (RR):

$$(I)\ RR = \left\{ Z \mid Z \leq -Z_{\frac{\alpha}{2}} = -2.241403 \text{ ó } Z \geq Z_{\frac{\alpha}{2}} = 2.241403 \right\}$$

$$como\ Z_{calc} = -0.672338374 < Z_{crit} = 2.241403 \Rightarrow no\ se\ rechaza\ H_0: \mu = 12.35684$$

$$(II)\ RR = \left\{ \overline{X} \mid \overline{X} \leq \mu_0 - Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = 12.35684 - 2.241403\left(\frac{14.82215571}{\sqrt{600}}\right) = 11.00054021 \text{ ó } \right\}$$

$$\overline{X} \geq \mu_0 + \left(\frac{\sigma}{\sqrt{n}}\right) = 12.35684 + 2.241403\left(\frac{14.82215571}{\sqrt{600}}\right) = 13.71313979\}$$

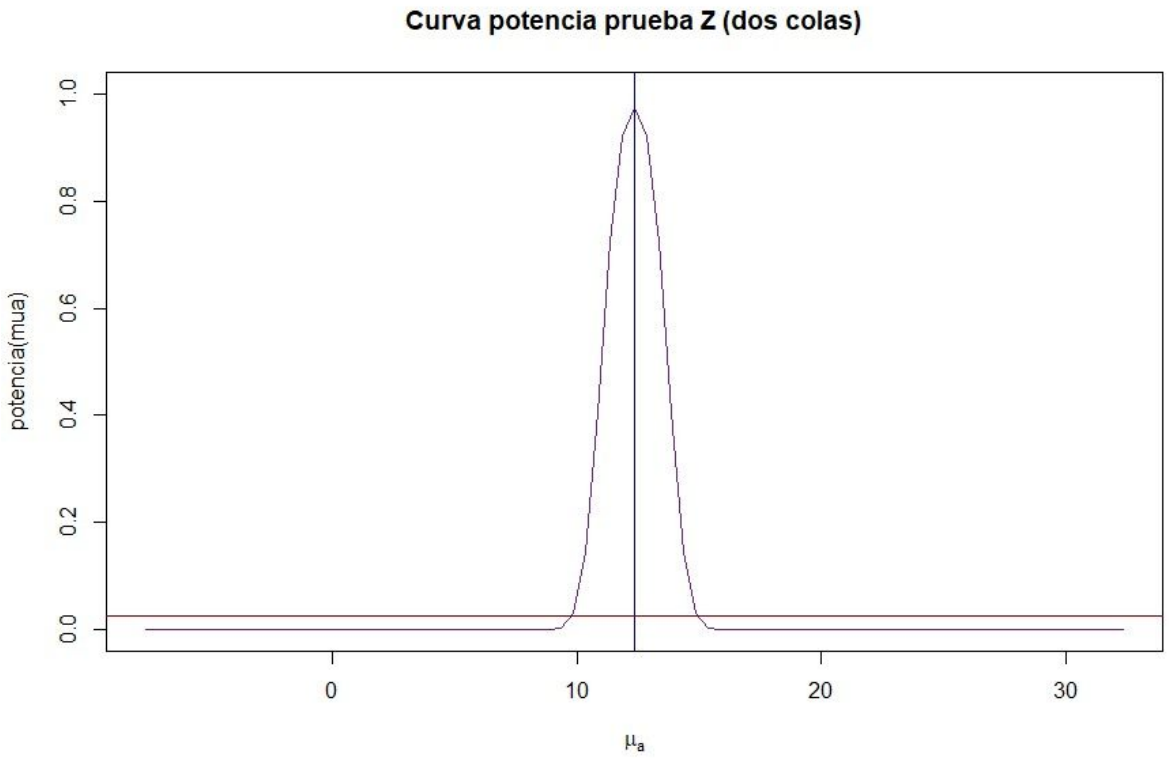
$$como\ \overline{X} = 12.35684 < 13.71313979 \Rightarrow no\ se\ rechaza\ H_0: \mu = 12.35684$$

Valor -p: 0.2506842*2= 0.5013684 *dato obtenido en R (Apéndice B.6.2)

Se concluye que la media de los libros es igual a 12.35684 de una muestra aleatoria de 600 libros vendidos de la plataforma Amazon.com

2. Curva potencia

Considerando la prueba bilateral anterior, dado que se conoció la región crítica, se propuso un conjunto de 10 valores distintos para la media según la hipótesis alternativa. Por lo cual se graficó la curva de potencia:



*datos obtenidos en R (Apéndice B.6.3)

De acuerdo a la gráfica podemos observar que entre más se alejan los valores de la alternativa μ_a de los valores nulos μ_0 , la potencia disminuye por lo que la probabilidad de equivocarse aumenta.

3. Comparación de dos poblaciones:

Tabla 6.1: Regiones de rechazo de la diferencia de medias con un nivel de significancia al 15%:

i,j	RR a dos colas para $\mu_i - \mu_j$	Rechazo H_0
Hardcover ,Paperback	$\mu_x - \mu_x \{T \leq -1.44$ o $T \geq 1.44\}$	No
Hardcover, Board book	$\mu_x - \mu_x \{T \leq -1.4403$ o $T \geq 1.4403\}$	No
Hardcover, Mass Market Paperback	$\mu_x - \mu_x \{T \leq -1.4403$ o $T \geq 1.4403\}$	No
Hardcover, Cards	$\mu_x - \mu_x \{T \leq -1.4403$ o $T \geq 1.4403\}$	No

Hardcover, Spiral-bound	$\mu_x - \mu_x \{T \leq -1.4403 \text{ o } T \geq 1.4403\}$	No
Hardcover, Hardcover-spiral	$\mu_x - \mu_x \{T \leq -1.4403 \text{ o } T \geq 1.4403\}$	No
Hardcover, Audio CD	$\mu_x - \mu_x \{T \leq -1.4403 \text{ o } T \geq 1.4403\}$	No
Paperback, Board book	$\mu_x - \mu_x \{T \leq -1.4406 \text{ o } T \geq 1.4406\}$	No
Paperback, Mass Market Paperback	$\mu_x - \mu_x \{T \leq -1.4406 \text{ o } T \geq 1.4406\}$	No
Paperback, Cards	$\mu_x - \mu_x \{T \leq -1.4407 \text{ o } T \geq 1.4407\}$	No
Paperback, Spiral-bound	$\mu_x - \mu_x \{T \leq -1.4407 \text{ o } T \geq 1.4407\}$	No
Paperback, Hardcover-spiral	$\mu_x - \mu_x \{T \leq -1.4407 \text{ o } T \geq 1.4407\}$	No
Paperback, Audio CD	$\mu_x - \mu_x \{T \leq -1.4407 \text{ o } T \geq 1.4407\}$	No
Board book, Mass Market Paperback	$\mu_x - \mu_x \{T \leq -1.4463 \text{ o } T \geq 1.4463\}$	No
Board book, Cards	$\mu_x - \mu_x \{T \leq -1.4527 \text{ o } T \geq 1.4527\}$	No
Board book, Spiral-bound	$\mu_x - \mu_x \{T \leq -1.4528 \text{ o } T \geq 1.4528\}$	No
Board book, Hardcover-spiral	$\mu_x - \mu_x \{T \leq -1.4528 \text{ o } T \geq 1.4528\}$	No
Board book, Audio CD	$\mu_x - \mu_x \{T \leq -1.4525 \text{ o } T \geq 1.4525\}$	No
Mass Market Paperback, Cards	$\mu_x - \mu_x \{T \leq -1.4528 \text{ o } T \geq 1.4528\}$	No
Mass Market Paperback, Spiral-bound	$\mu_x - \mu_x \{T \leq -1.453 \text{ o } T \geq 1.453\}$	No
Mass Market Paperback, Hardcover-spiral	$\mu_x - \mu_x \{T \leq -1.453 \text{ o } T \geq 1.453\}$	No
Mass Market Paperback, Audio CD	$\mu_x - \mu_x \{T \leq -1.4527 \text{ o } T \geq 1.4527\}$	No
Cards, Spiral-bound	$\mu_x - \mu_x \{T \leq -1.9243 \text{ o } T \geq 1.9243\}$	No
Cards, Hardcover-spiral	$\mu_x - \mu_x \{T \leq -1.9243 \text{ o } T \geq 1.9243\}$	No
Cards, Audio CD	$\mu_x - \mu_x \{T \leq -1.6994 \text{ o } T \geq 1.6994\}$	No
Spiral-bound, Hardcover-spiral	$\mu_x - \mu_x \{T \leq -2.2819 \text{ o } T \geq 2.2819\}$	No
Spiral-bound, Audio CD	$\mu_x - \mu_x \{T \leq -1.7782 \text{ o } T \geq 1.7782\}$	No
Hardcover-spiral, Audio CD	$\mu_x - \mu_x \{T \leq -1.7782 \text{ o } T \geq 1.7782\}$	No

*Apéndice 6.4

Apéndice:

. . .

Tema 1:

1.1. Eliminar valores nulos:

```
library(readr)
bestsellers <- read_csv("bestsellers.csv")
z<-!is.na(bestsellers$Reviews)&!is.na(bestsellers$`Price(USD)
`)&!is.na(bestsellers$Type)
x<-lapply(bestsellers, function(x) x[z])
df <- data.frame(Reviews = x$Reviews,
                 `Price(USD)` = x$`Price(USD)`,
                  Type = x$Type)
write.csv(df,'bestsellers_without_naValues.csv')
```

1.2. Resumen de los datos:

```
summary(bestsellers)
```

-Desviación estándar:

```
sd(bestsellers_without_naValues$Reviews)
sd(bestsellers_without_naValues$Price(USD))
```

-Modas:

```
install.packages("modeest")
library(modeest)
mfv(bestsellers_without_naValues$Reviews)
mfv(bestsellers_without_naValues$Price(USD))
mfv(bestsellers_without_naValues$Type)
```

1.3.Coef Asimetría:

```
skewness(bestsellers_without_naValues$Price.USD.)
skewness(bestsellers_without_naValues$Reviews)
```

-Kurtosis

```
kurtosis(bestsellers_without_naValues$Price.USD.)
kurtosis(bestsellers_without_naValues$Reviews)
```

1.4. Covarianza:

```
cov(bestsellers_without_naValues$Price.USD.,
bestsellers_without_naValues$Reviews)
```

-Correlación:

```
cor(bestsellers_without_naValues$Price.USD.,
bestsellers_without_naValues$Reviews)
```

Tema 2:

2.1. Muestreo Aleatorio Simple (MAS) de variable x2(Price):

```
library(readr)
library(dplyr)

bests_n <- read_csv("bestsellers_without_naValues.csv")
head(bests_n)
require(dplyr)
muestreo1 <- sample_n(bests_n,size=6)
nrow(muestreo1)
muestreo1
```

2.2. Muestras posibles de tamaño 3

```
library(gtools)
muestra <- c(11.9, 0.25, 5.43, 4.86, 13.9, 21.0)
posib_m <- combinations(6,3,muestra)
posib_m
```

2.3. Media muestral de las posibles muestras

```
library(gtools)
muestra <- c(11.9, 0.25, 5.43, 4.86, 13.9, 21.0)
posib_m <- combinations(6,3,muestra)
posib_m
mues_df <- data.frame(posib_m)
mues_df
mues_df$Media_muestral <- rowMeans(mues_df)
mues_df
```

2.4. Varianza muestral de las posibles muestras

```
#Se realizó la varianza muestra por muestra
mue <- c(0.25, 4.86, 5.43)
var(mue)
.
.
.
mue <- c(11.9, 13.9, 21.0)
var(mue)
```

2.5. Valor esperado de $\overline{X}$

```
library(gtools)
muestra <- c(11.9, 0.25, 5.43, 4.86, 13.9, 21.0)
posib_m <- combinations(6,3,muestra)
posib_m
mues_df <- data.frame(posib_m)
mues_df
mues_df$Media_muestral <- rowMeans(mues_df)
mues_df
Verif <- colSums(mues_df)
```

```
Verif
#obtenemos el valor de la suma de la media muestral y lo
dividimos entre 20
```

2.6. Muestreo Aleatorio Simple para X_3

```
library(readr)
library(dplyr)
bests_n <- read_csv("bestsellers_without_naValues.csv")
head(bests_n)
require(dplyr)
muestreon <- sample_n(bests_n,size=300)
nrow(muestreon)
muestreon
#Quedarse con columna de Reviews
colrev <- select(muestreon,Reviews)
colrev
```

2.7. `mean(muestreon$Reviews)`

Tema 5:

5.1 Proporción de cada tipo de libro

```
library(readr)
bestsellers <- read_csv("bestsellers_without_naValues.csv")
tipos<-table(bestsellers$Type)
n<-length(bestsellers$X1)
proporcion<-lapply(tipos, function(x) x/n)
```

5.2 Coeficiente de variación muestral para X_2 y X_3

```
library(readr)
bestsellers <- read_csv("bestsellers_without_naValues.csv")
x2<-bestsellers$Price.USD.
x3<-bestsellers$Reviews
x2mean<-mean(x2)
x2sd<-sd(x2)
x3mean<-mean(x3)
x3sd<-sd(x3)
cv<-c(x2sd/x2mean*100,x3sd/x3mean*100); cv
```

5.3 Gráfica de una posible distribución asociada para X_2:

```
library(readr)
bestsellers <- read_csv("bestsellers_without_naValues.csv")
x2<-bestsellers$Price.USD.

xb <- mean(x2)
s2 <- mean((x2-mean(x2))^2)
al_mom <- xb^2 / s2
```

```
la_mom <- xb / s2

log.vero <- function(theta) - sum(log(dgamma(x2, theta[1], theta[2])))
Lmax.1 <- nlm(log.vero, c(al_mom, la_mom))
Lmax.1

alMVnlm<-Lmax.1$estimate[1]
laMVnlm<-Lmax.1$estimate[2]

x <- seq(min(x2), max(x2), length = 300)

f<-dexp(x,rate=1/mean(x2))
fmom<-dgamma(x,shape=al_mom,rate=la_mom)
fMVnlm<-dgamma(x,shape=alMVnlm,rate=laMVnlm)

hist(x2,prob=TRUE,breaks = 30,main = "Histograma de X2",col="khaki")
lines(x, f, col = "red", lwd = 2)
lines(x, fmom, col = "black",lwd = 2)
lines(x, fMVnlm, col = "chartreuse4", lwd = 2)
legend("topright", c(expression(Exp(lambda)), paste(round(1/mean(x2),3))),
       lty = c(0, 1, 1), col = c("red"), box.lty = 0, lwd = 2)
legend("bottomright", c(expression(Gamma(paste(alpha,",",lambda))),
paste(round(al_mom,3),",",round(la_mom,3)),
       paste(round(alMVnlm,3),",",round(laMVnlm,3))),lty = c(0, 1, 1), col =
c("chartreuse4","black"), box.lty = 0, lwd = 2)
```

5.4 Cálculo de los valores poblacionales para X_2 y X_3:

```
library(readr)
bestsellers <- read_csv("bestsellers_without_naValues.csv")
x2<-bestsellers$Price.USD.
x3<-bestsellers$Reviews

N<-length(x2)
sumaX2<-sum(x2)
sumaX2Cuadrado<-sum(x2^2)
sumaX3<-sum(x3)
sumaX3Cuadrado<-sum(x3^2)

meanx2<-sumaX2/N
varx2<-(1/N)*(sumaX2Cuadrado-N*meanx2^2)
meanx3<-sumaX3/N
varx3<-(1/N)*(sumaX3Cuadrado-N*meanx3^2)
```

5.5 Intervalos de confianza para las proporciones de X_1:

```
library(readr)
bestsellers <- read_csv("bestsellers_without_naValues.csv")
x1<-bestsellers$Type
tipos<-table(x1)
n<-length(x1)

i.c<-lapply(tipos,function(x,alpha=0.1)
  c(x/n-qnorm(alpha/2,lower.tail = F)*sqrt((x/n*(1-(x/n)))/n),
    x/n+qnorm(alpha/2,lower.tail = F)*sqrt((x/n*(1-(x/n)))/n))
); i.c
```

5.6 Intervalos de confianza para la diferencia de muestras:

```
library(readr)
bestsellers <- read_csv("bestsellers_without_naValues.csv")
x1<-bestsellers$Type
x2<-bestsellers$Price.USD.
N<-length(x1)

combinaciones<-combn(x1[!duplicated(x1)], m=2, simplify=FALSE)

for(i in combinaciones){

    elemento_1<-x2[x1==i[1]]
    elemento_2<-x2[x1==i[2]]
    mean_elemento_1<-mean(elemento_1)
    mean_elemento_2<-mean(elemento_2)
    var_elemento_1<-var(elemento_1)
    var_elemento_2<-var(elemento_2)
    n<-length(elemento_1)
    m<-length(elemento_2)
    s2_pooled<-((n-1)*var_elemento_1+(m-1)*var_elemento_2)/(n+m-2)
    i.c<-function(alpha=0.15)
      c(mean_elemento_1-mean_elemento_2-qt(p=alpha/2,df=n+m-2,lower.tail =
  F)*sqrt(s2_pooled)*sqrt(1/n+1/m),
        mean_elemento_1-mean_elemento_2+qt(p=alpha/2,df=n+m-2,lower.tail =
  F)*sqrt(s2_pooled)*sqrt(1/n+1/m))

    i.c_iteracion<-i.c(0.15)

    if(!is.na(sum(i.c_iteracion))){
      print(paste(i[1],i[2],sep = ","))
      print(i.c_iteracion)
    }
  }
```

Tema 6:

6.1 Promedio de muestra aleatoria de 600 libros vendidos

```
library(readr)
library(dplyr)

bests_n <- read_csv("C:/Users/karen/Desktop/Estadística
I/PROYECTO/Semiannual_Statistical_I_Project-main/bestsellers_
without_naValues.csv")
head(bests_n)
require(dplyr)

set.seed(04022021)
muestraph <- sample_n(bests_n,size=600)
nrow(muestraph)
muestraph

mean(muestraph$Price.USD.)
```

6.2 Valor-p

```
pnorm(-0.672338374, lower.tail=T)
```

6.3 Curva potencia

```
n<-600
mu0<-12.35684
var0<-219.6963
alpha<-0.025

RR_Izquierda<-mu0-qnorm(p=alpha/2,lower.tail =
F)*sqrt(var0/n)
RR_Derecha<-mu0+qnorm(p=alpha/2,lower.tail = F)*sqrt(var0/n)

mua<-seq(-20+mu0,20+mu0,by=0.5)

potencia <- function(x){
  beta1<-pnorm(q=(RR_Izquierda-mua)/sqrt(var0/n),lower.tail =
T)
  beta2<-pnorm(q=(RR_Derecha-mua)/sqrt(var0/n),lower.tail =
F)
1-(beta1+beta2)
}

plot(mua,potencia(mua),type="l",ylim=c(0,1),
     main="Curva potencia prueba Z (dos colas)",
xlab=expression(mu[a]),col="darkorchid4")
abline(v=mu0,col="navyblue")
abline(h=alpha,col="red4")
```

6.4 Prueba de hipótesis bilateral para la diferencia de medias

```
library(readr)
bestsellers <- read_csv("bestsellers_without_naValues.csv")
x1<-bestsellers$Type
x2<-bestsellers$Price.USD.
N<-length(x1)

combinaciones<-combn(x1[!duplicated(x1)], m=2,
simplify=FALSE)

for(i in combinaciones){

  elemento_1<-x2[x1==i[1]]
  elemento_2<-x2[x1==i[2]]
  mean_elemento_1<-mean(elemento_1)
  mean_elemento_2<-mean(elemento_2)
  var_elemento_1<-var(elemento_1)
  var_elemento_2<-var(elemento_2)
  n<-length(elemento_1)
  m<-length(elemento_2)
s2_pooled<-((n-1)*var_elemento_1+(m-1)*var_elemento_2)/(n+m-2
)

T_calc<-(mean_elemento_1-mean_elemento_2)/(s2_pooled*sqrt(1/n
+1/m))
  RR<-function(alpha=0.15)
c(-qt(p=alpha/2,df=n+m-2,lower.tail =
F),qt(p=alpha/2,df=n+m-2,,lower.tail = F))

  RR_iteracion<-i.c(0.15)

  if(n>1&m>1&!is.na(sum(RR_iteracion))){
    print(paste(i[1],i[2],sep = ","))
    print(paste("T_calc:",T_calc))

print(paste("RR:",expression(mu[x]-mu[y]),"<=",round(RR_iteracion[
1],4),"   ;
",expression(mu[x]-mu[y]),">=",round(RR_iteracion[2],4)))
  }
}
```

Extras:

1.Datos Obtenidos:

```
https://www.amazon.com/-/es/gp/bestsellers/{año_venta}/books/
ref=zg_bsar_pg_2?ie=UTF8&language=en_US&pg={página=1/2}
```

2. Repositorio:

`https://github.com/richi1325/Semiannual_Statistical_I_Project.git`

3.Tableau Reader

`(Software para visualización e interacción de gráficas)`
`https://www.tableau.com/es-mx/products/reader`
`Archivo de Github Amazon.twbx`
`https://github.com/richi1325/Semiannual_Statistical_I_Project/blob/main/Amazon.twbx`

4.Dataset:

Mendoza, J. R. (4 de febrero de 2021). *Amazon top 100 bestselling books by year*. Obtenido de Kaggle: https://www.kaggle.com/ricardomendozavillar/amazon-top-100-bestselling-books-1995-2020

5.Complementarias:

Santibáñez, J. (19 de octubre de 2017). *Estimación por máxima verosimilitud*. Obtenido de Departamento de Probabilidad y estadística IIMAS UNAM: http://sigma.iimas.unam.mx/jsantibanez/Cursos/Inferencia/2018_1/Ejemplos/02_EMV.html