

II Jornadas Internacionales de Estadística Aplicada 5 y 6 de Diciembre de 2019

Análisis de ventas en la herrería “Hierro y Barro”

Autor: Estanislao Escudero

Institución: Facultad de Ingeniería, Universidad Católica de Salta. Salta Capital.

Datos de contacto: tanoescudero13@gmail.com; teléfono: 3875336819

Resumen

Un aspecto esencial en todo local comercial es la determinación del volumen de ventas. Tener conocimiento acerca de la demanda que experimenta el local comercial permite una efectiva planificación a largo y corto plazo, una organización del negocio y una proyección del mismo. En este trabajo se muestra cómo es posible utilizar las herramientas estadísticas para dichos fines y con ello poder diseñar un plan de producción. En el trabajo se realiza un análisis descriptivo de un lote de datos extrayendo características generales del mismo. Además se utilizaron métodos bayesianos de inferencia estadística para estimar parámetros poblacionales y se efectuó una comparación con los métodos de inferencia clásicos. El complemento de la estadística descriptiva y la inferencial permite efectuar un análisis integral de la situación del local comercial.

Palabras clave: ventas, muestreo, estadística, descripción, bayesiano.

Introducción

En este trabajo se desarrolló un análisis de las ventas producidas en la herrería “Hierro y Barro”, local comercial que se dedica a la venta de muebles e iluminaria de hierro forjado. Para entender por qué se realizó este trabajo es necesario comprender también la estructura general de ventas en la herrería. El aspecto más destacable en este sentido es que gran parte de las ventas se producen por encargo; es decir que el cliente encarga la fabricación del artículo que desea adquirir y tras un cierto tiempo-establecido en función de la complejidad del artículo y de la disponibilidad para su fabricación- se fabrica y entrega el producto. Además, no solo se venden artículos presentes en el catálogo de Hierro y Barro sino que también se venden artículos fuera de catálogo encargados por particulares. Se seleccionaron para el análisis 3 categorías de principal interés (arañas, faroles y apliques) por representar estos el mayor volumen de ventas y además se incluyó un breve análisis de los artículos “fuera de catálogo”.

Los datos usados en el análisis son el número de ventas mensuales de los artículos antes mencionados durante los últimos 3 años. En este trabajo se realizó, en primer lugar, un análisis descriptivo de dichos datos con lo que se buscó conseguir una imagen general del negocio a partir de la información directa provista por los datos al organizarlos.

Además en este trabajo se realizaron inferencias puntuales y por intervalo de ciertos parámetros utilizando la estadística Bayesiana. La principal razón que justifica el uso de métodos bayesianos es el amplio conocimiento que se tiene sobre las ventas producidas en el negocio. La herrería funciona hace más de 30 años y es por eso que los datos a priori con los que se cuenta, si bien no resultan de un análisis formal y son de cierta manera subjetivos, tienen una fundamentación muy sólida basada en la observación a largo plazo. Por ende no solo se justifica sino que en mi opinión sería una forma incorrecta de abordar el problema de la estimación el no utilizar dicha información.

La motivación del trabajo se presenta de manera natural; se deseaba saber cuál es la cantidad esperada de ventas de ciertos artículos de interés para poder planificar adecuadamente la producción. Esto no solo haría posible reducir los plazos de entrega de los productos sino que también otorga la capacidad de planificación y proyección a futuro del negocio.

Metodología

Todo el trabajo se fundamenta principalmente en la recolección de los datos de ventas de los artículos en cuestión. Como ya se explicó los artículos de interés fueron faroles, apliques y arañas debido a que estos suponen el mayor volumen de ventas en el local comercial. La información recolectada representa la cantidad de artículos de cada tipo vendidos mensualmente durante los últimos 3 años. Elegí este período para mantener la información lo más actualizada posible, de manera tal que el estudio refleje las verdaderas tendencias de compra al tiempo que se cuente con información suficiente como para

realizar un análisis estadístico significativo. Por otra parte, se consideraron las ventas mensuales porque como ya se explicó, se busca que el trabajo permita la planificación de la producción. La contemplación de un periodo de tiempo menor, como por ejemplo una semana, no solo trae consigo una enorme variabilidad sino que no permitirá una correcta organización. Lo mejor es planificar la producción en función de las ventas mensuales y almacenar los artículos en stock, para así poder hacer frente a la demanda en el negocio y reducir el tiempo de entrega de los artículos.

Un aspecto a destacar es que cada categoría de artículos incluye muchos modelos de artículos diferentes. Por ejemplo la categoría “faroles” es sumamente general ya que incluye dentro de sí una gran variedad de faroles de distinta clase. Lo mismo ocurre con las demás categorías. No se ha realizado una distinción en los artículos con el objetivo de que el trabajo no pierda generalidad. Cabe destacar que el monto de ventas en cada categoría es, en función de lo anterior, la suma de la cantidad de ventas de cada uno de los artículos que integra dicha categoría.

En primer lugar se presentarán los datos tal como fueron recolectados y luego se los organizará utilizando tablas de frecuencias, gráficos de bastones, histogramas, gráficos de caja y gráficos de barras¹. Posteriormente se efectúa un análisis cualitativo de la información así presentada.

En una segunda etapa del trabajo se realizan inferencias utilizando métodos bayesianos. Específicamente se realiza una estimación puntual y una estimación por intervalo para la media de la cantidad de faroles vendidos mensualmente (suponiendo una distribución a priori normal para la media poblacional y por tomar una muestra de tamaño mayor a 30 también se considerará que la verosimilitud tiene una distribución normal). Se incluye también una estimación puntual para la proporción de artículos fuera de catálogo vendidos, usando la estadística bayesiana y siguiendo un modelo beta-binomial.

Desarrollo

Primer Parte: Estadística Descriptiva.

En la siguiente tabla se exponen los datos tal como fueron recolectados a partir de las planillas de ventas. Incluyen los datos de ventas mensuales de los años 2017, 2018 y 2019. Sin embargo es destacable que los datos de 2019 fueron recolectados hasta el mes de Octubre puesto que al momento de realización del trabajo no se contaba con información sobre los meses siguientes.

Datos de ventas mensuales de los artículos seleccionados en el período 2017-2019														
Año 2019					Año 2018					Año 2017				
Mes	Artículo				Mes	Artículo				Mes	Artículo			
	Faroles	Apliques	Arañas	Artículos fuera de catálogo		Faroles	Apliques	Arañas	Artículos fuera de catálogo		Faroles	Apliques	Arañas	Artículos fuera de catálogo
1	28	7	5	1	1	28	17	6	6	1	55	12	9	0
2	76	10	0	8	2	29	0	0	8	2	49	20	9	1
3	145	5	6	4	3	27	0	4	4	3	71	2	0	9
4	18	22	0	5	4	33	24	8	4	4	14	12	5	1
5	114	6	7	6	5	18	13	6	2	5	45	41	12	6
6	35	10	6	5	6	10	0	0	3	6	31	34	4	3
7	133	0	0	5	7	34	16	4	5	7	50	34	24	2
8	73	40	6	6	8	54	11	2		8	41	1	4	3
9	35	29	11	3	9	47	12	6	1	9	50	14	3	11
10	34	0	4	25	10	26	14	5	2	10	58	7	4	7
11					11	26	35	0	3	11	48	5	0	9
12					12	58	0	11	4	12	56	7	1	4

A continuación realizaré un breve análisis descriptivo del conjunto de datos obtenidos.

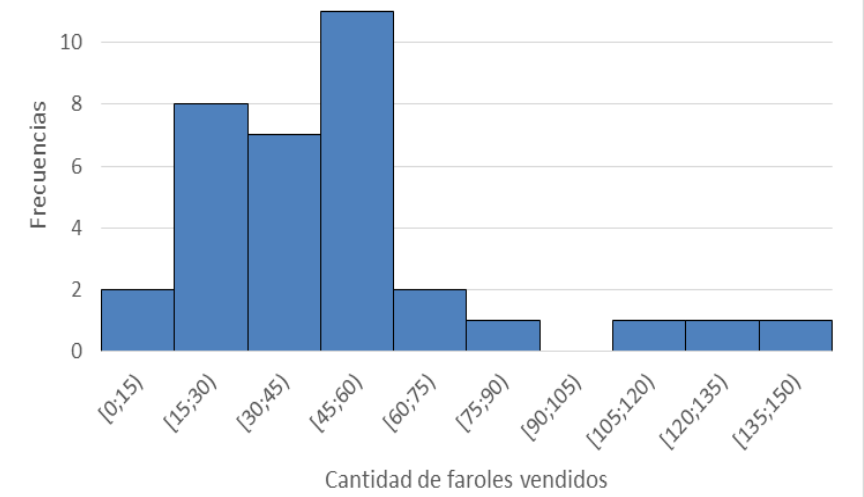
La primer categoría a considerar es la de faroles. La media de los datos sin agrupar es $\bar{x}_F = \frac{1}{34} \sum_{i=1}^{34} x_i = 48,5$ faroles por mes. La mediana de los datos ordenados sin agrupar es $\widetilde{x}_F = \frac{(x_{17}+x_{18})}{2} = 43$ faroles. Sin agrupar, los datos son plurimodales. La varianza de los datos es de $s^2 = \frac{1}{34} \sum_{i=1}^{34} (x_i - \bar{x}_F)^2 = 956,076$ y la desviación estándar es de $s = \sqrt{s^2} = 30,92$ faroles por mes. A partir de estos valores puede decirse que la distribución de los datos es muy dispersa por tener una desviación

¹ Todas las tablas y gráficos de este trabajo fueron realizados con el software “Microsoft Excel”.

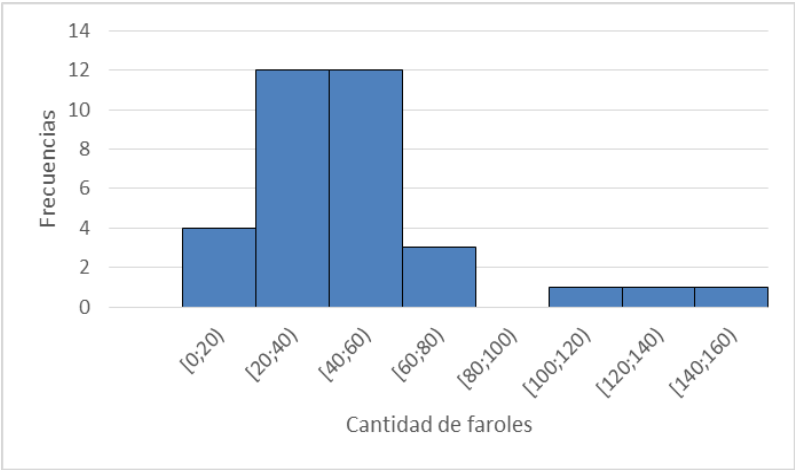
estándar relativamente elevada. Sin embargo la relativa cercanía entre la media y la mediana podría sugerir una distribución aproximadamente simétrica. Para complementar este análisis procedo a realizar un análisis gráfico del conjunto de datos.

Debido a la gran cantidad de categorías se han utilizado histogramas en lugar de gráficos de bastones con el objeto de mejorar la visibilidad de la distribución de datos. Además debido a la alta variabilidad de la cantidad de artículos vendidos en esta categoría he realizado distintos histogramas con diferente longitud de intervalo de clases a fin de observar mejor la distribución de ventas. A continuación se presenta la tabla de frecuencias del número de meses en los cuales se vendió una cantidad determinada de faroles, acompañada por su correspondiente histograma.

Faroles-Tabla de frecuencias		
Cantidad de faroles	Frecuencia	Frecuencia acumulada
[0;20)	4	4
[20;40)	12	16
[40;60)	12	28
[60;80)	3	31
[80;100)	0	31
[100;120)	1	32
[120;140)	1	33
[140;160)	1	34
Totales	34	34



Del histograma podemos ver que la distribución está concentrada entre 15 y 60 faroles por mes y no se puede extraer información clara sobre la simetría de la distribución. A continuación se presenta un segundo histograma generado con intervalos de clase de diferente longitud.



Faroles-Tabla de frecuencias		
Cantidad de faroles	Frecuencia	Frecuencia acumulada
[0;15)	2	2
[15;30)	8	10
[30;45)	7	17
[45;60)	11	28
[60;75)	2	30
[75;90)	1	31
[90;105)	0	31
[105;120)	1	32
[120;135)	1	33
[135;150)	1	34
Totales	34	34

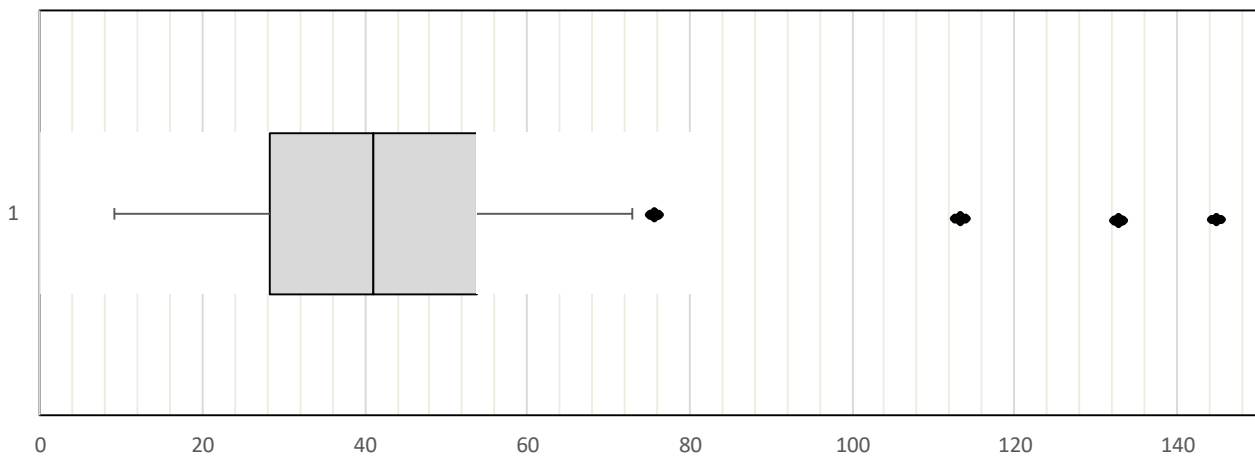
Del histograma anterior podemos ver que la distribución está concentrada entre 20 y 60 faroles por mes. En este gráfico podemos apreciar que la forma general de la distribución es semejante a la del histograma anterior. Así, podemos comenzar a considerar una distribución aproximadamente simétrica respecto a los 42 faroles por mes.

En primer lugar quiero destacar que se realizaron múltiples histogramas con intervalos de clase de diferente longitud con el objetivo de obtener una visualización correcta de la distribución puesto que son las características esenciales de la misma las que permanecen de un gráfico a otro. A partir de los gráficos anteriores, resulta evidente que la mayor cantidad de faroles vendidos está concentrada entre los 25 y 60 faroles aproximadamente. Por otra parte no podemos concluir nada acerca de la simetría de la distribución. El primer y segundo histograma nos sugieren cierta simetría con respecto a los 42 faroles por mes aunque no es claro y podría tratarse de asimetría positiva. Lo que sí resulta evidente es que la

cantidad de meses en los que se vendieron más de 70 faroles y menos de 15 fueron muy escasos y constituyen eventos extraños.

Como complemento final del análisis gráfico presento un diagrama de caja de la cantidad de meses en los que se vendió una cantidad determinada de faroles.

Diagrama de caja- Ventas de faroles-



En el eje horizontal se encuentra representada la cantidad de faroles vendidos en un mes determinado. Del boxplot podemos analizar las 5 medidas resúmenes de la distribución. Estas son la mediana, de 43 faroles por mes; los cuartiles (el cuartil inferior de 28,25 faroles y el cuartil superior de 55,75 faroles); el rango intercuartil, de 12,75 faroles; y el rango de la distribución (el mínimo observado es de 10 faroles y el máximo de 145 faroles siendo el rango de 135 faroles). Quiero destacar que la longitud máxima admitida de los “bigotes” es de 1,5 RI, es decir 1,5 del rango intercuartil.

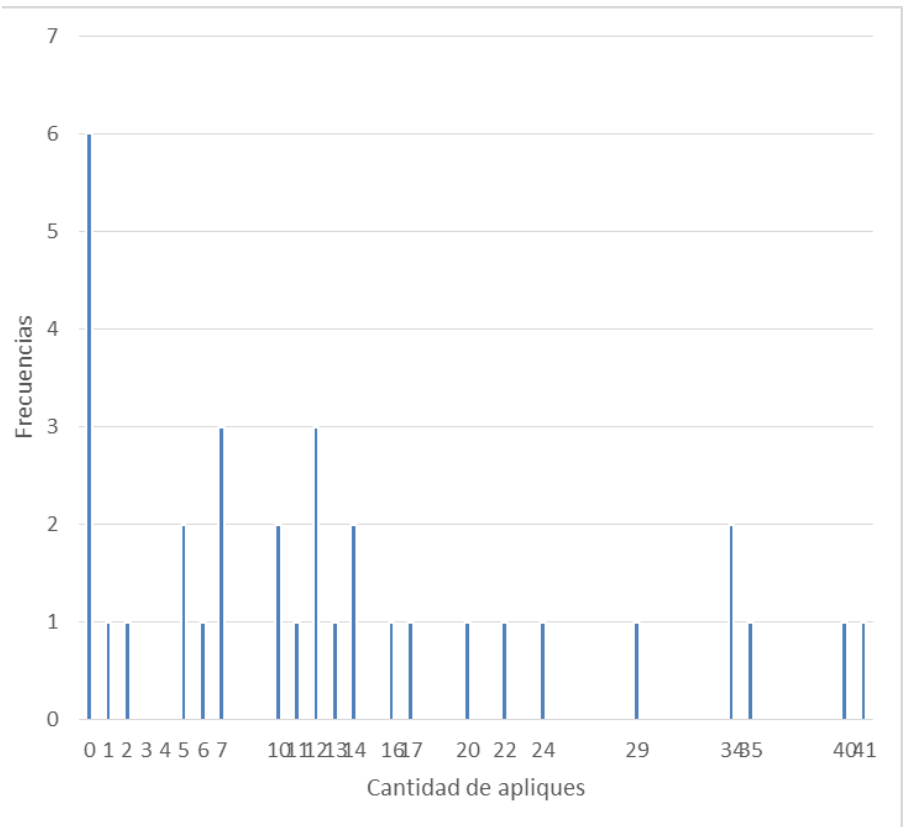
El análisis del gráfico nos muestra que en realidad la distribución de datos es bastante simétrica con respecto a la mediana de los datos, esta conclusión deriva de observar que el área de la caja a ambos lados de la mediana es similar y que longitud de los bigotes también es similar. Además, recordando que la media es de 48,5 faroles por mes vemos que es un valor cercano a la mediana- característica de las distribuciones simétricas-. Por otra parte se puede notar que nos encontramos con datos outlet o atípicos, únicamente en valores por encima de la mediana. La cantidad mínima de faroles vendidos en un mes fue de 10 faroles. Vemos que se tienen 3 outlet severos (114, 133 y 145 faroles), estos resultan claramente de meses atípicos, muy fuera de lo común en los que las ventas fueron demasiado altas. La presencia de estos valores atípicos elevados es la principal causa de diferencia entre la media y la mediana.

A continuación se presenta un análisis similar al anterior para la categoría de “apliques”. La media de los datos sin agrupar es $\bar{x}_A = \frac{1}{34} \sum_{i=1}^{34} x_i = 13,529$ apliques por mes. La mediana de los datos ordenados sin agrupar es $\widetilde{x}_A = \frac{(x_{17} + x_{18})}{2} = 11,5$ apliques. Sin agrupar, los datos son unimodales, la moda es de 0 faroles. La varianza de los datos es de $s^2 = \frac{1}{34} \sum_{i=1}^{34} (x_i - \bar{x}_A)^2 = 151,286$ y la desviación estándar es de $s = \sqrt{s^2} = 12,299$ apliques por mes. La mínima cantidad de apliques vendidos en un mes es de 0 apliques y la máxima de 41 apliques. El hecho de tener un rango tan extenso pero una desviación estándar tal que $(\bar{x}_A - s; \bar{x}_A + s)$ no cubre el rango total de la variable sino que solo cubre el mínimo y no el máximo nos indica que la distribución es asimétrica y está concentrada más cerca del mínimo que del máximo. Por otra parte, la semejanza entre la media y la mediana podría indicar que existe una región de gran concentración de los datos. Obviamente las dos afirmaciones anteriores resultan de un análisis sumamente superficial; para obtener mejores resultados procederé a realizar un análisis gráfico del bloque de datos.

A continuación se presenta la tabla de frecuencias del número de meses en los cuales se vendió una cantidad determinada de apliques, acompañada por el correspondiente gráfico de bastones.

Del análisis del gráfico de bastones es evidente que hubo una gran cantidad de meses en los que no se vendió ningún aplique, sin embargo también hubo muchos meses en los que sí se vendieron muchos apliques, por ejemplo hubo 9 meses en los que se vendieron más de 20 apliques. También es interesante notar que hubo 16 meses en los cuales se vendieron entre 7 y 22 apliques. Es decir que casi el 50 % de todos los datos están concentrados entre 2 y 22 apliques por mes.

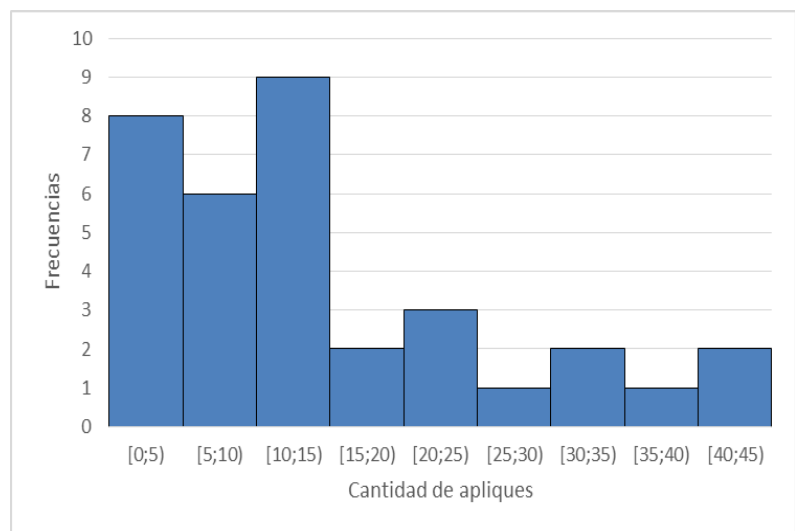
Apliques- Tabla de frecuencias		
Cantidad de apliques	Frecuencias absolutas	Frecuencias acumuladas
0	6	6
1	1	7
2	1	8
5	2	10
6	1	11
7	3	14
10	2	16
11	1	17
12	3	20
13	1	21
14	2	23
16	1	24
17	1	25
20	1	26
22	1	27
24	1	28
29	1	29
34	2	31
35	1	32
40	1	33
41	1	34
Totales	34	34



Con el fin de presentar la información de una manera más concisa, procedí a construir un histograma para el conjunto de datos. El análisis de estos gráficos es más simple que el correspondiente al gráfico de bastones. Podemos ver que la distribución tiene asimetría positiva, estando la producción concentrada entre 0 y 14 faroles por mes.

A continuación presento un segundo histograma que presenta una característica muy interesante.

Apliques-Tabla de frecuencias		
Cantidad de apliques	Frecuencia	Frecuencia acumulada
[0;5)	8	8
[5;10)	6	14
[10;15)	9	23
[15;20)	2	25
[20;25)	3	28
[25;30)	1	29
[30;35)	2	31
[35;40)	1	32
[40;45)	2	34
Totales	34	34



Como puede verse en el gráfico, y a partir de la tabla, la moda, con los datos así agrupados, está entre 10 y 15 faroles por mes. A partir de la fórmula de interpolación lineal:

$$Mo = L_i + \frac{f_{i+1}}{f_{i+1} + f_{i-1}} \times c$$

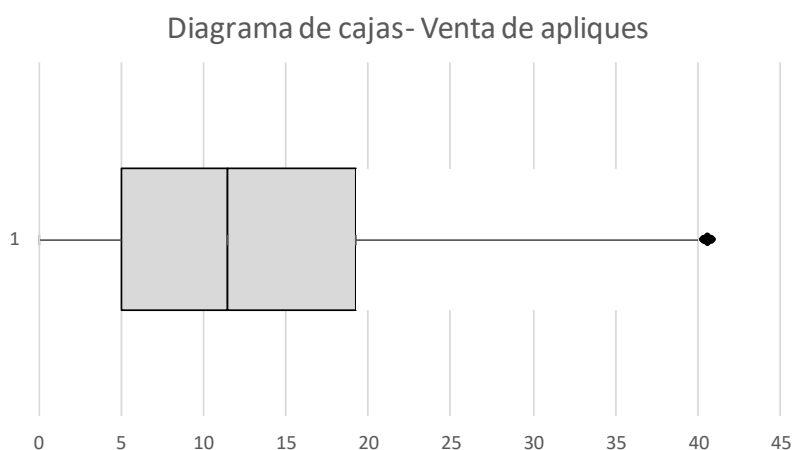
Donde Mo es el valor de la moda, L_i el valor del límite inferior del intervalo modal (el que tiene una mayor frecuencia), f_{i+1} la frecuencia absoluta del intervalo siguiente al intervalo modal, f_{i-1} la frecuencia absoluta del intervalo anterior al modal y c la longitud del intervalo modal. Se obtiene a partir de los datos, $Mo= 11,5$ faroles.

Es interesante destacar este aspecto porque nos muestra que diferentes intervalos de clase, pueden llevarnos a interpretar los datos de distinta manera y es por ello que un análisis correcto no puede basarse en una única herramienta estadística.

Como en el caso de los faroles, la realización de distintos tipos de gráficos para la distribución encuentra su justificación en la búsqueda de las características esenciales de la distribución. Como resulta evidente, esta tiene una marcada asimetría positiva.

La última herramienta gráfica a utilizar será un diagrama de caja de la cantidad de meses en los que se vendió una cantidad determinada de apliques.

En el eje horizontal se encuentra representada la cantidad de apliques vendidos en un mes determinado. Del boxplot podemos analizar las 5 medidas resúmenes de la distribución. Estas son la mediana, de 11,5 apliques por mes; los cuartiles (el cuartil inferior de 5 apliques y el cuartil superior de 19,25 apliques); el rango intercuartil, de 14,25 apliques; y el rango de la distribución (el mínimo observado es de 0 apliques y el máximo de 41 apliques siendo el rango de 41 apliques). Quiero destacar que la longitud máxima admitida de los “bigotes” es de 1,5 RI, es decir 1,5 del rango intercuartil.

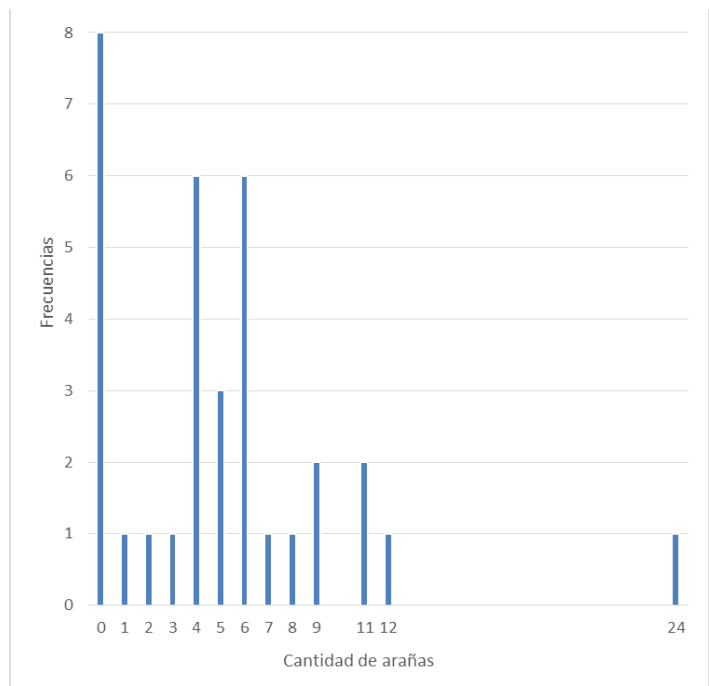


Podemos ver que el 50% de los datos se concentran entre 5 y 19 apliques por mes (en el análisis del gráfico de barras se había dicho que se encontraba entre 7 y 22 apliques por mes). Por otra parte vemos también que la distribución presenta una marcada asimetría positiva puesto que la longitud del bigote izquierdo es considerablemente menor a la del bigote derecho. Otro hecho a observar es que el 25% de los datos inferiores a la mediana se concentra en una región ligeramente más acotada que el 25% superior(a la mediana). Esto se observa a partir del ancho de las cajas a ambos lados de la mediana. Por último podemos destacar la presencia de un único outlet; sin embargo el valor de este es de 41 apliques en un mes y el último valor abarcado por el bigote es de 40 faroles por mes. Así, en realidad no es un valor “tan alejado”.

A continuación se presenta un análisis similar a los anteriores para la categoría de “arañas”. La media de los datos sin agrupar es $\bar{x}_a = \frac{1}{34} \sum_{i=1}^{34} x_i = 5,059$ arañas por mes. La mediana de los datos ordenados sin agrupar es $\widetilde{x}_a = \frac{(x_{17} + x_{18})}{2} = 4$ arañas. Sin agrupar, los datos son unimodales, la moda es de 0 arañas.

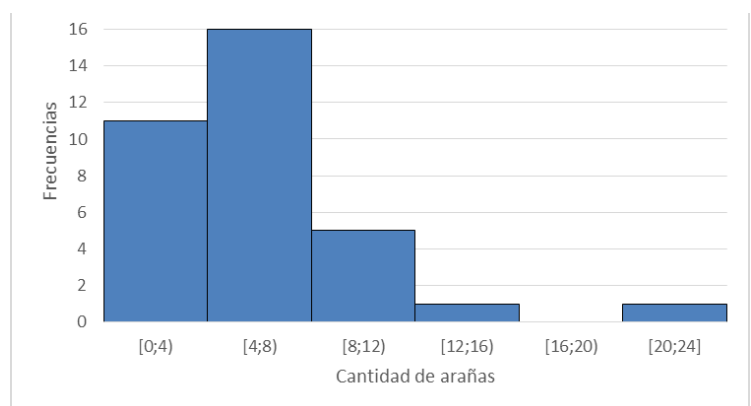
A continuación se presenta la tabla de frecuencias del número de meses en los cuales se vendió una cantidad determinada de arañas, acompañada por el correspondiente gráfico de bastones. Del análisis del gráfico de bastones es evidente que hubo una gran cantidad de meses en los que no se vendió ninguna araña (específicamente 8 meses). También podemos apreciar que hubo 4 meses en los cuales el número de arañas vendidas está entre 11 y 24, es decir una cantidad de ventas muy por encima del promedio.

Arañas- Tabla de frecuencias		
Cantidad de arañas	Frecuencias absolutas	Frecuencias acumuladas
0	8	8
1	1	9
2	1	10
3	1	11
4	6	17
5	3	20
6	6	26
7	1	27
8	1	28
9	2	30
11	2	32
12	1	33
24	1	34
Totales	34	34



Con el fin de presentar la información de una manera más concisa, procedí a construir un histograma para el conjunto de datos.

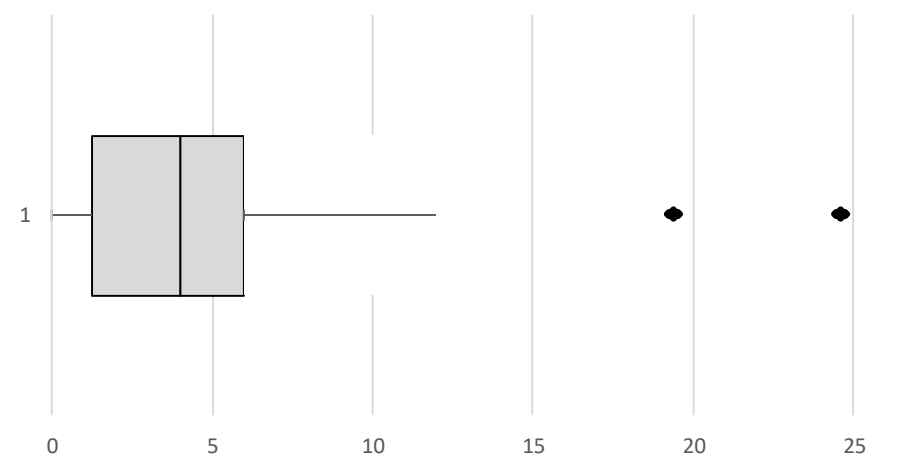
Arañas-Tabla de frecuencias		
Cantidad de apliques	Frecuencia	Frecuencia acumulada
[0;4)	11	11
[4;8)	16	27
[8;12)	5	32
[12;16)	1	33
[16;20)	0	33
[20;24]	1	34
Totales	34	34



Como puede apreciarse, se tiene una distribución con asimetría positiva. Es notable que la mayor parte de la distribución se concentra entre los 4 y los 12 arañas por mes.

La última herramienta gráfica a utilizar será un diagrama de caja de la cantidad de meses en los que se vendió una cantidad determinada de arañas.

Diagrama de caja- Venta de arañas



En el eje horizontal se encuentra representada la cantidad de arañas vendidas en un mes determinado. Del boxplot podemos analizar las 5 medidas resúmenes de la distribución. Estas son la mediana, de 4 arañas por mes; los cuartiles (el cuartil inferior de 1,25 arañas y el cuartil superior de 6 arañas); el rango intercuartil, de 4,75 arañas; y el rango de la distribución (el mínimo observado es de 0 arañas y el máximo de 24 arañas siendo el rango de 24 arañas). Quiero destacar que la longitud máxima admitida de los “bigotes” es de 1,5 RI, es decir 1,5 del rango intercuartil.

El análisis del gráfico nos muestra que en realidad la distribución de datos es bastante asimétrica, esta conclusión deriva de observar que el ancho de la caja a ambos lados de la mediana es muy diferente y así como también sucede con la longitud de los bigotes. Por otra parte vemos que nos encontramos con datos outlet o atípicos, únicamente en valores por encima de la mediana. Vemos que se tienen 2 outlet severos (19 y 24) estos resultan claramente de meses atípicos, muy fuera de lo común en los que las ventas fueron demasiado altas. La presencia de estos valores atípicos elevados es naturalmente la principal causa de diferencia entre la media y la mediana (la media es superior).

El último paso del análisis descriptivo a realizar es una breve comparación entre los lotes de datos. Fue con el objetivo de compararlos en última instancia que todas las tablas y gráficos se hicieron en función de la frecuencia de meses en los que se vendió una cantidad determinada de artículos de cada categoría. Gracias a esto es que las muestras tienen la misma cantidad de elementos (el número de meses -34-) y esto hace más fácil de comparar las ventas en las distintas categorías a pesar de que en ellas se haya producido una venta efectiva de una cantidad muy diferente de artículos.

Gráfico comparativo de diagramas de cajas

Diagrama de caja- Ventas de faroles-

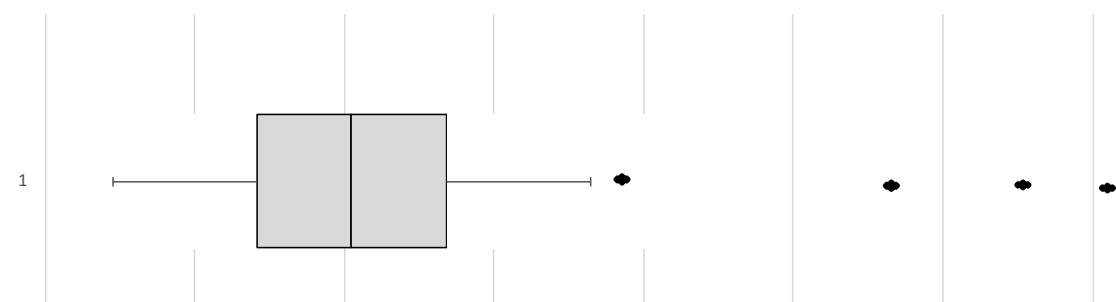


Diagrama de caja- Venta de apliques-

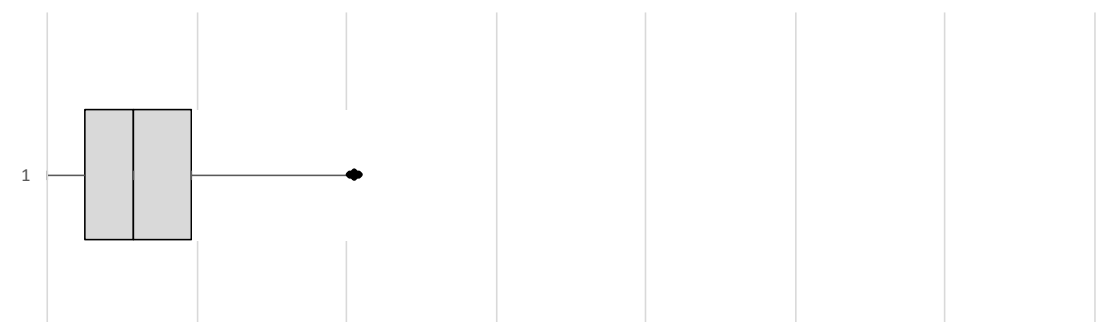
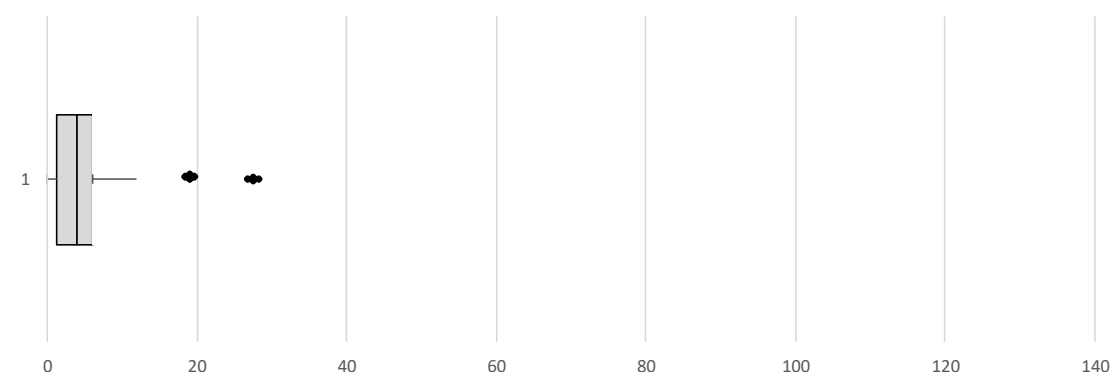


Diagrama de caja- Venta de arañas

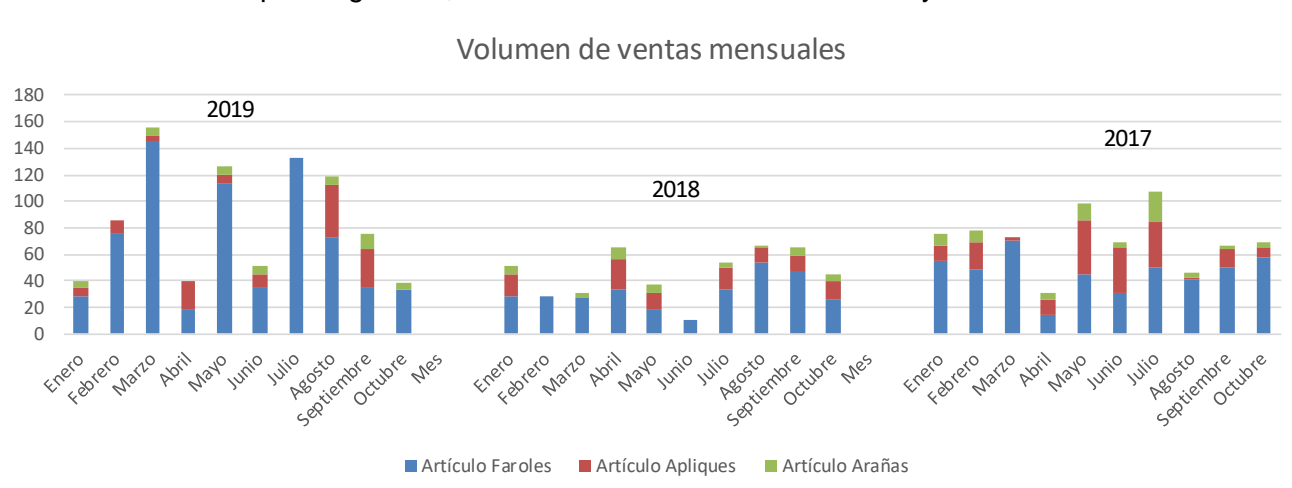


El primer detalle a considerar es la posición relativa de las diferentes cajas. Es evidente que la mediana de ventas de faroles es superior a la de las otras categorías y que la mediana de la venta de

arañas es la menor. En cuanto a la dispersión, salta a la vista que la distribución de los faroles es la más dispersa y la de las arañas la menos dispersa. Un hecho destacable es que el cuartil inferior de los faroles es superior al máximo de las arañas; esto quiere decir que en el 75% de los meses en los que se vendieron faroles, se vendieron más artículos que en el 100% de los meses en los que se vendieron arañas. Comparando los gráficos correspondientes a faroles y apliques podemos ver que el mínimo de faroles es superior al cuartil inferior de los apliques; por lo tanto en el 100% de los meses en los que se vendieron faroles se vendieron más artículos que en el 25 % de los meses en los que se vendieron apliques. De manera similar, el mínimo de faroles es superior al cuartil superior de las arañas, así en el 100% de los meses en los que se vendieron faroles se vendieron más artículos que en el 75 % de los meses en los que se vendieron arañas. Comparando el gráfico correspondiente a los apliques con el de las arañas podemos observar que el cuartil inferior de los apliques es casi igual al cuartil superior de las arañas. Lo anterior quiere decir que en el 75% de los meses en los que se vendieron apliques se vendieron más artículos que en el 75% de los meses en los que se vendieron arañas. Por otra parte, la presencia de outlets severos únicamente en la distribución de los faroles nos muestra que en el período desde 2017 se produjeron grandes ventas que implicaron la compra de grandes cantidades de faroles pero no de artículos de otras categorías. Estos meses particulares donde las ventas fueron exuberantes corresponden a ventas puntuales realizadas a grandes hoteles y son sucesos en general muy extraños.

El último análisis a realizar corresponde a una evaluación de las ventas en un nivel general en función del tiempo. El siguiente es un gráfico de barras del volumen de ventas mensuales efectuadas.

A simple vista pareciera que el volumen neto de ventas fue mayor en el año 2019 que en el año 2017 y en este último mayor que en el 2018. Efectivamente, en 2019 se vendieron 865 artículos en las categorías consideradas; en 2018 ,454; y en 2017 se vendieron 715 artículos. Por otra parte, comparado las barras para los diferentes meses, observamos que los meses de Marzo, Julio, Agosto y Septiembre son los meses en los que en general, el volumen de ventas totales es mayor.



Por último, es observable que además de lo dicho anteriormente, no pareciera haber otra tendencia común entre los 3 periodos considerados.

Segunda Parte: Estadística Bayesiana

En esta segunda parte del trabajo se desarrollará la estimación puntual y por intervalo de la media de una población utilizando una técnica basada en la estadística Bayesiana. Antes expondré una breve descripción del método Bayesiano.

Considérese el problema de calcular un estimado puntual del parámetro θ para la población con distribución $f(x|\theta)$, dado θ . Denótese con $\pi(\theta)$ la distribución a priori de θ . Supóngase que se observa una muestra aleatoria de tamaño n denotada con $\mathbf{X}=(X_1,X_2,\dots,X_n)$. A partir del teorema de Bayes es fácil demostrar que la distribución de θ dado \mathbf{X} es dada por:

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{g(\mathbf{x})},$$

Donde $g(\mathbf{X})$ es la distribución marginal de \mathbf{X} . En este caso se suele llamar a $f(x|\theta)$ función de verosimilitud y a $\pi(\theta)$ la distribución a priori.

La distribución marginal de \mathbf{X} en la definición anterior se puede calcular usando la siguiente fórmula:

$$g(\mathbf{x}) = \begin{cases} \sum_{\theta} f(\mathbf{x}|\theta)\pi(\theta), & \theta \text{ es discreta} \\ \int_{-\infty}^{\infty} f(\mathbf{x}|\theta)\pi(\theta) d\theta, & \theta \text{ es continua} \end{cases}$$

Si \bar{x} es la media de una muestra aleatoria de tamaño n tomada de una población normal con varianza conocida σ^2 , y la distribución a priori de la media poblacional es una distribución normal con media conocida μ_0 y varianza conocida σ_0^2 , se puede demostrar² que la distribución a posteriori de la media poblacional es también una distribución normal con media μ^* y desviación estándar σ^* , donde:

$$\mu^* = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n} \bar{x} + \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n} \mu_0 \quad y \quad \sigma^* = \sqrt{\frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}}.$$

En este caso, consideraremos que nuestra población está constituida por el número de ventas de faroles producidas en un mes determinado. El primer paso es determinar la distribución a priori del parámetro. A partir de una encuesta realizada al dueño del local comercial, quien se encarga también de planificar la producción, se obtuvo un valor medio del número de ventas de faroles mensuales de 40 faroles por mes. Además se obtuvo el dato de que la cantidad de faroles vendidos en un mes puede ser mayor o menor a 40 indistintamente, con lo cual supondremos que el parámetro tiene una distribución simétrica alrededor de su media y por simplicidad supondremos que dicha distribución es normal. Otro dato muy importante es que históricamente fue muy poco común que la cantidad de faroles vendidos en un mes determinado sea menor a 20, e intuitivamente se determinó que la probabilidad de ocurrencia de dicho fenómeno sería menor al 5 %³. Si μ es el parámetro de la población, estamos suponiendo que $\mu \sim N(\mu_0; \sigma_0^2)$ donde $\mu_0 = 40$ y σ_0 debe ser determinado. Se tiene que la variable Z definida por:

$$Z = \frac{\mu - \mu_0}{\sigma_0}$$

Tiene una distribución normal estándar. El valor de Z que deja un área de 0,05 a su izquierda es $z=-1,65$. Así se tiene que el valor de Z correspondiente a $\mu=20$ es $z=-1,65$. Por lo tanto:

$$-1,65 = \frac{20 - 40}{\sigma_0}$$

Por lo tanto se obtiene $\sigma_0 = 12,12$. Así la distribución a priori es una distribución normal con media $\mu_0 = 40$ y desviación estándar $\sigma_0 = 12,12$.

Ahora es necesario determinar la función de verosimilitud. En primer lugar, supóngase que se toma una muestra aleatoria de tamaño n de una población f con media μ y desviación estándar σ . Dicha muestra está constituida por las n variables aleatorias X_1, X_2, \dots, X_n . No se conoce dicha distribución f , sin embargo por el teorema del límite central se sabe que la variable aleatoria $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ tiene una distribución aproximadamente normal con media μ y desviación estándar $\frac{\sigma}{\sqrt{n}}$ para un n lo suficientemente grande. En general se admite la validez del enunciado anterior para un valor de n mayor o igual a 30. Por otra parte, si n es lo suficientemente grande y la población es aproximadamente acampanada⁴, se puede aproximar σ la desviación estándar poblacional con s la desviación estándar de la muestra. Así, escribimos $\bar{X} \sim N(\mu; s^2)$, lo cual constituye evidentemente una aproximación.

A partir de la muestra recolectada⁵ se tiene $n=34$, $\bar{x}=48,5$ y $s=30,92$. Por lo tanto, la distribución a posteriori para la media poblacional es normal con media μ^* y desviación estándar σ^* donde:

$$\mu^* = \frac{\sigma_0^2}{\sigma_0^2 + \frac{s^2}{n}} \times \bar{x} = \frac{12,12^2}{12,12^2 + \frac{30,92^2}{34}} \times 48,5 = 40,707$$

² Véase "Probabilidad y estadística para ingeniería y ciencias", Novena edición, RONALD E. WALPOLE, RAYMOND H. MYERS, SHARON L. MYERS y KEYING YE, Pearson Education.

³ Nótese que en la muestra recolectada, en 4 meses que la cantidad de faroles vendidos fue menor a 20. En una muestra de 34 meses, esto representa un 11% aproximadamente. Sin embargo como se mencionó, decir que la probabilidad de ocurrencia de este evento es de un 5 % aproximadamente resulta de hechos históricos.

⁴ Nótese que del análisis descriptivo del lote de datos de ventas de faroles se concluyó que la distribución de los datos recolectados era simétrica.

⁵ Los datos que utilizaré como datos muestrales son de hecho los mismos datos observacionales utilizados en la primer sección del desarrollo.

$$\sigma^* = \sqrt{\frac{\sigma_0^2 s^2}{n\sigma_0^2 + s^2}} = \sqrt{\frac{12,12^2 \times 30,92^2}{34 \times 12,12^2 + 30,92^2}} = 4,858$$

Resulta evidente que la estimación puntual para la media poblacional de la venta mensual de faroles es el valor esperado de μ , esto es μ^* . Así el valor esperado de la venta mensual de faroles es de 40,707~41 faroles por mes. Nótese la gran diferencia obtenida mediante el método clásico y el método bayesiano. A través de la teoría clásica la estimación puntual para la media poblacional fue de $\bar{x}=48,5$. Podemos ver entonces el gran peso que tuvo en la estimación la información a priori utilizada.

A continuación procederé a construir un intervalo Bayesiano, o intervalo de verosimilitud para μ . Un intervalo bayesiano del $100(1-\alpha)\%$ para μ , dado que este tiene una distribución normal con media μ^* y desviación estándar σ^* es:

$$\mu^* - z_{\alpha/2} \times \sigma^* < \mu < \mu^* + z_{\alpha/2} \times \sigma^*$$

Por ende, un intervalo bayesiano del 99% para μ , la media poblacional de ventas mensuales de faroles es:

$$\begin{aligned} \mu^* - z_{0,01/2} \times \sigma^* &< \mu < \mu^* + z_{0,01/2} \times \sigma^* \\ 40,707 - 2,33 \times 4,858 &< \mu < 40,707 + 2,33 \times 4,858 \\ 29,388 &< \mu < 52,026 \end{aligned}$$

La interpretación del intervalo bayesiano es muy simple, quiere decir que existe una probabilidad del 99% de que la media poblacional se encuentre en el intervalo. Así se tienen un 99% de seguridad de que el valor esperado de la cantidad mensual vendida de faroles estará entre 29 y 52 faroles aproximadamente.

A modo ilustrativo, procedí también a calcular el intervalo de confianza para la media poblacional a partir de métodos clásicos. Utilizando como estimador de la media poblacional a la media muestral se obtuvo, como ya se mencionó, una estimación puntual de 48,5 faroles por mes. Por otra parte, utilizando el teorema del límite central y usando la aproximación $s \sim \sigma$, el intervalo de confianza del 99% para la media poblacional obtenido fue:

$$43,197 < \mu < 53,802$$

La interpretación de este intervalo es sumamente diferente. No se tiene una probabilidad del 99% de que la media poblacional se encuentre en el intervalo obtenido sino que se tiene una probabilidad del 99% de haber seleccionado una muestra que produzca un intervalo que contenga al parámetro. Si se seleccionó una muestra que cumple las condiciones anteriores; entonces el parámetro se encuentra en el intervalo. La diferencia puede parecer sutil, pero es sumamente relevante. El intervalo bayesiano nos permite hacer inferencias sobre la probabilidad de encontrar al parámetro en un intervalo determinado puesto que para construirlo se ha trabajado con la distribución del parámetro mientras que el intervalo de confianza nos proporciona información sobre probabilidades relativas a la muestra. Una clara ventaja de los métodos bayesianos está implícita en lo anterior. Nos permite construir una distribución para el parámetro mismo.

El último punto en este trabajo es la estimación puntual de la proporción de artículos fuera de catálogo vendidos. Esto se hará a partir de la estadística bayesiana siguiendo modelo beta-binomial. En primer lugar realizaré una breve descripción del modelo.

Una variable aleatoria, tiene distribución beta de parámetros α, β en $[0, 1]$ (y se representa como $X \sim \text{Be}(\alpha, \beta)$), con $\alpha, \beta > 0$ si su función de densidad es:

$$f(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} I_{0 < x < 1}$$

Recuérdese que la función gamma se define como:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

Cuando $\alpha = \beta = 1$, se obtiene la distribución uniforme y, en general, cuando $\alpha = \beta$ la distribución es simétrica alrededor de $1/2$. Cuando $\alpha < \beta$ se concentra la probabilidad hacia la izquierda, y a la inversa cuando $\alpha > \beta$.

Supóngase ahora un experimento que consiste en observar n casos independientes, registrándose el número de casos favorables o éxitos que se presentan. Evidentemente, si X es la variable aleatoria que indica el número de éxitos en n ensayos Bernoulli independientes, con una probabilidad de éxito p , entonces $X \sim \text{Bi}(n, p)$. Así:

$$P(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

Se realiza, por tanto, el experimento y supongamos que se producen x éxitos. Nuestro interés se centra en estimar la proporción p de éxitos dada la muestra. Resulta inmediato que la verosimilitud es binomial. Por otra parte, si consideramos que el parámetro p tiene una distribución beta de parámetros α , β ; en nuestro caso se tiene que:

$$f(p|\alpha, \beta) = \begin{cases} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} & \text{para } 0 \leq p \leq 1 \\ 0 & \text{resto} \end{cases}$$

Además, se puede demostrar que:

$$E(p) = \frac{\alpha}{\alpha + \beta}, \quad Var(p) = \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)},$$

A partir del teorema bayesiano para determinar la distribución a posteriori es fácil demostrar que la distribución a posteriori de p sigue una distribución beta de parámetros $(x + \alpha)$ y $(n - x + \beta)$. Por lo tanto la media a posteriori (la estimación puntual de p) es:

$$p^* = \frac{x + \alpha}{n + \alpha + \beta}$$

Así, para realizar la estimación bayesiana puntual del parámetro p es necesario determinar a partir de una muestra los valores de x y n y los valores de α y β a partir de la distribución a priori.

Se quiere determinar la proporción de artículos fuera de catálogo vendidos. Así, se recolectaron los datos del número de artículos fuera de catálogo vendidos en 34 meses. Si X_1, X_2, \dots, X_{34} son variables aleatorias independientes que representan el número de artículos fuera de catálogo vendidos en un mes determinado entonces cada X_i representa la cantidad de artículos vendidos en el mes i -ésimo). Cada X_i tiene una distribución binomial $X_i \sim \text{Bi}(x_i, n_i, p)$. Es decir que si bien la probabilidad de éxito en cada una de las v.a. es la misma, el tamaño de la prueba (la cantidad de ensayos, es decir la cantidad de artículos vendidos) tomada en cada mes es diferente. La propiedad reproductiva de la distribución binomial nos permite asegurar que la variable aleatoria X , la cantidad de artículos defectuosos vendida en el período desde 2017 tiene una distribución binomial de la forma:

$$X \sim \text{Bi}(x_1 + x_2 + \dots + x_{34}, n_1 + n_2 + \dots + n_{34}, p)$$

Esta será ésta la función de verosimilitud utilizada. Escribimos $X \sim \text{Bi}(x, n, p)$, con $x = x_1 + x_2 + \dots + x_{34}$ y $n = n_1 + n_2 + \dots + n_{34}$.

A partir de la muestra seleccionada,⁶ se tiene $n = 2605$ artículos. Por otra parte, la cantidad de artículos defectuosos vendidos en el período fue de $x = 166$.

A partir de una encuesta realizada al dueño del local comercial, quien se encarga también de planificar la producción, se obtuvo que la proporción de ventas de artículos fuera de catálogo en un mes determinado fue de aproximadamente un 15%. Además se preguntó si alguna vez la proporción habría sido mayor al 50% y se dijo que nunca había ocurrido algo así. Por otra parte, en la encuesta se dijo que sería más probable que la proporción sea menor al 15 % antes que mayor. Toda la información recolectada sirve para modelar una distribución beta para el parámetro. En primer lugar, el valor esperado de la distribución sería evidentemente 0,15, por lo que se tiene que:

$$E[p] = 0,15 = \frac{\alpha}{\alpha + \beta}$$

Además, se sabe que sería más probable que $p < 0,15$ antes que la alternativa $p > 0,15$. Por lo tanto debería concentrarse la probabilidad hacia la izquierda. Así $\alpha < \beta$. El siguiente paso es la determinación de los parámetros. En realidad, como se tiene una única ecuación y dos incógnitas, su determinación se efectuará a partir de un proceso de tanteo usando la información conocida y luego se verificarán los resultados a partir del cálculo de probabilidades de que $p > 0,5$ que consideraré muy baja, tomaré dicha probabilidad de 0,001. Es decir que se buscan α y β tales que: $P(p < 0,15|\alpha, \beta) = 0,0001$,

Se tiene que:

$$0,15 = \frac{\alpha}{\alpha + \beta} \rightarrow \beta = \frac{\alpha - 0,15\alpha}{0,15}$$

⁶ Los datos que utilizaré como datos muestrales son de hecho los mismos datos observacionales utilizados en la primer sección del desarrollo.

Además, tenemos que $\alpha < \beta$. Así, dando a α una serie de valores obtenemos diferentes valores de β . Utilizando una distribución beta con dichos pares de parámetros y una calculadora de integrales, se obtuvo la probabilidad de que $p > 0,5$ dado el par específico. Los resultados se muestran en la tabla siguiente.

Pares de valores de α y β y la correspondiente probabilidad de ocurrencia								
α	β	Probabilidad de que $p > 0,5$	α	β	Probabilidad de que $p > 0,5$	α	β	Probabilidad de que $p > 0,5$
0,05	0,283333	0,144407104	0,5	2,833333	0,057184943	1	5,666667	0,019686266
0,1	0,566667	0,133539272	0,55	3,116667	0,051297511	1,5	8,5	0,007041899
0,15	0,85	0,121471712	0,6	3,4	0,046035559	2	11,33333	0,002583661
0,2	1,133333	0,109645799	0,65	3,683333	0,041332075	2,5	14,16667	0,000964281
0,25	1,416667	0,098593267	0,7	3,966667	0,037126449	3	17	0,000364304
0,3	1,7	0,088486935	0,75	4,25	0,033364334	3,5	19,83333	0,000138892
0,35	1,983333	0,079347958	0,8	4,533333	0,029997252	4	22,66667	5,33282E-05
0,4	2,266667	0,07113255	0,85	4,816667	0,026982085	4,5	25,5	2,05912E-05
0,45	2,55	0,063770835	0,9	5,1	0,024280545	5	28,33333	7,98731E-06

Resulta evidente que el par de parámetros que mejor se acerca a producir el resultado esperado ($P(p < 0,15 | \alpha, \beta) = 0,0001$ mientras que $E[p] = 0,15$) es el par:
 $\alpha = 2,5$ y $\beta = 14,167$.

Finalmente se tiene que:

$$p^* = \frac{x + \alpha}{n + \alpha + \beta} = \frac{166 + 2,5}{2605 + 2,5 + 14,167} = 0,0643$$

Es decir que la proporción de artículos fuera de catálogo vendidos es de aproximadamente un 6,4%. Nótese que debido al gran tamaño de la muestra, hay poca influencia de la información a priori con la que se cuenta en la estimación puntual. De hecho, el valor estimado de la proporción de defectuosos en el método clásico hubiese sido directamente $x/n=0,637$.

Conclusión

En el desarrollo de este trabajo se utilizaron diversos conceptos y herramientas estadísticas que permitieron efectuar un análisis global de un conjunto de datos. Los resultados puntuales del trabajo ya fueron expuestos en la sección “desarrollo del trabajo” con el objetivo de posibilitar que estos se comprendan de una manera más eficiente y efectiva permitiendo la visualización simultánea del desarrollo de un proceso y su resultado inmediato. Así en esta sección me abocaré a dar conclusiones generales sobre los métodos utilizados.

Un hecho que merece la pena destacar de la primera subsección del desarrollo es la forma en la cual se construyeron todas las tablas de frecuencias utilizadas. Para cada categoría se representó la frecuencia del número de meses en los cuales se vendió una cantidad determinada de artículos pertenecientes a la misma. Fue este hecho el que permitió la comparación directa y simple de las ventas en las distintas categorías⁷ a pesar de que la cantidad y tipo de artículos vendidos en cada una de ellas fuese diferente. Nótese que se podría haber construido las tablas de frecuencia utilizando diferentes conceptos, quizás más intuitivos o inmediatos⁸, pero la utilización de este sistema facilitó una comparación muy sencilla.

En el trabajo se expuso un gran defecto de la utilización de los histogramas. La construcción de intervalos de clase puede forzar a la aparición de características virtuales de la distribución, es decir características que no son reales. Es por ello que en el trabajo se usó un método que permitió utilizar histogramas sin resultar “engañados”. Simplemente se construyeron diversos histogramas con intervalos de clase de diferente longitud para así tener en cuenta únicamente aquellas características que prevalezcan de un gráfico a otro y de hecho son estas las características esenciales del lote.

También se pudo apreciar la gran versatilidad que ofrece la utilización de los gráficos de caja y como estos constituyen una alternativa eficaz a los métodos tradicionales de exposición de los datos. En este se pueden apreciar muchas de las características de la distribución que en los histogramas pueden quedar enmascaradas por la construcción de los intervalos de clase. Por ejemplo, los valores de la escala

⁷ Puesto que nos dio la capacidad de expresar todas las ventas en términos de la misma cantidad de elementos- los meses-.

⁸ Por ejemplo, podría haberse construido las tablas de frecuencia utilizando intervalos de clase continuos que expresen intervalos de tiempo y expresar la frecuencia de ventas de artículos en cada intervalo de tiempo.

que corresponden a la mediana, los cuartiles, los máximos y mínimos, etc. Además, los gráficos de caja permiten apreciar a los valores alejados de la distribución. Una de las grandes utilidades de estos tipos de gráficos es la facilidad que brindan a la hora de comparar distintos lotes de datos; posibilidad que no otorgan los histogramas.

Sin embargo, también resulta evidente que la mera utilización de gráficos de caja sin la contemplación de los histogramas resultaría incompleta ya que un gráfico de caja no nos permite identificar intervalos de mayor frecuencia tal como se mostró en el trabajo.

También pudo apreciarse como es que un análisis superficial a partir de datos meramente cuantitativos también puede proporcionar mucha información correcta de la distribución de los datos. Así, como se mostró en el análisis de cada categoría, el conocimiento de la media, el rango, la mediana y la desviación estándar de un lote de datos brinda mucha más información sobre el mismo que 4 meros datos puntuales. Estos pueden interpretarse para determinar posibilidad de valores alejados, zonas de concentración central, etc.

Por otra parte se utilizaron gráficos de barra acumuladas para mostrar el volumen de ventas. Esta herramienta, si bien es muy simple, es también muy potente puesto que permite un análisis comparativo inmediato de las variables en cuestión.

En cuanto, a la segunda subsección del desarrollo se utilizaron métodos bayesianos de estadística inferencial. La realidad es que la principal dificultad que se presenta en estos métodos es la de la determinación de la distribución a priori. Por ejemplo, en el caso de los faroles, se presentaron ciertas razones lógicas para poder suponer una distribución a priori normal para el parámetro en cuestión. Sin embargo, también es cierto que la distribución podría haber sido otra y se tiene un total desconocimiento acerca de la misma. La principal dificultad es que son pocos los datos reales con los que se cuenta para poder modelar dicha distribución.

A pesar de todo lo anterior, tenemos un resultado bastante interesante. Como pudo observarse en la determinación de la proporción de artículos defectuosos la suposición de que el parámetro seguía una distribución beta, si bien se contaba con ciertos datos que fomentaban su uso, fue simplemente una suposición. Sin embargo, cuando se utiliza una muestra lo suficientemente grande, es observable que poco influye en el resultado la forma general de la distribución a priori.

Por último, la utilización de los métodos bayesianos es totalmente justificada. Los métodos clásicos se ven contundentemente limitados por la información muestral principalmente porque se privan del conocimiento previo que el investigador pueda tener acerca de la materia en la que se desarrolla. Suponer que el investigador es ajeno a la materia es muchas veces incorrecto y así se muestra en este trabajo. Como ya se expuso, se contaban con supuestos formulados durante más de 30 años. Si bien nunca respondieron a un estudio formal, estos datos no son para nada descartables.

También es posible encontrar que las suposiciones subjetivas son incorrectas o, por lo menos, no tan acertadas. Así resultó ser el caso de la estimación de la proporción, donde se pudo apreciar que la información a priori: la proporción de artículos fuera de catálogo es de 0,15, no concordaba ni siquiera en cercanía con la información muestral: la proporción de artículos fuera de catálogo es de aproximadamente 0,6. Esto es evidentemente resultado de un conocimiento subjetivo incorrecto, una suposición generada por mera impresión y que no responde a la realidad. Una forma de evitar este fenómeno podría haber sido realizar una encuesta sobre la estimación de diferentes personas de los parámetros en cuestión. Así, todavía se estaría utilizando información subjetiva, pero no se haría depender el análisis de impresiones particulares erróneas. Así, en este trabajo, si bien se mostró como el uso de la estadística Bayesiana se ve sumamente justificado, presenta ciertas dificultades a salvar.

En este trabajo se muestra cómo es posible utilizar las herramientas estadísticas para poder diseñar un plan de producción. Además se pudo apreciar que el complemento de la estadística descriptiva y la inferencial permite efectuar un análisis integral de la situación del local comercial.

Bibliografía

- Ronald E. Walpole, Raymond H. Myers, Sharon I. Myers y Keying Ye. (2012). *Probabilidad y estadística para ingeniería y ciencias*, 9ª Edición. Ed. Pearson educación.
- Dailos Castellano Marrero (2015). *Introducción a la Estadística Bayesiana- Trabajo de fin de Grado de Estadística Aplicada Curso 2014/2015* (PDF). Recuperado de: https://ddd.uab.cat/pub/tfg/2015/137782/TFG_DailosCastellanoMarrero.pdf
- Notas de clases teóricas de la cátedra de "Probabilidad y Estadística I" de la Universidad Católica de Salta.
- Notas de clases teóricas de la cátedra de "Probabilidad y Estadística II" de la Universidad Católica de Salta.