Universidad De La Habana

MATCOM

Proyecto de Estadística

Fase 1

Autores:
Olivia González Peña C411
Juan Carlos Casteleiro Wong C411
Juan Carlos Vázquez García C412

Contents

Introducción

Nuestro trabajo se centra en el análisis e interpretación de datos a partir de los estadísticos estudiados en clases. Consta de un primer ejercicio en el que trabajamos con una población normal generada por nosotros y otros dos ejercicios que se nutren de un set de datos que detallamos más adelante e incluimos en el compactado. Para hallar las soluciones a las interrogantes planteadas en los tres ejercicios, le dedicamos a cada uno nuestra solución programada en el lenguaje de programación R. Para la solución de cada uno de los ejercicios, los tres integrantes del equipo debatimos la forma en la que se debía afrontar el problema, dejando a cada uno la escritura del código de uno de ellos (1: Juan Carlos Casteleiro, 2: Juan Carlos Vázquez, 3: Olivia). Asimismo, cada contratiempo o duda que se presentaba (sobre todo en el pre procesamiento de datos) la compartimos entre todos.

Ejercicio 1

2.1 Problema

Genere una Población Normal de tamaño 500, seleccione 8 muestras de tamaños varios (Mucho mayor que 30, mayor que 30, 30, 20), 4 muestras con remplazo y 4 sin remplazo.

a. Calcule para cada una de las muestras los Estadísticos Descriptivos, de la Conferencia 1.

b. Calcúlelos en la población inicial. Analice las diferencias.

c. Grafique los resultados

d. Para cada muestra calcule los intervalos de confianza para la media y la varianza

e. Analice las diferencias en los resultados de las muestras de tamaños similares

2.2 Solución

Para darle un poco de sentido a los datos generados, se simula que se refieren al consumo diario de azúcares añadidos de una persona adulta con alimentación regular para una población de 500 individuos con tales características. La información se basa en los resultados arrojados por el estudio Anibes (2013), que estima que el consumo medio es de 34 gramos. De esta población se extraen cuatro muestras sin reemplazo y cuatro con reemplazo, de tamaños 20, 30, 50 (mayor que 30) y 150 (mucho mayor que 30) en ambos casos. Cada una de estas se analiza a través de los estadísticos descriptivos estudiados en clases: Media, Moda, Mediana, Varianza, Desviación Estándar, Coeficiente de Variación y Cuartiles; e igualmente se procede con la población.

2.2.1 Resultados y Observaciones

MUESTRAS SIN REEMPLAZO						
n	MEDIA	MEDIANA	MODA	VARIANZA	D.E	C.V
20	34.53348	34.61435	34.71044	0.97572	0.98778	0.02860
30	33.97636	33.85711	33.73591	0.98057	0.99024	0.02914
50	33.82508	33.97537	34.08459	0.93388	0.96637	0.02856
150	34.15048	34.20484	34.31579	0.84458	0.91901	0.02691
500	33.97779	34.01595	34.27691	1.03406	1.01688	0.02992

MUESTRAS CON REEMPLAZO						
n	MEDIA	MEDIANA	MODA	VARIANZA	D.E	C.V
20	33.75620	33.15964	32.97848	1.32679	1.15186	0.03412
30	33.70979	33.47262	33.35642	1.20272	1.09668	0.03253
50	33.86883	33.65513	33.32666	1.15822	1.07620	0.03177
150	33.94290	33.97552	33.34395	0.93280	0.96581	0.02845
500	33.97779	34.01595	34.27691	1.03406	1.01688	0.02992

MUESTRAS SIN REEMPLAZO					
n	0%	25%	50%	75%	100%
20	32.49134	34.21119	34.61435	35.05234	36.80392
30	31.80522	33.54149	33.85711	34.69973	35.84861
50	31.82690	33.17716	33.97538	34.35024	36.06757
150	31.29760	33.55993	34.20485	34.76229	37.36077
500	31.12623	33.27891	34.01596	34.63213	37.37660

MUESTRAS CON REEMPLAZO					
n	0%	25%	50%	75%	100%
20	31.83794	32.94853	33.15965	34.80543	35.71643
30	31.77987	33.19152	33.47263	34.38747	37.15599
50	32.19990	33.02910	33.65514	34.52197	36.06757
150	31.78206	33.21863	33.97552	34.64517	36.80392
500	31.12623	33.27891	34.01596	34.63213	37.37660

Las muestras de mayor tamaño poseen características más similares a la población. Los valores como el de la media asociados a cada muestra difieren más con respecto a la media real de los datos mientras más pequeña es la muestra.

En todos los casos, los coeficientes de variación cumplen que: 0% <= CV <= 11%, por lo cual se puede afirmar que se trabaja con muestras y población con datos muy homogéneos.

Analizando las medidas de tendecia central de muestras del mismo tamaño con y sin remplazamiento los valores de las primeras suelen ser mas pequeños, exceptuando el valor de la media para la muestra de tamaño 50. Por otro lado las medidas de dispersión se comportan de manera opuesta en todos los casos.

De manera general, al comparar las medidas de posición de muestras de igual tamaño, los valores correspondientes a las muestras con remplazamiento suelen ser menores, con excepciones, como por ejemplo en el tercer cuartil de las muestras de tamaño 50.

Intervalo de Confianza para la Media ($\alpha = 0.05$)

- Muestra sin reemplazo:
  - para n = 20: [33.32799; 34.61204]
  - para n = 30: [33.72821; 34.41383]
  - para n = 50: [33.80829; 34.45376]
  - para n = 150: [33.88090; 34.24871]
  - para n = 500 (población): [33.96774; 34.15012]
- Muestra con reemplazo:
  - para n = 20: [33.47177; 34.79042]
  - para n = 30: [33.55481; 34.09010]
  - para n = 50: [33.75557; 34.22821]
  - para n = 150: [33.91929; 34.21805]
  - para n = 500 (población): [33.96774; 34.15012]

Intervalo de Confianza para la Varianza ($\alpha = 0.05$)

- Muestra sin reemplazo:
  - para n = 20: [0.5643046; 2.0814780]
  - para n = 30: [0.6219441; 1.7720808]
  - para n = 50: [0.6516466; 1.4501766]
  - para n = 150: [0.6813893; 1.0746872]
  - para n = 500 (población): [0.9168548; 1.1753953]
- Muestra con reemplazo:

- para n = 20: [0.7673488; 2.8304214]
- para n = 30: [0.7628422; 2.1735361]
- para n = 50: [0.8081876; 1.7985433]
- para n = 150: [0.7525594; 1.1869369]
- para n = 500 (población): [0.9168548; 1.1753953]

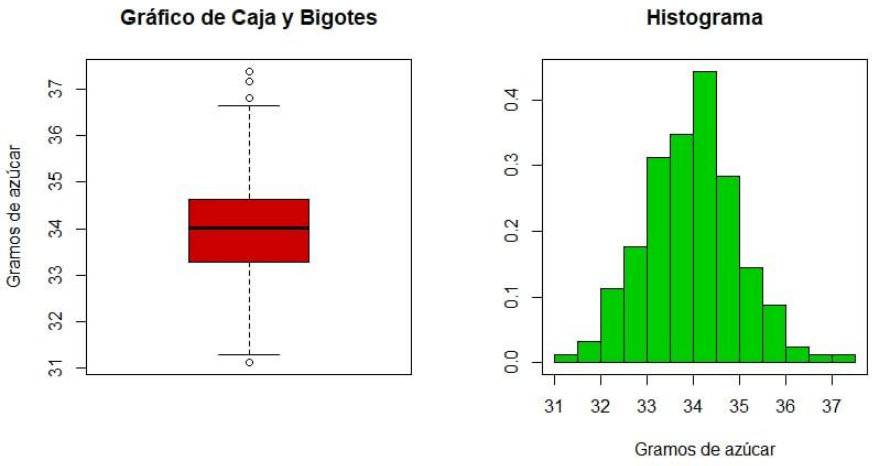
Asimismo, el intervalo de confianza para la media, es más restringido al intervalo de confianza para población, a medida que el tamaño de la muestra es mayor. En todos los casos los intervalos obtenidos son relativamente pequeños.

Es necesario aclarar que las especificidades expuestas se refieren a una ejecución en particular de nuestro código, pero de manera general los resultados obtenidos se comportan de manera similar.

Fue necesario instalar el paquete "modeest" para hacer uso de su función "mlv" para realizar el cálculo de la moda, ya que, aunque en las clases prácticas anteriores habíamos tenido que hacer nuestra propia implementación , esta función es sugerida en múltiples páginas visitadas por su eficiencia y versatilidad [1].

2.3 Gráficos

Los gráficos siguientes corresponden a los datos de la población generada.



Ejercicio 2

3.1 Problema

De acuerdo a su set de datos:

a. Utilice los Estadísticos Descriptivos estudiados en la Conferencia 1. Para describir el comportamiento de tres de sus variables. Seleccione las que sean más importantes y explique porque seleccionó estas.

b. Grafique los resultados.

c. Interprete los resultados en términos del problema.

3.2 Solución

Para la realización de este ejercicio y el siguiente, nos apoyamos en el archivo "IMBD.csv" (que deberá situarse en la misma carpeta que el código que brindamos como solución). El mismo hace referencia a una selección de las "mejor valoradas" 118 películas realizadas en inglés en el período comprendido entre los años 2010 y 2016, e incluye en su información datos sobre las mismas tales como nombre, año de realización, presupuesto destinado, total de votos obtenidos para la valoración, puntuación alcanzada, duración (en minutos) de la película, entre otros (55 en total).

Para realizar un análisis de dichos datos, escogimos las variables: Total de Votos, Duración y Rating; ya que en múltiples ocasiones cuando hemos tenido la necesidad de decidir cuál película ver, vamos en busca de estos principalmente.

Resultó necesario hacer un pre procesamiento para los datos asociados a la duración de las películas, ya que no solo aparecían los minutos numéricamente expresados, sino que se acompañaban de la especificación de la unidad en que fue medida: ¨min¨. Para ello, hicimos uso de la función "ConvertTime" que realizamos basándonos en la función "ConvertCurrency" [2].

Hallamos los estadísticos descriptivos mencionados en el primer ejercicio, esta vez asociados a los datos de cada una de las tres variables escogidas.

3.2.1 Resultados y Observaciones

variable	MEDIA	MEDIANA	MODA
rating	7.87288	7.8	7.78829
total votes	372835.02542	333251.5	312187.71447
runtime	126.67010	126	108

variable	VARIANZA	D.E	C.V
rating	0.05601	0.23666	0.03006
total votes	69312898254.6746	263273.42869	0.70613
runtime	383.07753	19.57236	0.15451

variable	0%	25%	50%	75%	100%
rating	7.5	7.7	7.8	8.0	8.8
total votes	26016	206593.8	333251.5	508417.5	1609713
runtime	91	113	126	138	180

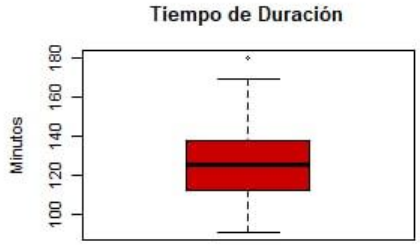
Los datos asociados al rating tienes valores próximos a 7.8 como muestran sus medidas de tendecia central. El coeficiente de variación refleja que dichos datos son muy hemogéneos.

Por otra parte, los datos asociados a la cantidad de votos son muy heterogéneos, promediando 372835 votos por películas.

Finalmente, la duración de las películas promedia sobre los 126 minutos aunque la duración más frecuente es de 108 minutos, en este caso los datos se clasifican como homogéneos teniendo en cuenta el coeficiente de variación (aunque casi alcanzando el valor para ser clasificados como heterogéneos).

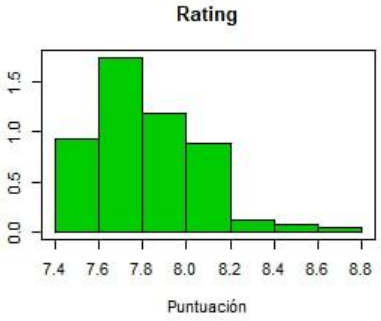
Para la realización de este ejercicio utilizamos nuevamente la función "mlv" del paquete "modeest".

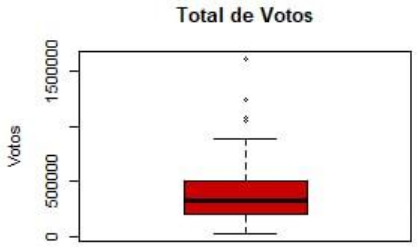
3.3 Gráficos

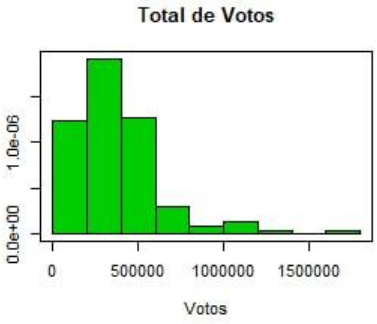

Tiempo de Duración
Minutos
100
120
140
160
180


Tiempo de Duración
0.000
0.010
100
120
140
160
180
Minutos


Rating
Puntuación
7.6
8.0
8.4
8.8


Rating
0.0
0.5
1.0
1.5
7.4
7.6
7.8
8.0
8.2
8.4
8.6
8.8
Puntuación


Total de Votos
Votos
0
500000
1500000


Total de Votos
0.0e+00
1.0e-06
0
500000
1000000
1500000
Votos

Ejercicio 3

4.1 Problema

¿Existen diferencias entre el precio del presupuesto de películas de acción y dramas?

4.2 Solución

Para la realización de este ejercicio, nos apoyamos en el set de datos mencionado en el ejercicio anterior. Esta vez, nos centramos en los géneros que caracterizan a las películas y en el presupuesto que se destina para su realización. Se entiende que cada película puede tener asociados hasta tres géneros (Aventuras, Comedia, Drama, Terror, etc.) y un presupuesto, todos estos pueden estar o no especificados.

Aclarar que hay casos en que una misma película tiene en unos de los géneros Drama y en otro Acción, el presupuesto destinado a dicha película está contemplado, por tanto, en el de cada uno de los géneros.

Fue necesario el pre procesamiento de los datos, pues en el caso del presupuesto, no solo aparecían las cifras numéricamente expresadas, sino que se acompañaban de la especificación de la moneda: ¨$¨; asimismo, las unidades de millar y de millón de los datos aparecían separadas por comas, caracteres no deseados para el entendimiento y manejo de las cifras por el lenguaje utilizado. Para dar solución a esto, hicimos uso de la función "ConvertCurrency" [2]. En algunos casos, también nos encontramos información asociada al presupuesto de una película que no trataba sobre este realmente (ver ejemplo en la imagen a continuación). Ante estos casos, decidimos aprovechar que la función ConvertCurrency devuelve los datos con tipo numérico ("as.numeric") y como estos no lo son se introducirían NA en estas casillas. Como en el pre procesamiento de la información tratamos estas casillas omitiéndolas del set, obtenemos el resultado esperado.

```
> my_data$Budget
  [1] $20,000,000
  [2] $18,000,000
  [3] $8,000,000
  [4] $12,000,000
  [5] $8,900,000
  [6] $44,500,000
  [7] $47,000,000
  [8] $6,000,000
  [9] $3,000,000
 [10] Opening Weekend:        56,215,889            (USA)      (7 November 2014)
 [11] $18,000,000
 [12] $13,000,000
 [13] $4,000,000
 [14] $40,000,000
 [15] $250,000,000
 [16] $170,000,000
 [17] Opening Weekend:        93,824            (USA)      (8 July 2016)
 [18] $55,000,000
 [19] $35,000,000
 [20] $5,000,000
 [21] $170,000,000
 [22] $58,000,000
 [23] Opening Weekend:        56,397,125            (USA)      (9 July 2010)
 [24] Opening Weekend:        10,739            (USA)      (16 March 2012)
 [25] $10,000,000
```

Se desea conocer si el presupuesto destinado a la realización de películas de Drama es igual al destinado para las películas de Acción, para lo cual debemos realizar una prueba de hipótesis de media de igualdad contra diferencia. Pero, al desconocer los valores de las varianzas y por tanto la igualdad o no entre ellas, primeramente, planteamos una prueba de hipótesis de igualdad contra diferencia de varianzas, utilizando la función var.test() de R. Luego de obtener como resultado que las varianzas son distintas, se procede a realizar la primera prueba antes mencionada que es, de hecho, la respuesta final de nuestro ejercicio. Para ello hicimos uso de la función t.test() de R.

Referencias

[1] - https://r-coder.com/moda-r/
[2] - https://stackoverflow.com/questions/7337824/read-csv-file-in-r-with-currency-column-as-numeric