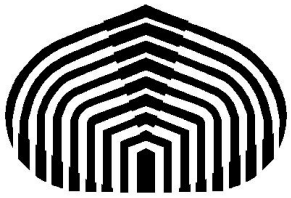


Universidad Simón Bolívar

Departamento de Cómputo Científico y Estadística

CI3321 – Estadística para Ingeniería

Trimestre Enero-Marzo 2012

INFORME I

Integrantes: Fátima Querales 07-41388

Johanna Chan 08-10218

Vicente Santacoloma 08-11044

Sartenejas, 24 de febrero de 2012

INTRODUCCIÓN

En el informe que se presenta a continuación se realizará un estudio y análisis de datos correspondientes al porcentaje del consumo total de proteínas de una muestra de 30 países europeos. Este consumo está distribuido entre nueve variables, que representan las principales fuentes de proteínas como los son las carnes rojas, carnes blancas, huevos, leche, pescado, cereales, almidón, nueces, frutas y vegetales.

Para la realización de este proyecto se utilizó como herramienta imprescindible el lenguaje de programación funcional R, con el cual, gracias a su gran repertorio de funciones y comandos se pudo realizar todos los cálculos necesarios para el análisis y graficación. Además, este lenguaje fue de suma utilidad para la verificación de algunos resultados y propiedades; y para la implementación de un programa capaz de calcular el intervalo de confianza para la diferencia de proporciones.

RESUMEN

A continuación se realizará en primer lugar un análisis descriptivo para cada una de las variables de la muestra. Para tal fin se utilizarán histogramas, diagramas de boxplot, así como también medidas numéricas descriptivas de localización, dispersión y posición.

Posteriormente, se proporcionarán dos tipos de agrupamientos para los países de la muestra. El primero se hará en base al consumo de proteínas característico de cada país, y el segundo tendrá como criterio la región geográfica a la cual pertenece cada uno.

Luego, utilizando ésta última agrupación, se calculará un intervalo de confianza para determinar si dos proporciones dadas son estadísticamente diferentes; y además se proporcionará un ordenamiento para establecer cuales zonas tienen mayor consumo para dos fuentes de proteínas (pescado y nueces).

Finalmente, se tendrá lugar a una serie de razonamientos para dar un criterio sobre cual de las dos agrupaciones ya mencionadas es mejor.

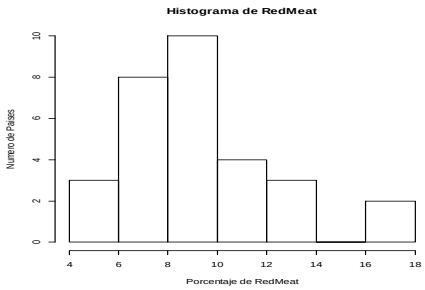
DESARROLLO

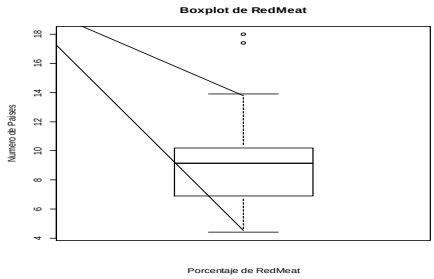
Pregunta 1: Realice un análisis descriptivo de cada variable contenida en el archivo de datos.

REDMEAT

RedMeat = Variable cuantitativa que representa el porcentaje total de proteína proveniente de carnes rojas para cada uno de los países.

Mínimo	1er. Quantile	Mediana	Promedio	3er Quantile	Máximo	Desviación
4.400	6.950	9.150	9.333	10.170	18.000	3.276598





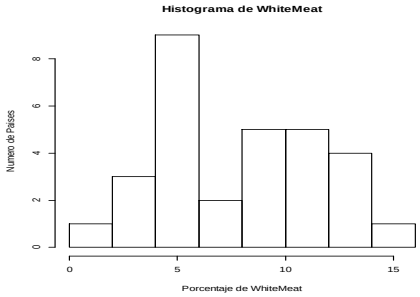
En el histograma se puede apreciar que el consumo de proteínas provenientes de carnes rojas para la mayoría de los países es menor al 10%. Este histograma posee un sesgo hacia la izquierda.

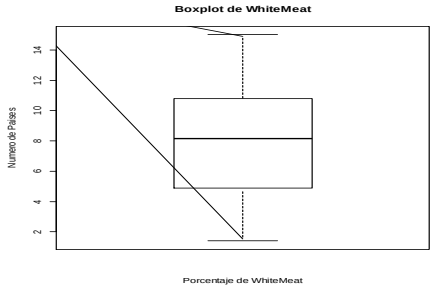
El boxplot corrobora los resultados obtenidos en el histograma, donde la mayoría de los países tienen un consumo de carnes rojas por debajo del 10%. Se puede observar además que los datos típicos no se encuentran distribuidos simétricamente alrededor de la mediana; y que se tienen dos países con un consumo de carnes rojas atípico entre un 16% y 18%.

WHITEMEAT

WhiteMeat = Variable cuantitativa que representa el porcentaje total de proteína proveniente de carnes blancas para cada uno de los países.

Mínimo	1er. Quantile	Mediana	Promedio	3er Quantile	Máximo	Desviación
1.400	4.925	8.150	8.053	10.650	15.000	3.693399





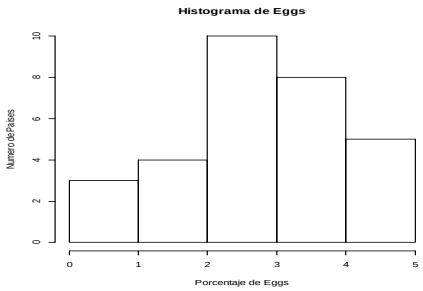
En el histograma se puede apreciar que el consumo de proteínas provenientes de carnes blancas en aproximadamente 9 países se establece en 5%. Además casi 10 países tienen un consumo entre 9% y 11%. Este histograma posee un sesgo hacia la derecha.

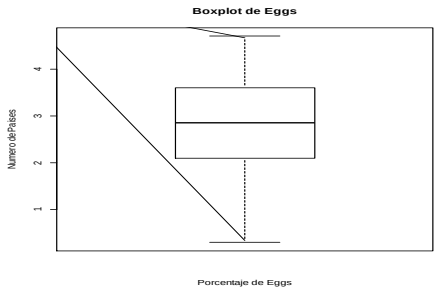
El boxplot corrobora los resultados obtenidos en el histograma, donde la mayoría de los países tienen un consumo de carnes rojas que va del 5% al 11%. Se puede observar además que los datos típicos no se encuentran distribuidos simétricamente alrededor de la mediana; y que no se tienen datos atípicos.

EGGS

Eggs = Variable cuantitativa que representa el porcentaje total de proteína proveniente de huevos para cada uno de los países.

Mínimo	1er. Quantile	Mediana	Promedio	3er Quantile	Máximo	Desviación
0.300	2.125	2.850	2.740	3.575	4.700	1.194124





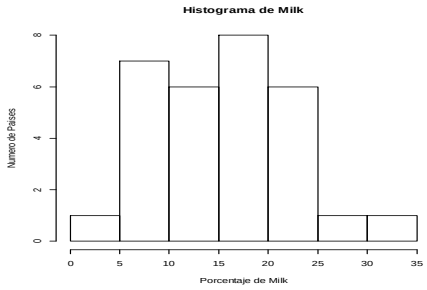
En el histograma se puede apreciar que el consumo de proteínas provenientes de huevos para la mayoría de los países es mayor al 2%. Este histograma posee un sesgo hacia la derecha.

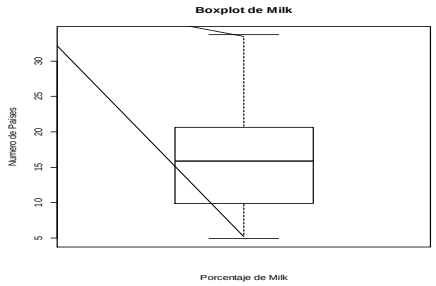
El boxplot corrobora los resultados obtenidos en el histograma, donde la mayoría de los países tienen un consumo de huevos por encima del 2%. Se puede observar además que los datos típicos se distribuyen de forma casi simétrica alrededor de la mediana. No hay observaciones que se puedan considerar como atípicas.

MILK

Milk = Variable cuantitativa que representa el porcentaje total de proteína proveniente de la leche para cada uno de los países.

Mínimo	1er. Quantile	Mediana	Promedio	3er Quantile	Máximo	Desviación
4.90	10.20	15.85	16.14	20.42	33.70	6.983063





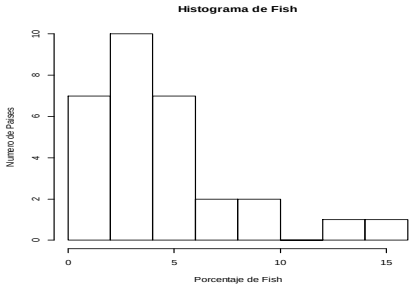
En el histograma se puede apreciar que la mayoría de los países tienen un consumo de proteínas provenientes entre de la leche de entre un 10% y 25%. Este histograma posee un leve sesgo hacia la izquierda.

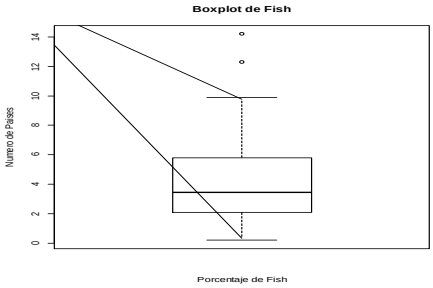
El boxplot corrobora los resultados obtenidos en el histograma, donde la mayoría de los países tienen un consumo de huevos por encima del 10% y por debajo del 25%. No hay distribución simétrica de datos típicos alrededor de la mediana y no se observan datos atípicos.

FISH

Fish = Variable cuantitativa que representa el porcentaje total de proteína proveniente del pescado para cada uno de los países.

Mínimo	1er. Quantile	Mediana	Promedio	3er Quantile	Máximo	Desviación
0.200	2.125	3.450	4.423	5.775	14.200	3.469938





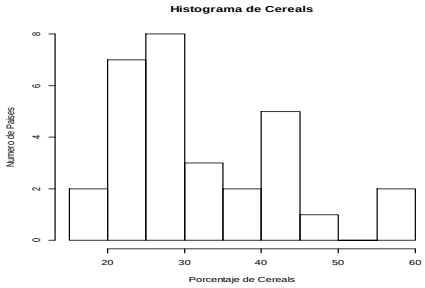
En el histograma se puede observar que el consumo de proteínas provenientes del pescado es menor al 6% para la mayoría de los países. Además se tiene una muy pequeña cantidad de países con un consumo entre 11% y 16%. El histograma posee un leve sesgo orientado hacia la izquierda.

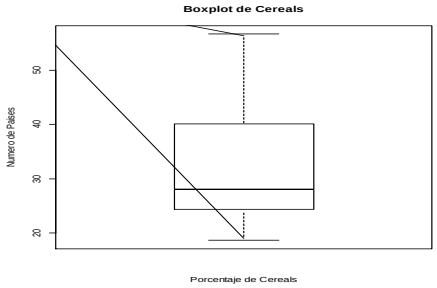
En este boxplot se observa que los datos típicos no se distribuyen de manera simétrica alrededor de la mediana. Se observa además que existen dos datos atípicos mayores que el resto de los datos con valores de 12% y 14%. Relacionando este boxplot con el histograma se aprecia que claramente la mayoría de los países tienen un consumo de pescado por debajo del 6%.

CEREALS

Cereals = Variable cuantitativa que representa el porcentaje total de proteína proveniente de cereales para cada uno de los países.

Mínimo	1er. Quantile	Mediana	Promedio	3er Quantile	Máximo	Desviación
18.60	24.38	28.05	32.01	39.28	56.70	10.33737





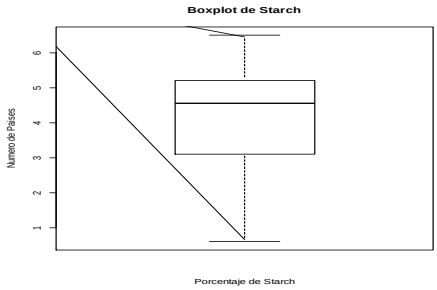
En el histograma se puede apreciar que el consumo de proteínas provenientes de cereales es alto para casi todos los países. Además el histograma posee un sesgo orientado hacia la izquierda.

El boxplot revela que los datos típicos no están distribuidos simétricamente alrededor de la mediana. No se tienen datos atípicos.

STARCH

Starch = Variable cuantitativa que representa el porcentaje total de proteína proveniente de almidón para cada uno de los países.

Mínimo	1er. Quantile	Mediana	Promedio	3er Quantile	Máximo	Desviación
0.60	3.20	4.55	4.22	5.20	6.50	1.531373



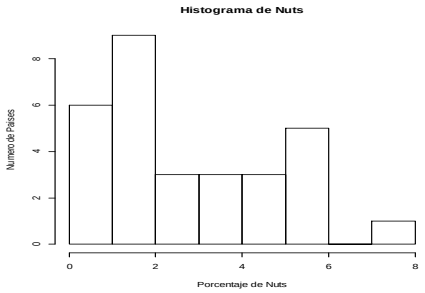
En el histograma se puede apreciar que el consumo de proteínas provenientes de almidón para la mayoría de los países es mayor al 3%. El histograma posee un sesgo a la derecha.

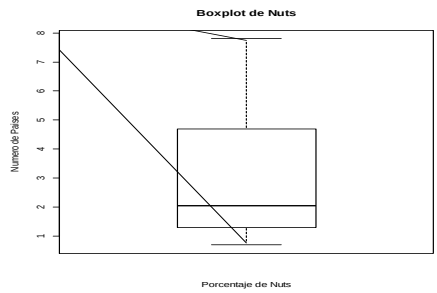
El boxplot nos señala que los datos típicos no se encuentran distribuidos de forma simétrica. Para el consumo de almidón no se tienen datos atípicos.

NUTS

Nuts = Variable cuantitativa que representa el porcentaje total de proteína proveniente de nueces para cada uno de los países.

Mínimo	1er. Quantile	Mediana	Promedio	3er Quantile	Máximo	Desviación
0.700	1.325	2.050	2.867	4.600	7.800	1.977517





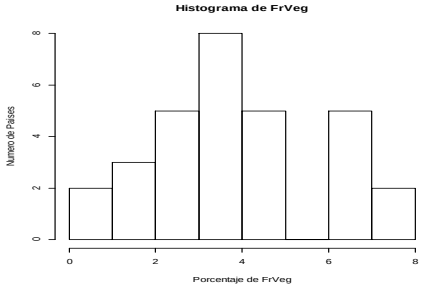
El histograma nos muestra como dato interesante que no se tienen países con un consumo entre 6% y 7%. Además éste posee un sesgo hacia la izquierda.

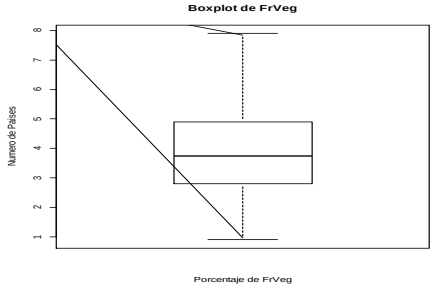
De acuerdo a lo visto en el diagrama de boxplot los datos se encuentran distribuidos asimétricamente con respecto a la mediana. No se observa ningún dato atípico. El grueso de los datos típicos se encuentra ubicado entre aproximadamente el 1% y el 5%.

FRVEG

Nuts = Variable cuantitativa que representa el porcentaje total de proteína proveniente de frutas y vegetales para cada uno de los países.

Mínimo	1er. Quantile	Mediana	Promedio	3er Quantile	Máximo	Desviación
0.900	2.825	3.750	3.970	4.750	7.900	1.873987

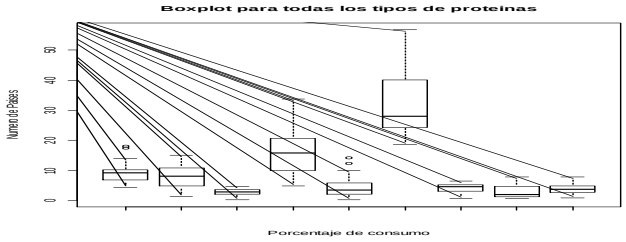




El histograma nos muestra como un dato interesante que no se tienen países que tengan un consumo de frutas y vegetales entre un 5% y 6%. Además éste presenta un leve sesgo a la izquierda.

Para la variable actual de estudio el diagrama de boxplot nos revela que los datos típicos están distribuidos casi de forma simétrica. No se observa ningún dato atípico.

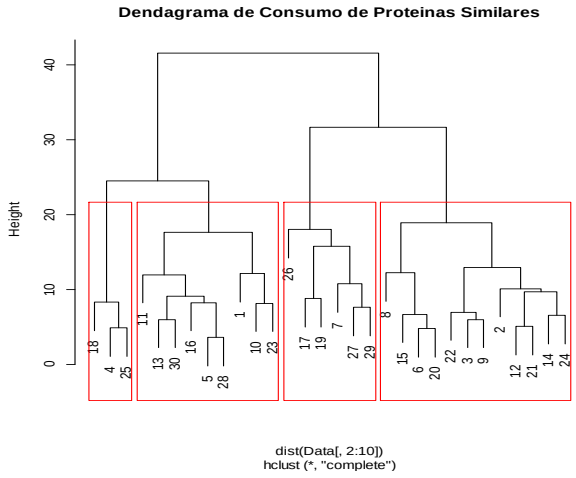
BOXPLOT DE TODAS LAS PROTEÍNAS



En este boxplot que incluye a todas las variables se aprecia el claro predominio de los cereales y la leche como fuente principales de proteínas para los países de la muestra.

<u>Pregunta 2</u>: Agrupe a los países de acuerdo a su consumo de proteínas característico. Justifique (desde un punto de vista estadístico) el número de grupos formados y sus integrantes.

Para obtener una agrupación como la aquí requerida, R cuenta con la función hclust, la cual permite agrupar las variables de acuerdo a un criterio. En este caso se agruparon los países por el consumo de proteínas que sean similares, obteniéndose el siguiente dendagrama:



En esta estructura todos los países que estén en nodos cercanos serán los más parecidos en cuanto a su consumo. Como estas diferencias pueden en algunos casos no ser tan distantes, se generaron 4 nuevos grupos a partir de los anteriores denotados por los rectángulos en rojo; esto con el fin de tener una mejor apreciación del consumo de proteínas de los países.

De este dendagrama vale la pena destacar que se tienen a los países Yugoslavia (25), Bulgaria (4) y Rumania (18) en un mismo grupo. Además en otro destacan, como parte de ese grupo, a los países Finlandia (8), Noruega (15), Dinamarca (6) y Suecia (20).

No es de extrañar que los países nombrados anteriormente, para esos dos grupos, tengan un consumo similar de proteínas puesto que ellos limitan geográficamente y además comparten factores ambientales, culturales y atmosféricos. El primer grupo pertenece a Europa del este y el segundo contiene a los todos los países nórdicos.

En algunos casos el factor cultural es quizás más determinante que el geográfico, ya que por ejemplo los países España (19) y Francia (9) que son fronterizos están en grupos diferentes.

Pregunta 3: Agregue al archivo de datos una nueva variable llamada "zona" utilizando para ello la división que se presenta en la figura 1. Cada país debe tener una zona asignada. Utilice la información para responder los siguientes ítems:

a. Determine el porcentaje del consumo total de proteínas para cada zona propuesta en el mapa.

Zona	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	FrVeg
1	10.93	10.1	3.5100	19.55	6.17	22.63	4.9300	1.6700	2.940
2	8.766667	8.5	2.9666	19.35	3.383	34.75	5.2000	1.6000	4.216
3	7.1125	5.925	1.6875	10.875	2.05	43.1	2.5375	4.9125	3.375
4	10.2	7.033333	2.6333	14.26667	5.716	30.13	4.3000	3.4000	6.233

En este cuadro está representado el consumo promedio de cada proteína para todos los países que conforman cada una de las zonas. Por ejemplo el consumo promedio de carnes rojas para la Zona 1 es de 10.93%. Un dato resaltante es que para todas las zonas prevale el consumo de proteínas provenientes de los cereales por sobre todas las demás.

b. Presente un código de R que ejecute un intervalo de confianza que permita determinar si dos proporciones son estadísticamente diferentes. ¿Se puede aplicar el código para responder si la proporción de países de la zona IV que consumen más del 9% de carnes rojas es significativamente diferente a la proporción de países de la zona III que consumen más del 9% de carnes rojas?. Justifique

El código del intervalo de confianza, así como el del filtro de los países de la zona IV y III de acuerdo a la condición indicada y de todo el procedimiento está ubicado en el anexo.

Para poder aplicar el intervalo de confianza para la diferencia de proporciones es necesario que se cumplan las siguientes condiciones:

n1, n2 >= 30
n1p1, n1q1 >= 5
n2p2, n2q2 >= 5

En este caso ninguna se cumplió puesto que el tamaño de las dos muestras n1=6 y n2=8, que representan la cantidad de países que conforman la zona iv y iii respectivamente, eran muy pequeñas.

c. Proporcione un ordenamiento (ranking de mayor a menor) de las zonas para determinar aquellas que consumen mayor cantidad de buenas proteínas: pescados y nueces.

Pescado

Zona	Consumo
1	6.17
2	5.716667
3	3.383333
4	2.05

La zona 1 es la que tiene el mayor consumo promedio de proteínas provenientes del pescado. Por su parte la zona 4 es la que tiene el menor consumo.

Nueces

Zona	Consumo
3	4.9125
4	3.4
1	1.67
2	1.6

La zona 3 es la que tiene el mayor consumo promedio de proteínas provenientes de nueces. Por su parte la zona 2 es la que tiene el menor consumo.

d. Indique cuál de los dos criterios de formación de grupos (el utilizado en la pregunta 2 versus el de la zona geográfica, pregunta 3) es mejor. Justifique.

Las agrupaciones generadas en la pregunta 2 mediante el comando hclust, proporcionan una información más exacta y precisa, ya que dependiendo del argumento con que se ejecute esta función, podremos ir teniendo diferentes enfoques de los datos. Por ejemplo con los argumentos: Ward tendremos grupos compactos y esféricos mientras que con enlace completo formaremos grupos de datos similares.

Por otra parte, agrupando los datos por zona geográfica, podremos en algunos casos tener información de interés, dependiendo de como se realice esta agrupación. Vale destacar que esta agrupación no depende de los datos, sino de los países que se desee agrupar por su cercanía geográfica. Es por esto que no se asegura que los países pertenecientes a un mismo grupo vayan a tener alguna relación. Por ejemplo, si agrega a una zona conformada por países que tengan acceso al mar, uno que no lo tenga, pese a estar geográficamente cercano a los otros, se puede estar obteniendo una información que desde el punto de vista del análisis puede resultar errónea o nada representativa para el consumo de proteínas provenientes del pescado, pues es posible que este país que no tiene acceso al mar no tenga un consumo alto.

CONCLUSIÓN

Luego de la realización de este proyecto acerca del consumo de proteínas para una muestra de 30 países europeos se logró tener los siguientes resultados:

- La leche y los cereales son los alimentos que más proporcionan proteínas a los países de la muestra.
- Las nueces, el almidón, los huevos, las frutas y los vegetales son los que menos proporcionan proteínas a los países de la muestra.
- No se logró determinar un intervalo de confianza para la diferencia de proporciones puesto que no se cumplían las condiciones necesarias.
- La zona 1 es la que tienen un mayor consumo de proteínas provenientes del pescado.
- La zona 4 es la que tiene un menor consumo de proteínas provenientes del pescado.
- La zona 3 es la que tienen un mayor consumo de proteínas provenientes de nueces.
- La zona 2 es la que tiene un menor consumo de proteínas provenientes de nueces.
- La agrupación mediante el comando hclust que genera un dendograma, es mas eficaz y precisa que la de por zonas geográficas.
- Los histogramas y diagramas de boxplot resultaron de gran utilidad para el análisis de los diferentes tipos de proteínas, puesto que permite identificar rápidamente aspectos relevantes acerca de su distribución.

BIBLIOGRAFÍA

Dennis D. Wackerly, William Mendenhall, Richard L. Scheaffer. ***"Estadística Matemática con Aplicaciones"***. Editorial: CENGAGE Learning. Edicion: 7ma.