

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

Facultad de Matemáticas, Departamento de Estadística

**DIPLOMADO EN ESTADISTICA – VERSION 2023**

**2 (M-J)**

Control Multivariado Nombre: Martín Venegas Márquez

14 de noviembre de 2023

La base de datos **Proteinas.xlsx** contiene información respecto al consumo de proteínas según tipo de alimento en las dietas típicas de países de Europa (1991).

1. Describir las variables. Comente (comportamiento, anomalías, etc.)
2. ACP: Obtenga DOS Componentes Principales y grafique
   1. ¿Qué representa el primer componente? ¿retención de varianza con 2 CP?
   2. ¿Es posible construir una “agregación” que permita identificar regiones?
3. AFAC: Lleve a cabo un Análisis Factorial.
   1. ¿Qué representa el primer factor? ¿cuántos factores proponen?
   2. Con la solución de dos factores, realice una rotación varimax y grafique.
4. Clúster: Obtenga un número apropiados de “Cluster”, ¿cómo se agrupan los países?
5. En no más de dos párrafos, contraste y discuta los resultados obtenidos por cada uno de los métodos

Hint: No es necesario complicarse con el concepto de proteína, solo basta con mencionar la variable como proteína “apellido”, por ejemplo, proteína Rojas. Simplifique.

\*\*\* Datos de Proteinas –

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., and Ostrowski, E. (1994), A Handbook of Small Data Sets, London: Chapman & Hall.

Artículo original que fue replicado en 1991:

Weber, A. (1973). Agrarpolitik im Spannungsfeld der internationalen Ernaehrungspolitik. Institut fuer Agrarpolitik und marktlehre, Kiel.

Indicaciones:

* Redacte un informe (pregunta/respuesta) que permita responder cada punto indicado.
* NO coloque el script en el informe (solo tablas, gráficos, métricas de interés, valores resultantes que ayuden a respaldar sus respuestas, no relleno).
* Por ningún motivo el último taller es pauta para este control, recaude toda la información pertinente (teórica como práctica).
* El trabajo es individual, pero puede ser grupal también (hasta 3 personas, sin más excepciones).
* El informe completo se entrega en el buzón “Entrega Final” (disponible en la plataforma del curso). No olvide poner el nombre de los miembros del grupo en el encabezado del informe!.
* **Fecha de entrega: 14 de noviembre hasta las 18:30 hrs.**
* Se debe entregar si o si un avance el 09 de noviembre hasta las 21:30 hrs.

**Desarrollo prueba**

**Primera pregunta: Describir las variables. Comente (comportamiento, anomalías, etc.)**

Tabla 1. Descriptivos del consumo de proteínas por tipo de alimento

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Media** | **DE** | **p0** | **p25** | **p50** | **p75** | **p100** | **Histograma** |
| Rojas | 0.170 | 0.059 | 0.074 | 0.135 | 0.165 | 0.184 | 0.315 | ▅▇▅▂▂ |
| Blancas | 0.136 | 0.065 | 0.021 | 0.083 | 0.135 | 0.188 | 0.244 | ▃▇▁▇▅ |
| Huevos | 0.048 | 0.019 | 0.005 | 0.044 | 0.047 | 0.061 | 0.079 | ▃▃▇▇▆ |
| Leche | 0.299 | 0.126 | 0.083 | 0.193 | 0.308 | 0.409 | 0.593 | ▆▃▇▆▁ |
| Pescado | 0.072 | 0.060 | 0 | 0.033 | 0.056 | 0.099 | 0.248 | ▇▅▂▁▁ |
| Cereales | 0.567 | 0.194 | 0.326 | 0.427 | 0.492 | 0.706 | 1 | ▇▅▃▂▂ |
| Azucares | 0.072 | 0.028 | 0.007 | 0.051 | 0.079 | 0.097 | 0.112 | ▂▃▅▇▇ |
| Frutos | 0.050 | 0.035 | 0.008 | 0.023 | 0.038 | 0.079 | 0.135 | ▇▂▂▃▁ |
| Vegetales | 0.069 | 0.031 | 0.021 | 0.047 | 0.063 | 0.083 | 0.136 | ▃▇▃▂▂ |

*Nota: los valores fueron normalizados restándoles el valor mínimo y dividiéndolo por el rango. Los valores fueron truncados a 3 decimales.*

Previo a realizar los análisis multivariados, es prudente hacer un análisis descriptivo de los datos. En la Tabla 1 se presentan los estadísticos descriptivos del consumo de proteínas de los países europeos por tipo de alimento. Los valores fueron normalizados (escala de 1 a 0) ya que no fue posible obtener certeza de que los valores de las proteínas estaban en la misma escala. Se observa que el consumo medio de proteínas es mayor en los Cereales (0.567), siguiendo la Leche (0.299), luego las Rojas (0.170) y en cuarto lugar las Blancas (0.136). Sin contar los Cereales, el consumo de proteínas de parte de los países no sobrepasa el 0.5 por tipo de alimento. También, destaca que hay países que tienen valores altos (dentro de la distribución de cada tipo de alimento) en el consumo de proteínas provenientes de la Leche (Finlandia: 0.593), el Pescado (Portugal: 0.248) y los Frutos (Grecia: 0.135).

Al observar el detalle del consumo de las principales proteínas, Bulgaria es la que tiene mayor consumo de Cereales y Alemania occidental el menor consumo. Finlandia tiene el mayor consumo de Leches y Portugal el menor. En las carnes, Francia tiene el mayor consumo de Rojas y Yugoslavia el menor, en tanto, Austria tiene el mayor consumo de Blancas y Albania el menor.

**Segunda pregunta: ACP: Obtenga DOS Componentes Principales y grafique**

Tabla 2. Cargas y varianza explicada para solución de dos componentes para proteínas

|  |  |  |
| --- | --- | --- |
| **Variable** | **Componente 1** | **Componente 2** |
| Rojas | 0.302 | 0.056 |
| Blancas | 0.310 | 0.236 |
| Huevos | 0.426 | 0.035 |
| Leche | 0.377 | 0.184 |
| Pescado | 0.135 | -0.646 |
| Cereales | -0.437 | 0.233 |
| Azucares | 0.297 | -0.352 |
| Frutos | -0.420 | -0.143 |
| Vegetales | -0.110 | -0.536 |
| **Desviación estándar** | 2.001 | 1.278 |
| **Proporción de varianza** | 0.445 | 0.181 |
| **Proporción de varianza acumulada** | 0.445 | 0.626 |

*Nota: valores truncados a 3 decimales.*

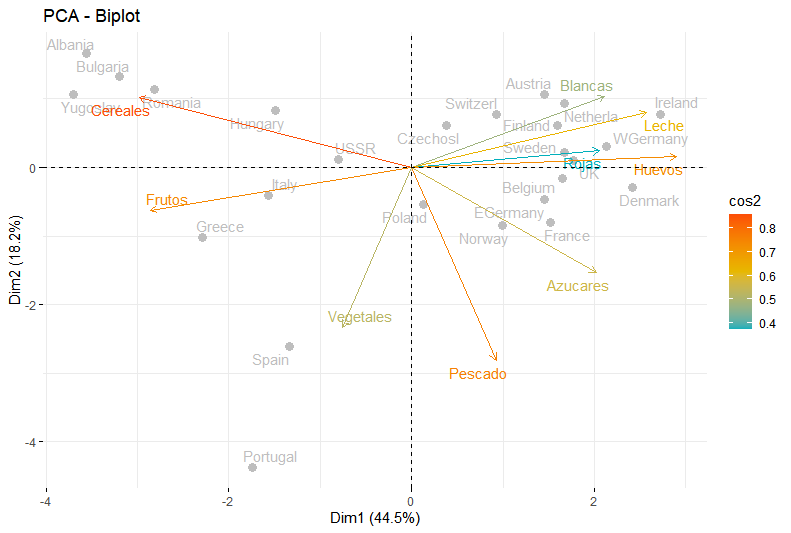
1. **¿Qué representa el primer componente? ¿retención de varianza con 2 CP?**

En la Tabla 2 se observa una solución de dos componentes utilizando la matriz de correlaciones para el consumo de proteínas de los países por tipo de alimento. Los valores absolutos de las cargas sugieren que las principales contribuciones a la primera componente provienen de los Cereales (0.437), los Frutos (0.420), los Huevos (0.377) y las carnes Blancas (0.310). En cambio, las contribuciones para la segunda componente se concentran en Pescado (0.646), Vegetales (0.536) y, en menor medida, en Azucares (0.352) y carnes Blancas (0.236). En términos sustantivos, y considerando el signo de las cargas, se observa que la primera componente caracteriza una dieta con un alto consumo de proteínas en base a huevos y carnes blancas, y un bajo consumo de proteínas en base a cereales y frutos. En tanto, la segunda componente sugiere una dieta con un bajo consumo de pescado y vegetales, y un mayor consumo de azucares y carnes blancas.

En lo que respecta a la varianza explicada por las componentes, se observa que la primera componente explica el 44.5% de la varianza en el consumo de proteínas. En tanto, la segunda componente explica el 18.1% de la varianza del consumo de proteínas de los países. En total, ambas componentes explican el 62.6% de la varianza, lo cual es un valor aceptable considerando las ganancias del análisis, a saber: reducción de dimensionalidad y posicionamiento de las variables y los países en un plano cartesiano.

1. **¿Es posible construir una “agregación” que permita identificar regiones?**

Sí, es posible construir una agregación que permita identificar regiones. Para esto se calculan los puntajes (*scores*) que tendría cada país para cada componente y se gráfica en un plano bidimensional. En concreto, en la Figura 1 se presenta un *biplot* para la solución de dos componentes para el consumo de proteínas de los países. En este gráfico se presentan, por un lado, las contribuciones de las variables (tipos de alimentos) a cada componente. Esto se realiza con el cociente al cuadrado (*cos2*), donde a mayor *cos2*, mejor representación de la variable en el componente. Por otro lado, el gráfico sitúa los países dentro del mapa, permitiendo describir su comportamiento en términos del consumo de proteínas.

Figura 1. Biplot para solución de dos componentes para consumo de proteínas por país.

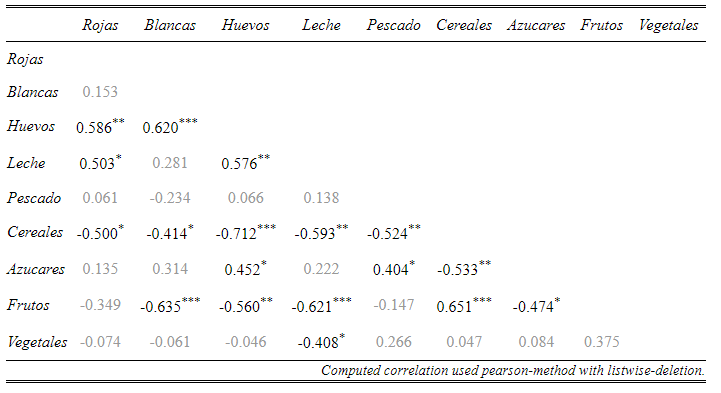
En términos de la contribución de las variables a los componentes, la Figura 1 muestra gráficamente lo que sugerían las cargas. Por un lado, el componente 1 se caracteriza por una dieta en donde el mayor consumo de proteínas viene de la leche, los huevos, las carnes blancas y, en menor medida, de carnes rojas. Así también, esta componente se caracteriza por menores niveles de proteína provenientes de cereales y frutos. En lo que respecta a la clasificación los países, esto implica que países con alto consumo de proteínas provenientes de alimentos como leche o huevos, tendrán un menor consumo proveniente de cereales o frutos, y viceversa. Por ejemplo, en una escala estandarizada de 0 a 1, Irlanda cuenta con un consumo proteína proveniente de la leche del 0.443 y un consumo proveniente de frutos del 0.024. En contraste, Albania cuenta con un consumo de proteínas provenientes de la leche de 0.153, y un consumo proveniente de cereales del 0.745. De esta manera, podemos afirmar que, en general, países como Irlanda, Países Bajos y Austria se caracterizan por este tipo de dieta, estando más cerca del consumo de proteínas en base a huevos, leche y carnes. En tanto, países como Albania, Bulgaria y Yugoslavia, también se caracterizan por este tipo de dieta, pero estando más cerca del consumo de proteínas en base a cereales.

Por otro lado, la segunda componente muestra un segundo tipo de dieta, la cual tiene por principal característica un bajo consumo de proteínas en base a pescado y vegetales. Este segundo componente agrupa el comportamiento de Portugal y España, en tanto sus valores originales en el consumo de proteínas en base a vegetales rondan el 0.1. Sin embargo, en comparación a la primera componente, esta segunda componente no es tan informativa, lo cual va de la mano con el porcentaje de varianza que explica (18.2%).

En suma, la solución con dos componentes presentada permite reducir la dimensionalidad del análisis, pasando de nueve variables a dos, explicando un poco más del 60% de la varianza. La primera componente representa una dieta caracterizada por un alto consumo de proteínas en base huevos, leches o carnes y menor consumo de proteína en base a cereales o frutos. La segunda componente, un tanto más difuso en términos de sus características, representa un menor consumo de vegetales y pescado.

**Tercera pregunta: AFAC: Lleve a cabo un Análisis Factorial.**

Tabla 3. Correlaciones del consumo de proteínas por tipo de alimento.



Previo a realizar un análisis factorial, es importante determinar si los datos cumplen las condiciones para su realización. Para ello se analizarán las correlaciones entre las variables, se hará una prueba de KMO y una prueba de Bartlett. En primer lugar, la Tabla 2 sugiere a grandes rasgos que los datos son adecuados para un análisis factorial, en tanto existe una cantidad considerable de correlaciones por sobre los umbrales esperados para constructos relacionados (> 0.50). En segundo lugar, la prueba KMO da un valor general de 0.65, lo cual está unos cuantos puntos por arriba del umbral mínimo aceptado (> 0.60) para realizar una prueba factorial. En tercer lugar, la prueba de Bartlett entrega evidencia para rechazar la hipótesis nula de que la matriz de correlaciones es una matriz de identidad (valor p = 3,405614-e10 < 0,05). En conclusión, en base a la tabla de correlaciones, la prueba KMO y la prueba de Bartlett, es posible afirmar que los datos son adecuados para la realización de un análisis factorial.

Posterior a determinar la adecuabilidad de los datos, es necesario determinar de cuántos factores debe ser el modelo. En base a la Figura 2 se muestra la sugerencia de factores en base a un gráfico de sedimentación, tanto para un ACP, como para un AFE. La interpretación de los autovalores de para la línea de FA sugiere una solución de un factor, en tanto un factor estaría explicando la varianza de casi 4 variables aproximadamente. Por otro lado, la Figura 3 muestra un gráfico del análisis paralelo, el cual compara los factores sugeridos a partir de los datos observados en comparación a un conjunto de datos aleatorio sin correlación entre los datos. Según el análisis paralelo, se sugiere una solución de dos factores, en tanto la línea azul está por sobre la línea roja. Sin embargo, cabe destacar que el segundo factor tiene un autovalor menor a 1, es decir, explica menos varianza que una variable por si sola.

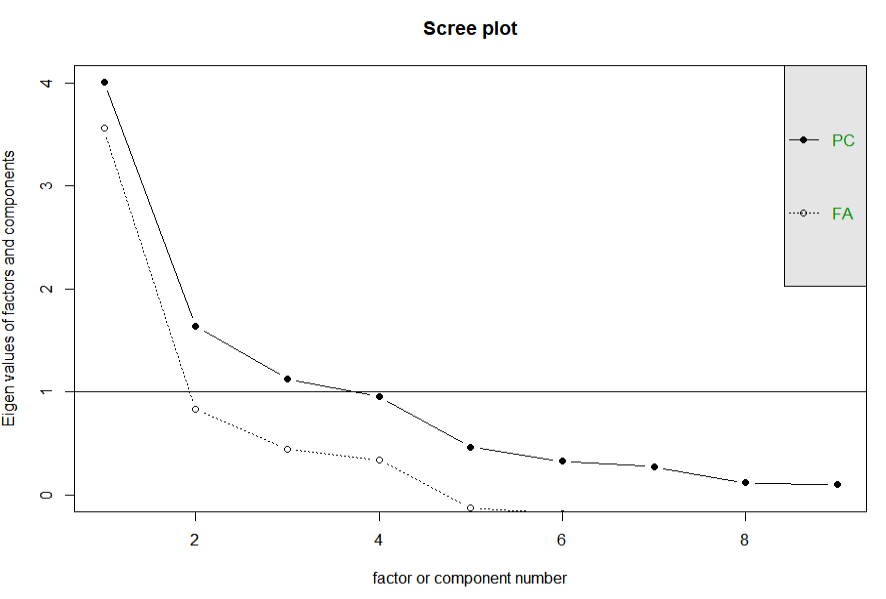


Figura 2. Gráfico de sedimentación.

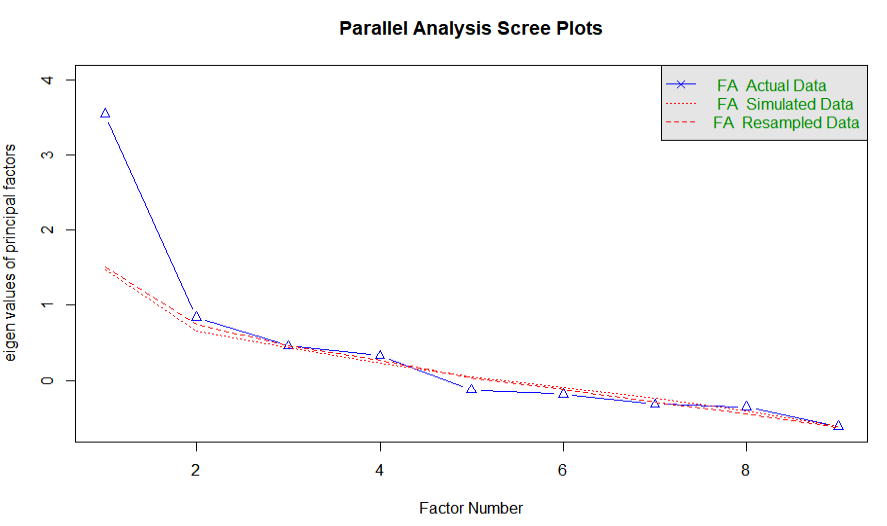


Figura 3. Gráfico de análisis paralelo

Considerando los resultados disimiles entre los gráficos de sedimentación y de análisis paralelo, se pondrán a prueba dos soluciones, una con un factor y otra con dos factores.

1. **¿Qué representa el primer factor? ¿cuántos factores proponen?**

Tabla 4. Solución de AFE con rotación *varimax* para 1 factor

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Factor 1** | **Comunalidad** | **Unicidad** |
| Rojas | 0.57 | 0.32 | 0.68 |
| Blancas | 0.58 | 0.333 | 0.67 |
| Huevos | 0.83 | 0.681 | 0.32 |
| Leche | 0.7 | 0.489 | 0.51 |
| Pescado | 0.28 | 0.081 | 0.92 |
| Cereales | -0.86 | 0.747 | 0.25 |
| Azucares | 0.55 | 0.298 | 0.7 |
| Frutos | -0.77 | 0.597 | 0.4 |
| Vegetales | -0.17 | 0.029 | 0.97 |
| **Proporción varianza explicada** | 0.4 |  |  |
| **Proporción varianza explicada acumulada** | 0.4 |  |  |

La Tabla 4 presenta una solución de análisis factorial exploratorio con rotación *varimax* para 1 factor. Los valores absolutos de las cargas factoriales estandarizadas sugieren que el factor único se ve compuesto en su mayoría por el consumo de proteínas en base a cereales, huevos, frutos y leche, en tanto presentan valores de 0.7 o mayores. En menor medida, las carnes (rojas y blancas) así como también las azucares también contribuyen al factor con cargas mayores a 0.5. La solución explica el 40% de la varianza, lo que es un valor bajo en base a los umbrales convencionales (> 60%), por lo que la solución de un factor podría no ser la mejor.

1. **Con la solución de dos factores, realice una rotación varimax y grafique.**

Tabla 5. Solución de AFE con rotación *varimax* para 2 factores.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Factor 2** | **Factor 1** | **Comunalidad** | **Unicidad** |
| Rojas | 0.56 | -0.11 | 0.33 | 0.672 |
| Blancas | 0.59 | -0.43 | 0.54 | 0.462 |
| Huevos | 0.84 | -0.2 | 0.74 | 0.257 |
| Leche | 0.68 | -0.07 | 0.47 | 0.534 |
| Pescado | 0.3 | 0.95 | 1 | 0.005 |
| Cereales | -0.9 | -0.27 | 0.88 | 0.116 |
| Azucares | 0.54 | 0.25 | 0.36 | 0.642 |
| Frutos | -0.76 | 0.09 | 0.58 | 0.415 |
| Vegetales | -0.14 | 0.32 | 0.13 | 0.874 |
| **Proporción varianza explicada** | 0.4 | 0.16 |  |  |
| **Proporción varianza explicada acumulada** | 0.4 | 0.56 |  |  |

Al observar la solución de AFE con dos factores (Tabla 5), vemos que se mantienen altas cargas factoriales estandarizadas en el factor 2, el cual se compone principalmente de cereales, huevos, frutos y leche, así en menor medida de carnes y azucares. En tanto, el primer factor parece componerse, en su mayoría, de las proteínas provenientes del pescado y en mucha menor medida de las carnes blancas. Esta solución sugiere que el consumo de proteínas de parte de los países manifiesta dos tipos de dietas. Una primera donde el consumo de proteínas proviene principalmente de fuentes de origen animal, especialmente de huevos y leche (factor 2) y, una segunda donde la proteína proviene principalmente del pescado. Dicho de otra forma, una dieta *ovo lacto* y una dieta *pescatariana*.

En términos de varianza explicada, ambos factores explican el 56% de la varianza, lo que está por debajo del umbral convencional (0.60). Considerando este escenario, cabría hacer análisis posteriores para ver si se puede mejorar el modelo, por ejemplo, extrayendo variables que no carguen lo suficiente en los factores.

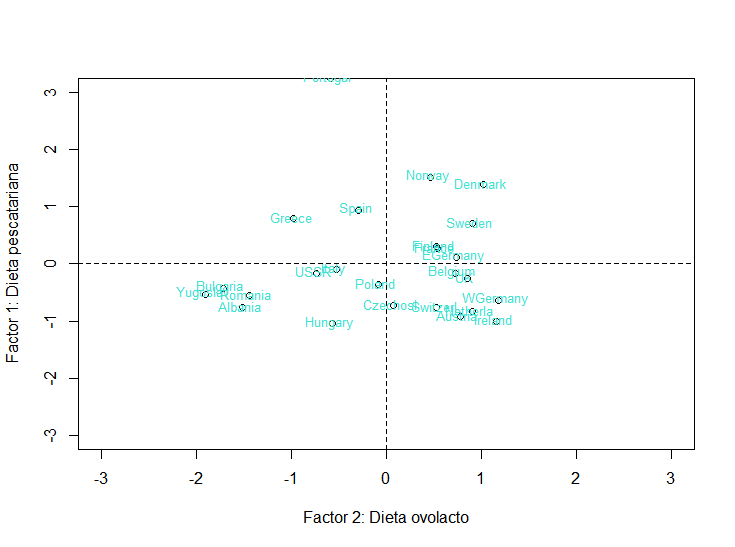


Figura 4: Puntajes factoriales para solución de dos factores con rotación *varimax*

En la Figura 4 se presenta un planto cartesiano que posiciona a los países en función de los puntajes factoriales (*scores*) del análisis exploratorio de dos factores. En general, se observan dos grupos de países, siendo los puntajes en relación al factor 2 (dieta ovo lacto o proteínas de fuente animal en general) las que generan la distinción. El primer grupo de países, correspondiente a Suiza, Irlanda, Bélgica y otros, representa un mayor consumo de este tipo de dieta, en tanto, un segundo grupo de países como Yugoslavia, Albania o Romania se caracterizan por un menor consumo de este tipo de dieta. El factor 1 (dieta pescatariana) es menos clara para separar a los países en relación al patrón de consumo de proteínas.

**Cuarta pregunta: Obtenga un número apropiados de “Cluster”, ¿cómo se agrupan los países?**

Al realizar un análisis de clúster, es importante tomar ciertas decisiones previas. Una primera decisión es si es necesario normalizar o estandarizar las variables a utilizar en el análisis. Generalmente, este procedimiento no es necesario si es que las variables comparten la misma escala, sin embargo, para el presente conjunto de datos no se sabe con certeza la unidad de medida de las proteínas. Por ende, se procedió a normalizar los datos restándole su valor mínimo y dividiéndolo por el rango. Una segunda decisión corresponde a la cantidad de clústeres que tendrá la solución. En base a la Figura 5, es posible afirmar que la solución óptima de clúster para los datos del consumo de proteínas de los países es de dos, según el método de *silhoulette*.

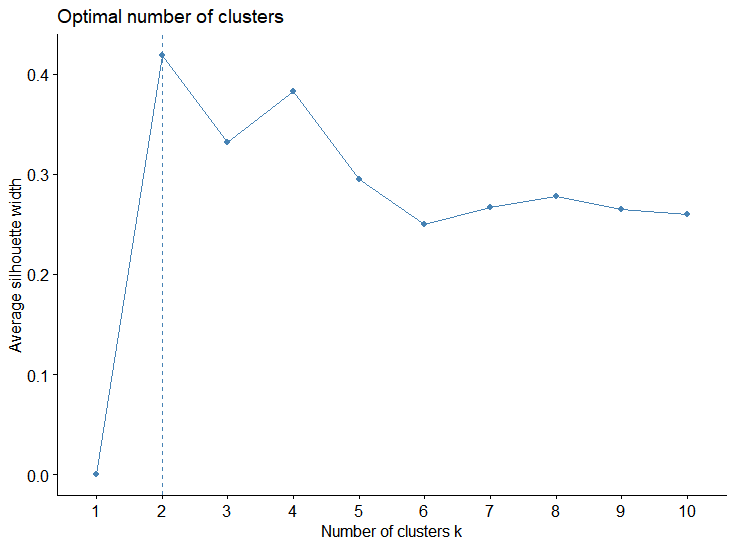


Figura 5. Número óptimo de clústeres.

Una vez conociendo el número estimado de clústeres a poner a prueba, una tercera decisión es el método por el cual se realizará la agrupación. En el último tiempo, los métodos más utilizados corresponden a aquellos no jerárquicos, en tanto pueden iterar para encontrar la mejor agrupación y permiten lidiar con la agrupación de casos poco distinguibles en un grupo en particular. Sin embargo, los métodos jerárquicos también tienen su merito al ser más claro el proceso de clasificación. Se pondrán a prueba ambos métodos.

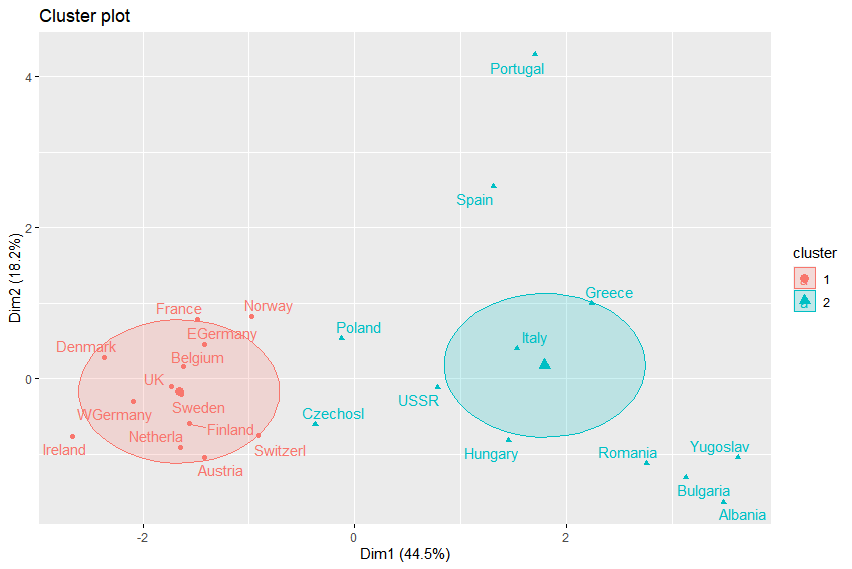


Figura 6. Análisis de clúster no jerárquico vía *k-means* para el consumo de proteínas de los países.

La Figura 6 muestra una solución de dos clústeres por vía del método no jerárquico de *k-means*. Acá se observa que hay un grupo bastante homogéneo en términos del consumo de proteínas, el cual agrupa a países como Finlandia, Reino Unido, Francia y Suiza. Por otro lado, hay un grupo un tanto más disperso que agrupa a países mediterráneos (ej. Grecia, Italia y España) y bálticos (Yugoslavia, etc.). Tomando en consideración antecedentes históricos, pareciese ser que el primer clúster agrupa, en su mayoría, a países con mayor industrialización, lo que podría explicar un mayor acceso a fuentes de proteínas de origen animal (huevos, leches, carnes) y por ende una mayor presencia en las dietas de estos países. En tanto, el otro clúster agrupa una cantidad de países que en mayor o menor medida han tenido una industrialización históricamente más lenta que los países del otro clúster, lo cual va de la mano con un acceso a fuentes de proteínas principalmente basadas en la agricultura, como los cereales, los frutos o los vegetales.

La realización de un análisis de clúster jerárquico (Figura 7) le da más sustento a la interpretación dada anteriormente. La Figura 7 muestra un primer clúster compuesto principalmente por países bálticos (Romania, Bulgaria y Yugoslavia) y un segundo gran clúster con los demás países. Si bien en términos de agrupación, este método pareciese no ser tan bueno, ya que entrega clúster bastante desproporcionados, también sugiere que los países bálticos son considerablemente distintos a los demás en términos de las fuentes de proteínas que contienen sus dietas. Esto va en línea con los antecedentes históricos, en tanto los países de Europa del Este mantuvieron una estructura principalmente agraria hasta fines del siglo XX, siendo países como Yugoslavia los mayores ejemplos.

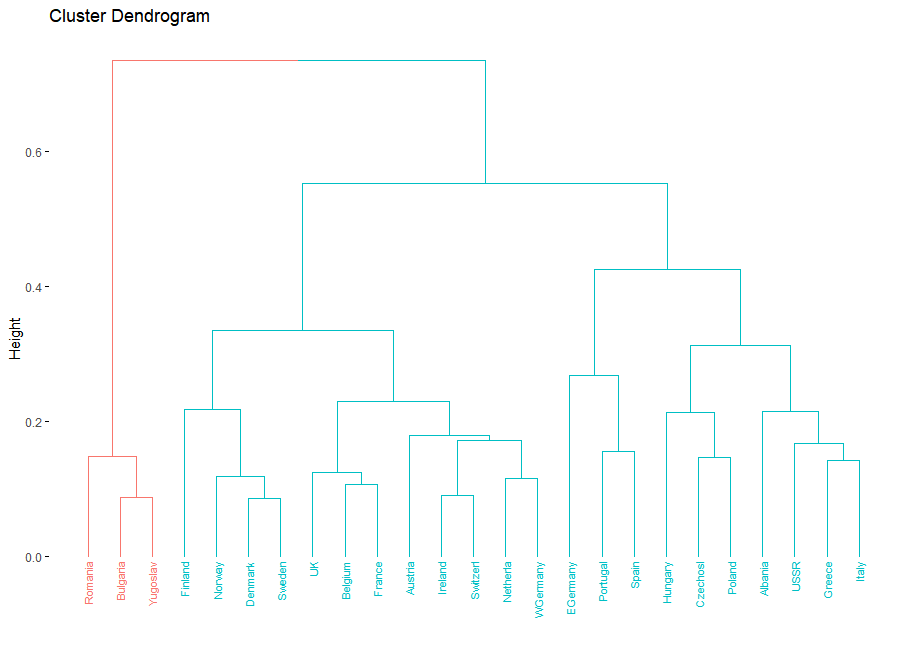
****

Figura 7. Análisis de clúster jerárquico por distancia euclidiana y método completo para el consumo de proteínas de los países.

En suma, ambos métodos de agrupación sugieren una diferencia entre los países en términos de las fuentes de proteínas en sus dietas y cómo eso va asociado a los cambios históricos (en términos de industrialización y estructura agraria) de cada grupo.

**Quinta pregunta: En no más de dos párrafos, contraste y discuta los resultados obtenidos por cada uno de los métodos**

En el presente control se aplicaron técnicas multivariadas a un conjunto de datos con mediciones del consumo de proteínas de los países según los tipos de alimentos. En concreto, se realizaron dos ejercicios: por un lado, la reducción en la cantidad de variables en base a ACP y AFE y, por otro lado, la agrupación de los países en relación a su consumo de proteínas en base a los puntajes de ACP y AFE, así como también en base a análisis de clúster. Se pueden contrastar los resultados obtenidos en relación al objetivo que tenía cada ejercicio. En lo que respecta a la reducción de variables, se observó que tanto ACP y AFE permitieron reducir la información de nueve variables en dos componentes/factores. Sin embargo, en ambas técnicas la explicación de varianza del segundo componente/factor fue considerablemente menor y tampoco permitía identificar un patrón claro en el consumo de proteínas. Si bien una solución de un componente/factor permitía hacer una lectura más clara del problema, tampoco daba cuenta de una gran cantidad de varianza (alrededor del 40% en ambos métodos). Por ello, si el objetivo es reducir complejidad, se sugiere avanzar en análisis posteriores eliminando variables que no aportan tanto a la construcción de los componentes o factores y así buscar obtener una solución de un componente o factor que explique más varianza que las soluciones propuestas en este control.

En lo que respecta a la agrupación de los países, todas las técnicas se encaminaron a un resultado similar: la existencia de un grupo de países caracterizado por una dieta con mayores niveles de proteína de origen animal y la existencia de un grupo de países caracterizado por una dieta con menores niveles de proteína animal y mayores niveles de proteínas de origen vegetal. Sin embargo, en términos de interpretabilidad, los resultados de ACP y AFE requerían de conocimientos previos y detalles propios de la técnica que pueden hacer más difícil la comunicación de los resultados. En contraste, el análisis de clúster, tanto no jerárquico, como jerárquico realizó una agrupación mucho más clara, permitiendo entregar más evidencia a las diferencias entre los grupos y proponer una interpretación al respecto (antecedentes históricos en la industrialización de los países). En suma, si bien los resultados de todas las técnicas siguieron un patrón parecido, el análisis de clúster parece ser el más idóneo en términos de cumplir el objetivo de agrupar a los países en grupos homogéneos y comunicar efectivamente los resultados.