

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

09 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:

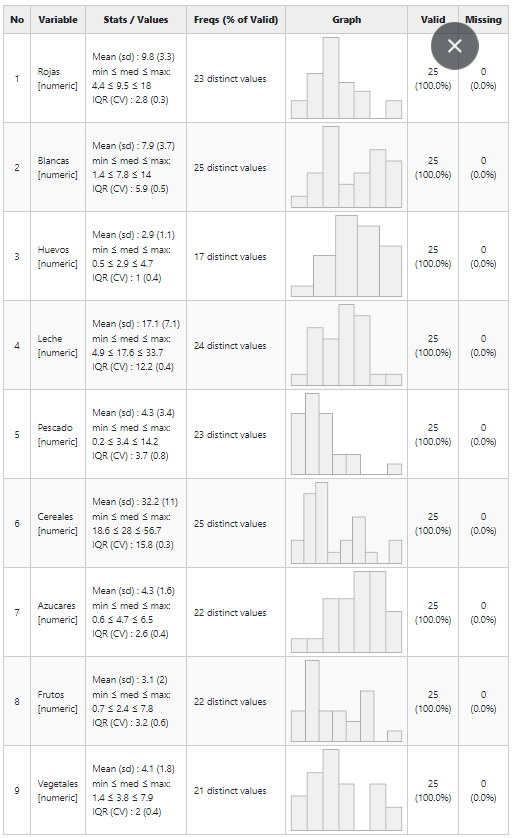
* Redacte un informe (pregunta/respuesta) que permita responder cada punto indicado.
* NO coloque el script en el informe (solo tablas, gráficos, métricas de interés, valores resultantes que ayuden a respaldar sus respuestas, no relleno).
* Por ningún motivo el último taller es pauta para este control, 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.**

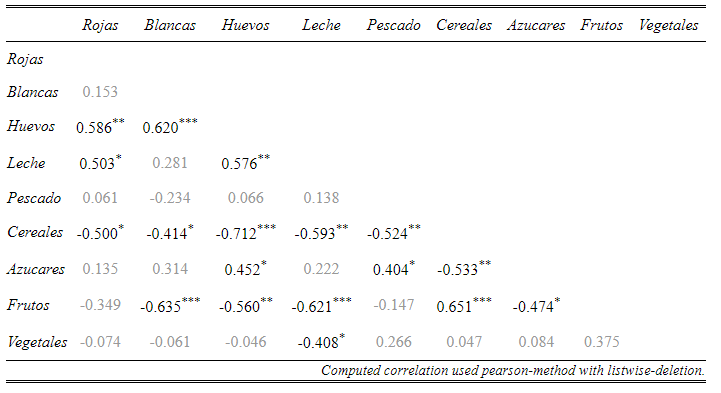
Previo a realizar los análisis multivariados, es prudente hacer un análisis descriptivo de los datos. En la Tabla 1 se presentan los descriptivos del consumo de proteínas de los países europeos por tipo de alimento. Se observa que el consumo medio de proteínas es mayor en los Cereales (32.2), siguiendo la Leche (17.2), luego las Rojas (9.8) y en cuarto lugar las Blancas (7.9). Al observar el detalle del consumo de proteínas de Cereales por país, se observa que el país con menor consumo de proteínas en esta categoría es Alemania Occidental con 18.6 y el país con mayor consumo de proteínas es Bulgaria con 56.6.

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

****

Tercera pregunta.

Tabla 2. 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 (> 0.5). En segundo lugar, la prueba KMO da un valor general de 0.65, lo cual está unos cuantos puntos por arriba del umbral esperado (> 0.6) 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 1 se muestra la sugerencia de factores en base a un gráfico de sedimentación, tanto para un PCS, como para un EFA. La solución con 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 2 muestra un gráfico paralelo, el cual compara los factores sugeridos a partir de los datos observados en comparación a una base de datos aleatoria 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 la varianza menor a un item.

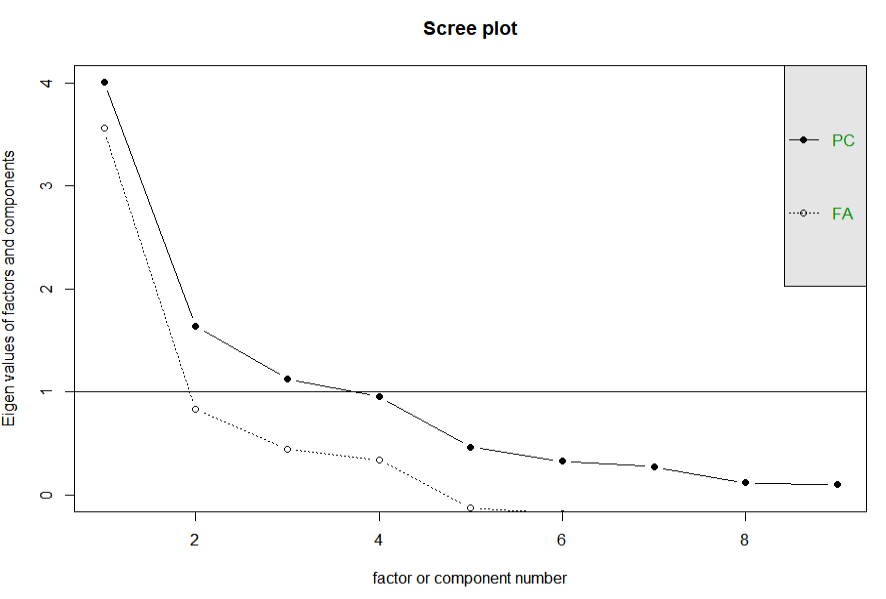


Figura 1. Gráfico de sedimentación.

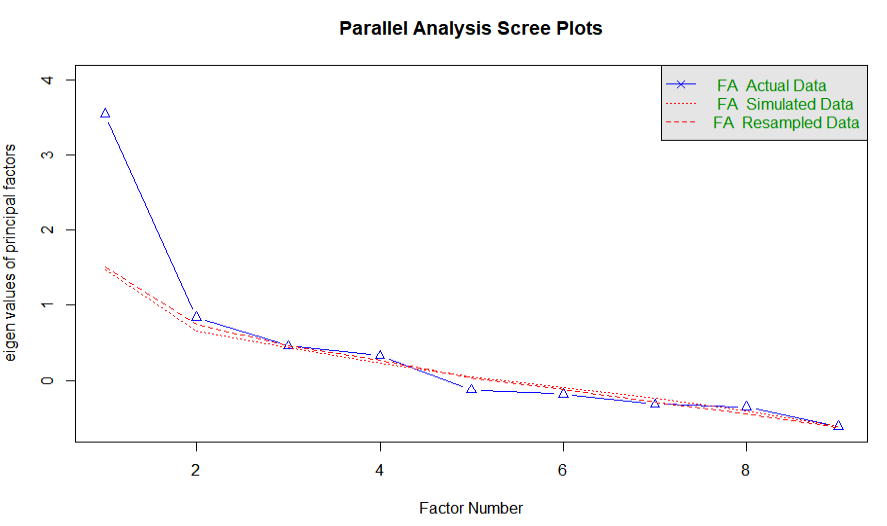


Figura 2. Gráfico de análisis paralelo

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

