# Archivo de Consultas Generales sobre Análisis Metagenómicos[1]

**Alejandro Navas González**

2023-06-16

[1]Alejandro Navas González

**1. ¿Por qué los modelos de correlación de Spearman y Pearson son subóptimos cuando se aplican sin la modificación de las abundancias relativas de los organismos?¿Cómo se haría eso?**

Los modelos de correlación de Spearman y Pearson son subóptimos cuando se aplican a las abundancias relativas de los organismos porque estos modelos asumen una distribución normal de los datos y la independencia entre las observaciones. Sin embargo, los datos de abundancia relativa de los organismos en un microbioma no suelen distribuirse normalmente y las observaciones no son independientes, ya que la abundancia de un organismo puede afectar a la de otros. Para modificar las abundancias relativas, se pueden utilizar transformaciones de los datos, como la transformación logarítmica o la transformación de raíz cuadrada, que pueden ayudar a normalizar la distribución de los datos y reducir la dependencia entre las observaciones.

**2. ¿Qué es un modelo de Markov?**

Un modelo de Markov es un tipo de modelo estadístico que se utiliza para representar sistemas que cambian con el tiempo. En un modelo de Markov, se asume que el estado futuro del sistema sólo depende de su estado actual y no de cómo llegó a ese estado. Esto se conoce como la propiedad de Markov o la suposición de "falta de memoria".

**3. ¿Qué es un boosted linear model (GBLM)?**

Un Boosted Linear Model (GBLM) es un tipo de modelo de regresión que utiliza el método de boosting para mejorar su rendimiento. El boosting es una técnica que combina múltiples modelos débiles (es decir, modelos que sólo son ligeramente mejores que una predicción aleatoria) para crear un modelo fuerte. En el caso de un GBLM, los modelos débiles son modelos lineales.

**4. Por lo general, es muy difícil determinar, sólo a partir de las mediciones de la abundancia relativa, si estas asociaciones negativas representan una verdadera anticorrelación (por ejemplo, un organismo que supera a otro) o el crecimiento excesivo de un organismo mientras que el resto de la población permanece invariable (lo que da lugar a una correlación negativa debido a la composición de estos datos). ¿Por qué?**

En los datos de microbiomas, las mediciones de abundancia relativa representan la proporción de un organismo en relación con el total de organismos en la muestra. Por lo tanto, si la abundancia de un organismo aumenta, necesariamente la abundancia relativa de al menos uno de los otros organismos debe disminuir, incluso si su abundancia absoluta no cambia. Esto puede dar lugar a correlaciones negativas que no reflejan interacciones reales entre los organismos.

**5. ¿Qué es el cálculo de disancias funcionales por Jaccard index of non-shared COG families?**

El cálculo de distancias funcionales por el índice de Jaccard de familias COG no compartidas es una medida de la disimilitud funcional entre dos muestras de microbiomas. Las familias COG (Clusters of Orthologous Groups) son grupos de genes que se cree que provienen de un ancestro común. El índice de Jaccard mide la disimilitud entre dos conjuntos como el complemento de la proporción de elementos que tienen en común.

**6. ¿ Qué son el método Simes y la corrección por Benjamini-Hochberg-Yekutieli false discovery rate (FDR)?**

El método de Simes y la corrección por Benjamini-Hochberg-Yekutieli para la tasa de falsos descubrimientos (FDR) son técnicas utilizadas para controlar la tasa de falsos positivos en las pruebas de hipótesis múltiples. El método de Simes es un procedimiento para ajustar los valores p en las pruebas de hipótesis múltiples. La corrección de Benjamini-Hochberg-Yekutieli es un procedimiento para controlar la FDR, que es la proporción esperada de errores de tipo I (falsos positivos) entre las hipótesis rechazadas.