Problema de estadística descriptiva - 29/10/2020

En bioinformática, se emplea una técnica de simulación llamada "docking" para evaluar fácilmente si moléculas de interés se podrían unir o no a proteínas u otras moléculas "blanco". Esto sirve para seleccionar de entre miles de candidatos posibles para el desarrollo de nuevos fármacos, por ejemplo, sólo aquellas moléculas que sean más promisorias.

Ver por ejemplo: Arcon et al, 2017, https://pubs.acs.org/doi/10.1021/acs.jcim.6b00678.

Para evaluar esto, se emplean funciones de "scoring", que intentan predecir la afinidad de estas moléculas por una sección blanco en una proteína de interés. Para evaluar si una función de scoring es adecuada o no, se deben calibrar con moléculas de las que ya se conozcan los parámetros adecuados. En este caso, se comparan tres funciones de scoring, evaluadas sobre una base de datos con 100 moléculas ya ensayadas. Es deseable que - en general, para las moléculas de interés, estas funciones tengan un valor que no diste de 1 en más de $\pm 0,1$ (1 representa una afinidad de referencia). - los valores deben tener una distribución aproximadamente normal (que apoyaría la hipótesis de que los valores de scoring sólo tienen un error aleatorio, y no hay ningún sesgo).

1) Calcular medidas de centralidad para los valores de cada una de estas funciones de scoring ensayadas sobre el banco de datos de referencia:

- media,
- mediana,
- media α-podada para $\alpha = 0.1, 0.2$.

Comparar los valores obtenidos para cada función. ¿Qué diferencias observa? ¿Hay alguna función de scoring que podríamos considerar peor que las demás?

2) Obtener los percentiles 10, 25, 50, 75 y 90 y los valores máximos y mínimos, para cada una de las funciones de scoring. Comparar los valores obtenidos.

3) Calcular medidas de dispersión para estos tres conjuntos de datos:

- desvío estándar,
- rango intercuartil o intercuartílico (IQR),
- MAD (mediana de la desviación absoluta).

Comparar los valores de dispersión obtenidos. ¿Cuál de las funciones parece tener valores menos dispersos?

4) Construir histogramas que permitan visualizar los valores de scoring para cada función. ¿Qué observaciones haría sobre la distribución de estos valores?. ¿Alguna de ellas parece bimodal? ¿En alguna de ellas parece haber valores atípicos o outliers?

¿Los valores de scoring se hallan en el rango deseado? ¿Hay alguna asimetría en la distribución de los valores de una función? ¿En algún caso el ajuste normal parece razonable?

5) Graficar los box-plots correspondientes. ¿Cómo se compara la información que dan estos gráficos con la obtenida con los histogramas? En base a los gráficos obtenidos, discutir simetría, presencia de outliers y comparar dispersiones.

6) Graficar los qqplots correspondientes. ¿En algún caso el ajuste normal parece razonable?

7) En base a todo el análisis anterior, ¿cuál sería la función de scoring que más se ajusta a los requerimientos pedidos?