Modelación Estadística

Actividada 5: Ajuste a distribución Normal y homogeneidad de varianzas

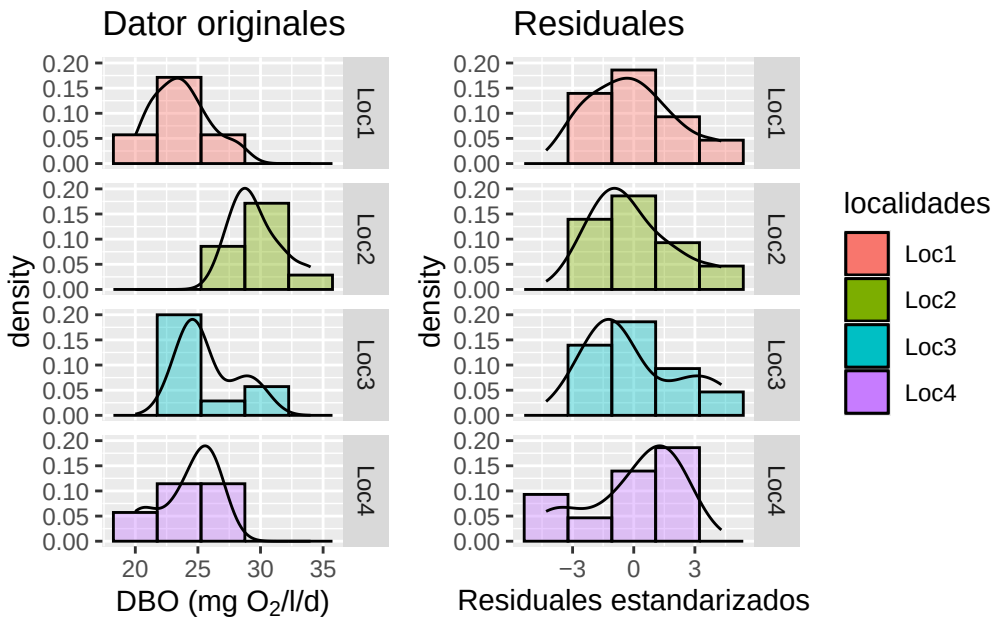
Prof. Edlin Guerra, Prof. María Muciño

Continuamos con el caso de la planta procesadora de celulosa en Uruguay. Recordemos que estamos interesados en evaluar la Demanda Bioquímica de Oxígeno (DBO) en cuatro localidades a lo largo de dos ríos: (1) río Uruguay arriba (zona carente de contaminación por los efluentes de la planta), (2) en los alrededores de descarga de la planta en el río Uruguay, (3) en la desembocadura del rió Uruguay sobre el río de la Plata, y (4) en la desembocadura del Río de la Plata sobre el océano Atlántico (Figura 1). Para validar la interpretación que hicimos del ANOVA, es necesario verificar que se cumplen los supuestos de normalidad y homogeneidad de varianzas.

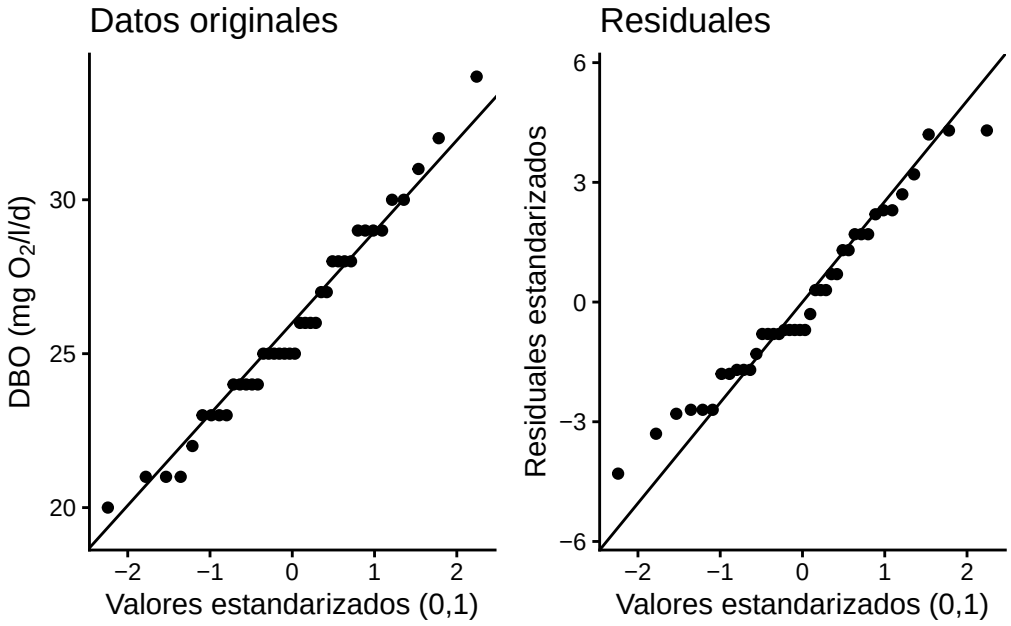
PARTE 1: Evaluación de Normalidad

1. Exploración gráfica

- Ajusta un modelo lineal de la DBO en función de la localidad como en la actividad anterior, nombrando al objeto mod1.
- Explora el objeto mod1 y extrae los residuos como una nueva columna de datos que combinarás con la matriz de datos original. Usando códigos dplyr, puedes estimar el promedio de los residuos por cada localidad ¿qué obtienes?
- Seguidamente, construye dos histogramas: uno para la DBO y un histograma de residuos, en ambos casos usando la función básica `hist`. ¿qué puedes apreciar? ¿en qué se diferencian ambos histogramas?
- Luego elabora histogramas estéticamente más elegantes con `ggplot2`, usando `geom_histogram` y `geom_density`, considerando cada localidad. Debes obtener algo así:



- Seguidamente, genera dos diagramas Q-Q de los valores de DBO y de los residuos usando la funciones `qqnorm` y `qqline` . Interpreta visualmente: ¿se parecen a una distribución normal? Como complemento exploratorio, repite los gráficos Q-Q con `ggplot2`, usando las funciones `geom_qq` y `geom_qq_line()`. Deberías obtener algo así:



2. Prueba de normalidad Shapiro-Wilk

¿Cuál es la hipótesis estadística nula de la prueba? ¿Qué significa rechazarla?. Realiza el cálculo manual de Shapiro-Wilk con la fórmula provista en la guía. Luego, aplica la prueba en R con la función `shapiro.test` de la paquetería básica de R (no deben instalar nada). ¿Coinciden los cálculos?

3. Procedimientos de D'Agostino y Pearson (usando paquete moments)

Calcula manualmente los coeficientes de asimetría y curtosis con las fórmulas provistas y complementa la prueba de D'Agostino y Pearson. ¿Qué forma tienen los residuos? ¿existen evidencias probabilísticas para señalar que la distribución es simétrica? ¿Qué forma tienen los residuos? ¿existen evidencias probabilísticas para señalar que la distribución es mesocúrtica?

Confirma que tu resultado fue correcto desarrollando este análisis con el paquete `moments`. Específicamente, puedes calcular simetría (`skewness`) y curtosis (`kurtosis`) y desarrollar una prueba de hipótesis para cada una. Alternativamente, puedes construir una prueba formal para ajuste a la normalidad con la función `jarque.test`. Preguntas: ¿Cómo se relacionan asimetría y curtosis con la normalidad? ¿coinciden los resultados de las pruebas de Shapiro-Wilk y los procedimientos de D'Agostino y Pearson?

PARTE 2: Evaluación de Homogeneidad de Varianzas

1. En la actividad previa hicimos varios gráficos exploratorios de dispersión de datos y tendencia central, que llamamos `fig1`, y que fuimos mejorando estéticamente con capas de información. Reconstruye este gráfico y analiza: ¿se parecen sus varianzas? ¿Qué observarías si las varianzas fueran muy diferentes?

2. Apliquemos una prueba de varianzas muy popular: la prueba de Cochran. Para ello, usa la función `cochran.tets{outliers}`. Pregunta: ¿Cuál es la hipótesis nula en esta prueba? ¿Cómo interpretarías un resultado significativo?

3. La prueba de Levene es un ANOVA aplicado a los valores absolutos (todos positivos) de los residuos del modelo. A diferencia de un ANOVA convencional, la hipótesis nula es específicamente sobre las varianzas de cada grupo, sometiendo a prueba que son todas iguales. Aplica un modelo ANOVA a los residuos absolutos y decide sobre la hipótesis nula. Seguidamente, aplica la función `leveneTest{car}` y compara el resultado con el anterior ¿coinciden los resultados? ¿hay evidencias suficientes para indicar que las varianzas de la DBO son iguales entre las localidades? ¿Por qué se considera a la prueba de Levene más robusta que la de Cochran?

PARTE 3: Reflexión Integrada

Resume los resultados obtenidos y discute:

1. ¿Se cumplen los supuestos de normalidad?
2. ¿Se cumple el supuesto de homogeneidad de varianzas?
3. ¿Qué implicaciones tienen los resultados para la validez del ANOVA aplicado en la actividad anterior?