Modelación Estadística. Lab2: Modelo lineal y Prueba de hipótesis simples

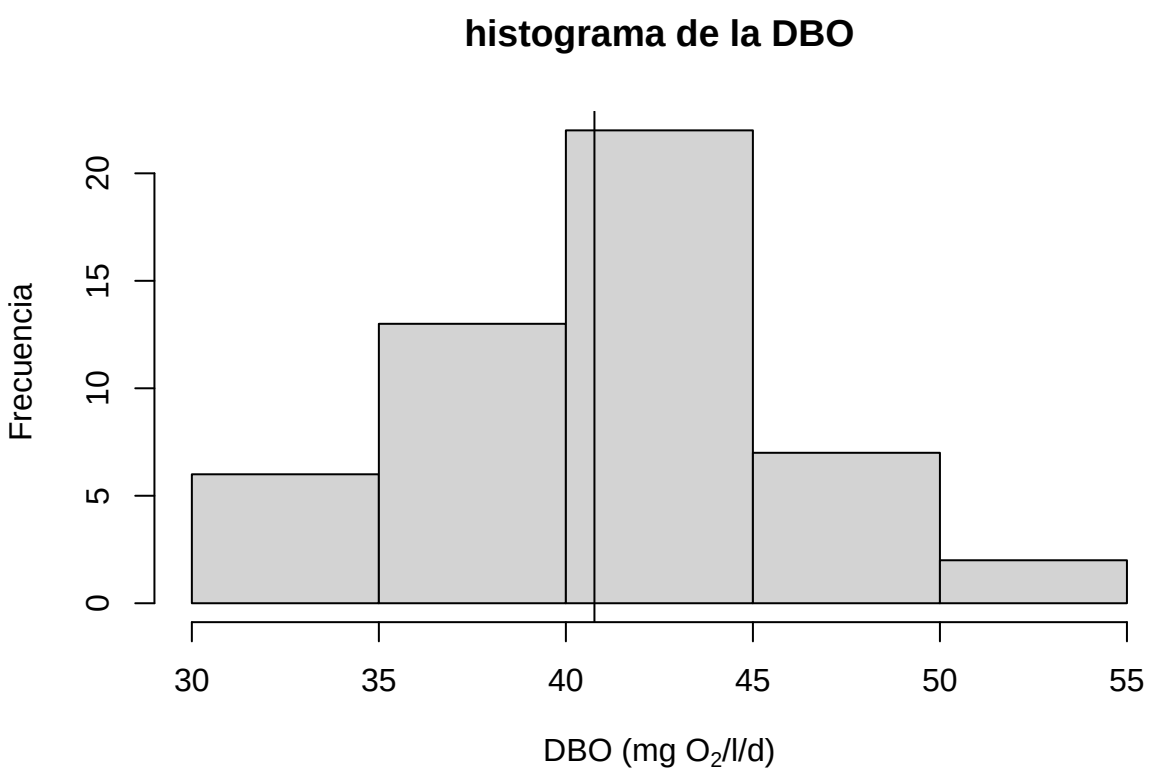
Edlin Guerra Castro

23/2/2022

Distribución t-student: Ejericio 1 - Una muestra

Caso hipotético. Se requiere evaluar si una planta procesadora de celulosa está respetando los límites máximos permisibles de contaminantes en las descargas de aguas residuales vertidas a un río cercano. La normativa ambiental Mexicana especifica varias variables y límites máximo permisible para cada una, entre las que destaca la Demanda Bioquímica de Oxígeno (DBO), cuyo promedio máximo permitido es de 30 ml O2/l por día. La normativa indica que el promedio debe ser estimado de una muestra considerando 50 mediciones aleatorias en un lapso de tres meses de completa operatividad de la planta.

1. Importa los datos del archivo **datos1.csv** al ambiente R con el nombre **datos1** y verifica sus características y estructura. Asegúrate que el archivo esté en el ambiente de trabajo.
2. Calcula la media (con la función `mean`) y desviación estándar (con la función `sd`) de la DBO en la muestra y obtén un histograma (con la función `hist`) para ver su distribución. ¿Te parece que tienen una distribución normal? Confirma que los resultados que obtienes y el gráfico queden así:



3. ¿cuál es el error estándar de la media? ¿Es un valor grande o pequeño? Confirma que este es el resultado:

```
## [1] 0.6638794
```

4. Define la hipótesis alternativa y luego la Hipótesis nula. Aplica una prueba de t para una sola muestra, usando el comando `t.test`, pero antes de efectuar la prueba debes definir el alfa para rechazar la Hipótesis nula. Utiliza la ayuda `help` para escribir los argumentos correctos de `t.test` (tip: los más importantes son *mu* y *alternative*)

5. A partir del resultado obtenido (output o salida) responde a las siguientes preguntas:

   a) ¿Cuál fue el valor de t que obtuviste?
   b) ¿Cuál es el valor de probabilidad asociado a ese valor de t?
   c) ¿Existen evidencias suficientes para considerar que DBO de las aguas redisuales superan los valores de la norma?
   d) ¿Cuáles son los límites inferior y superior del intervalo de confianza al 95%?

6. Vamos a calcular el intervalo de confianza paso por paso. Para ello requerimos contar con el valor de t que separa la curva en zonas de baja y alta probabilidad de ocurrir (dos colas, cada una de prob alfa/2). Este valor se obtiene con la función `qt` y requiere de tres argumentos esenciales (p = 0.025, df = 49, lower.tail = TRUE. Usa el `help` para tener más detalles de esta función).

   a) ¿Son un error los valores negativos de esta función? ¿Por qué?
   b) ¿Cómo puedes obtener un valor positivo?
   c) ¿Por qué razón usamos una probabilidade de 0.025, si el intervalo de confianza es de 95%?
   d) ¿Por qué usamos n-1 en vez de n? Ahora, multiplica el valor de t-student (absoluto) por el valor de error estándar calculado antes.
   e) ¿Qué obtuviste con esta operación?

f) ¿Cuáles son las unidades de esta medida?
g) Finalmente suma y resta el intervalo (al 95%) a la media para conocer los límites superior e inferior.
h) Compara los valores obtenidos con el resultado de la función usada para hacer la prueba de hipótesis.

7. Calcula el intervalo de confianza al 99%

   a) ¿En cuál de los dos intervalos (95% o 99%) tengo más confianza de encontrar la media si repito el experimento?
   b) ¿Cuál de los dos intervalos es más ancho?
   c) ¿Qué le sucedería al intervalo si en vez de 50 mediciones, tuviese solo 5 mediciones?
   d) Verifica tu conclusión (PISTA: usa la función `qt`)

8. ¿Qué sucedería con el IC al 95% si la DBO de los efluentes fuese mas variable? Obtén un histograma y verifica tu conclusión mediante una simulación usando la función `rnorm` (PISTA: mantén la muestra simulada en n=50, y la media en 40).

Distribución t-student: Ejericio 2 - Dos muestras

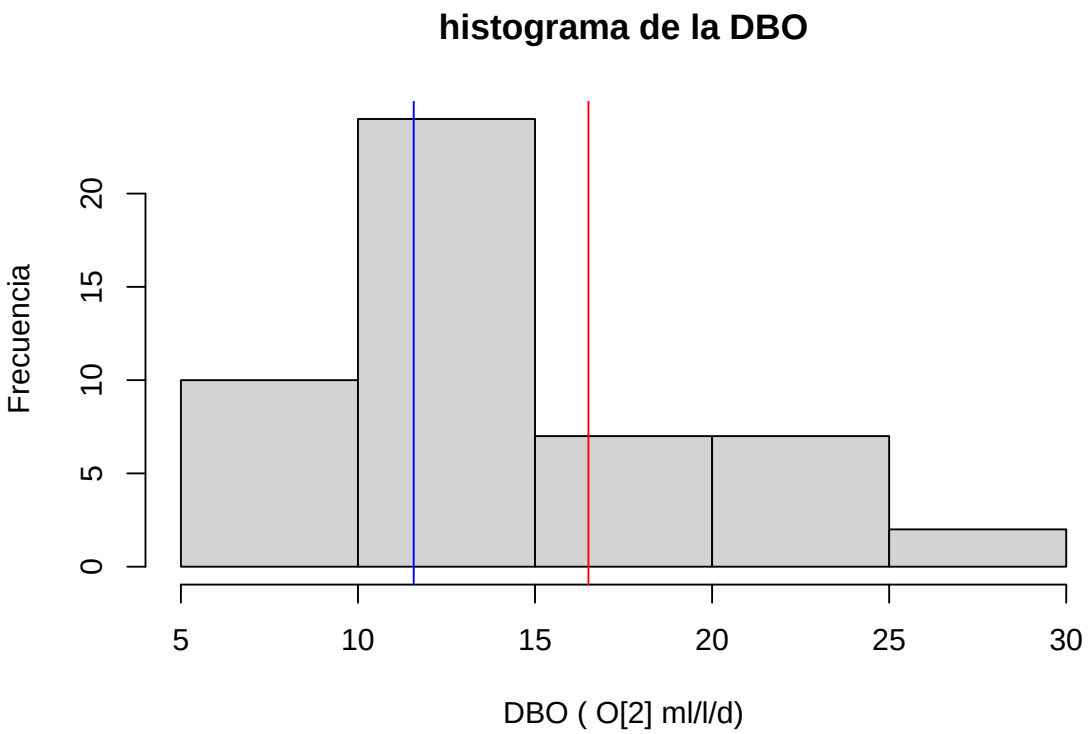
Continuamos con el caso hipotético. Con el análisis anterior se evidenció que la planta procesadora de papel vierte agua con mayor DBO de la que establece la normativa Mexicana. Sin embargo, esto no implica que haya ocurrido un impacto ambiental, ya que habría que comprobar el impacto (cambio en la DBO antes y después de la operatividad de la planta). Previendo esto, antes de la puesta en desarollo de la planta, se efectuó un Estudio Ambiental de Línea Base, en el que se midió la DBO tres meses antes de operaciones, y luego a los seis meses de operaciones.

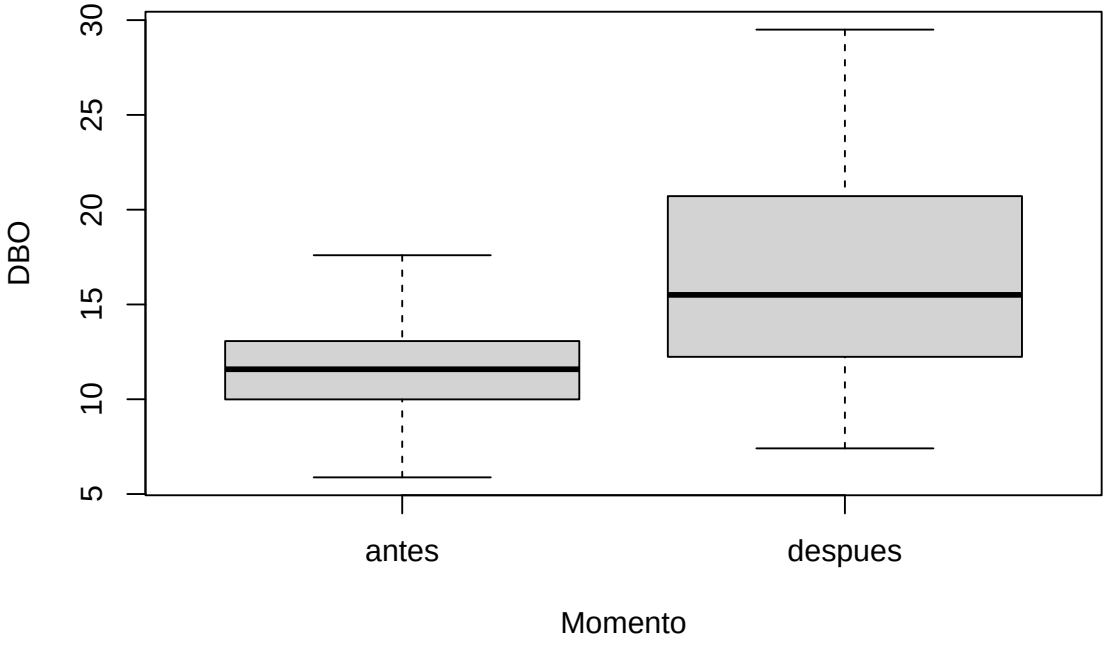
1. Importa los datos del archivo **datos2.csv** al ambiente R con el nombre **datos2** y verifica sus características y estructura. Asegúrate que el archivo esté en el ambiente de trabajo.

2. Obtén los nombres de las variables en el archivo y sigue las siguientes instrucciones:

   a) Identifica el vector que contiene la variable que se pretende analizar, y establece si ésta es contínua o discreta.
   b) Identifica el vector que contiene el factor 'origen'. ¿Cuántos niveles tiene dicho factor?
   c) ¿Cuál es la unidad de observación en este estudio?
   d) ¿Cuántas réplicas fueron medidas en cada nivel? ¿Consideras este diseño uno balanceado?

3. Contruye un histograma con `hist` para para la variable `DBO` sin distinguir el momento, pero agregue dos líneas verticales para cada promedio. Luego obtén un gráfico de cajas-bigotes agrupando datos según el *Momento*. PISTA: usa la función `boxplot`, busca cómo se usa esta función. ¿cuál representación gráfica te parece más informativa? Explica la respuesta. Confirma que obtienes estos gráficos:





4. Describe las características de ambos momentos en términos de sus medias, dispersiones y la forma aproximada de sus distribuciones. Ayudate calculando estos estadísticos con las funciones `summary`,

`mean`, y `sd`.

a) ¿Cuáles de estas características son parecidas y cuáles diferentes entre las dos muestras?
b) ¿cuál es el error estándar de la media de cada momento?
c) ¿qué significa esto respecto a cada momento?

5. Formula la hipótesis del modelo y la hipótesis nula, y define si se trata de una prueba de **una** o **dos** colas. Obten el valor crítico de la distribución de t-student para rechazar la hipótesis nula. Considera un alfa = 0.05, y grados de libertad = n1 + n2-2.

6. Aplica una prueba de t-student para dos muestras independientes (*S3 method for class formula*), y responde a las siguientes preguntas:

a) ¿Cuál es el valor de t obtenido? ¿Es grande o pequeño?
b) ¿Cuál es la probabilidad asociada al valor de t obtenido? ¿Es grande o pequeña?
c) ¿Cuál es la conclusión de la prueba que aplicaste?
d) ¿Podrías enlistar tres de las condiciones más importantes que deben cumplir los datos para interpretar adecuadamente e resultado de una prueba t clásica?

7. Considera que las varianzas de estas dos muestras son lo suficientemente similares como para asumirlas iguales, y aplica una prueba de t-student para varianzas iguales. (PISTA: revisa la informacion del `help` de `t.test`)

8. Compara los resultados de ambas pruebas y responde a las siguientes preguntas:

a) ¿Qué ocurre con el valor de t cuando asumes homogeneidad de varianzas aún cuando no la hay?
b) ¿Qué ocurre con el valor de probabilidad asociado al estadístico t cuando asumes homogeneidad de varianzas y no la hay?
c) ¿Qué ocurre con los grado de libertad?
d) ¿Cuál de las dos pruebas es más conservadora (i.e., tiene menos probabilidad de caer en errores tipo I)?
e) ¿Confías más en el resultado de la primera prueba o en el de la segunda? Explica tu respuesta.