Estadística (Q)
Listado de Ejercicios de las clases prácticas

---

1. Supongamos que tiramos dos dados (de 6 caras, equilibrados).

   (a) Describir un espacio muestral de este experimento que sea equiprobable.

   (b) Calcular la probabilidad de que en las dos tiradas hayan salido un 4 y un 1 en algún orden.

   (c) Calcular que en las tiradas hayan salido dos 5.

   (d) Calcular la probabilidad de que el número obtenido en la segunda tirada sea estrictamente mayor al obtenido en la primera.

2. Tenemos una bolsa con 7 bolitas blancas y 4 bolitas rojas. Sacamos 3 bolitas al azar con reposición.

   (a) Describir un espacio muestral de este experimento que sea equiprobable.

   (b) Hallar la probabilidad de que en las dos primeras extracciones hayan salido bolitas rojas y en la última haya salido una bolita blanca.

   (c) Calcular la probabilidad de que haya salido al menos una bolita roja en las tres extracciones.

   (d) Repetir los items (a) y (b) para extracciones sin reposición.

   Definición : número combinatorio

   Dados n objetos distintos, nos preguntamos cuántas formas distintas hay de elegir k de ellos (sin importar el orden de los elegidos). La respuesta a esta pregunta es

   $$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}.$$

   donde $n! = 1.2.3 \ldots n$. Por ejemplo, si tenemos 5 objetos distintos y solo nos dejan elegir 2 de ellos para llevar a una isla perdida, tenemos $\binom{5}{2} = 5!/(2!\,3!) = 10$ posibilidades. Si queda tiempo vamos a justificar esto.

3. En un campamento hay 12 chicos y los queremos dividir en tres equipos: rojo, blanco y azul. ¿Cuántas formas distintas hay de hacer la división si queremos que cada equipo tenga exactamente 4 personas?

4. El equipo A jugó 10 partidos en un torneo, en los que obtuvo 7 éxitos (o triunfos) y 3 fracasos (o derrotas).

   (a) ¿De cuántas maneras pudieron haber ocurrido estos resultados a lo largo del torneo, tomando en cuenta el orden?

   (b) Responder a la pregunta anterior si en total hubiera obtenido 6 éxitos y 4 fracasos.

   (c) Ídem con 9 éxitos y 1 fracaso.

   (d) Ídem con 10 éxitos.

5. Se analizan 100 muestras de tres variedades de jugo (A, B, C) y se las clasifica según su contenido energético (medido en kcal. por 100ml.) en dos niveles: alto o bajo. El resultado del conteo de muestras según esta clasificación se presenta, en forma incompleta, en la siguiente tabla

	A	B	C	Total
Alto	15		20	55
Bajo	5	25		
Total		45	35	100

(a) Completar la tabla.

(b) Se elige al azar una de estas muestras, ¿cuál es la probabilidad de que

i. la muestra seleccionada sea de la variedad C?

ii. la muestra seleccionada sea de la variedad A y resulte de bajo contenido energético?

iii. sabiendo que resultó de alto contenido energético, no sea de la variedad B?

iv. sabiendo que no resultó de la variedad A, sea de contenido energético bajo?

v. sea de la variedad A o de contenido energético bajo?

(c) Un fabricante asegura que la variedad de jugo A es la que tiene mayor contenido energético entre las variedades de jugo analizadas. Los resultados obtenidos, ¿apoyan esta afirmación? ¿Qué probabilidades deben ser comparadas para analizar lo dicho por el fabricante?

(d) ¿Son los eventos "la muestra elegida es de variedad A" y "la muestra elegida es de bajo contenido energético" independientes?

6. En una urna con 5 bolitas rojas, 3 blancas y 4 verdes, se realizan 2 extracciones de bolitas y se observa el color de las bolitas extraídas.

(a) Extracciones con reposición: Si las 2 extracciones se hacen con reposición (es decir, se devuelve al bolillero la bolita extraída antes de hacer la siguiente extracción), calcular las siguientes probabilidades.

i. Hallar la probabilidad de que la primera bolita extraída sea roja.

ii. Hallar la probabilidad de que la segunda bolita extraída sea roja sabiendo que la primera bolita extraída es roja.

iii. Hallar la probabilidad de que las dos primeras bolitas extraídas sean rojas

iv. Hallar la probabilidad de que la segunda bolita extraída sea roja.

v. Hallar la probabilidad de que se extraiga una bolita blanca y una roja.

vi. Hallar la probabilidad de que las dos bolitas extraídas sean del mismo color.

vii Hallar la probabilidad de que alguna de las bolitas sea verde.

(b) Extracciones sin reposición: Si las 2 extracciones se hacen sin reposición (es decir, una vez extraída una bolita ésta no se devuelve al bolillero), repetir el cálculo de las probabilidades anteriores.

7. El pájaro biguá se alimenta de peces. Sobrevuela la laguna hasta que, cuando tiene un pez a la vista, desciende repentinamente a su caza con una probabilidad de éxito de 0.3. Queda satisfecho apenas caza el primer pez y desciende a lo sumo tres veces por día. Sea X el número de descenso en que el pájaro biguá caza un pez ó cero si no captura ningún pez.

(a) Hallar la función de probabilidad puntual p_X y graficarla.

(b) Hallar la función de distribución acumulada F_X y graficarla.

(c) ¿Cuál es la probabilidad de que en un día determinado el pájaro coma?

(d) El pájaro biguá puede sobrevivir 3 días sin comer. Sabiendo que comió el jueves ¿cuál es la probabilidad de que el pájaro muera de hambre el lunes?

8. Un apostador tiene la opción de jugar a dos juegos de azar con los dados. El juego A le paga \$2 si al arrojar un dado sale 1 ó 5 y diez centavos si sale cualquier otro valor. El juego B le paga \$2 si el resultado del dado es 1 ó 5, \$1 si dicho resultado es un 3 y no recibe premio si sale cualquier otro valor. Además, el costo por jugar a cada juego es de \$1. Hallar la esperanza y varianza de la ganancia neta para cada juego. ¿A cuál le conviene jugar?

9. Sea X una v.a. con la siguiente función de distribución:

$$F_X(x) = \begin{cases} 0 & x < 1 \\ 0.4 & 1 \leq x < 2 \\ 0.7 & 2 \leq x < 5 \\ 1 & x \geq 5 \end{cases}$$

   (a) Calcular $P(1.5 < X \leq 5)$
   (b) Calcular $P(1 < X < 5)$
   (c) Calcular $P(X \geq 2)$
   (d) Hallar p_X.

10. Se arroja un dado 8 veces. Calcular la probabilidad de que salgan

   (a) exactamente 3 unos,
   (b) entre 3 y 5 cuatros,
   (c) menos de 7 cincos.

11. El número de cierto tipo de larvas en un estanque tiene una distribución de Poisson de párametro $\lambda = 3$ por cm^3 de agua.

   (a) Calcule la probabilidad de que una muestra de 1 cm^3 contenga 4 o más larvas.
   (b) Si ahora se toman en forma independiente 5 muestras de 1 cm^3 de volumen cada una. Cuál es la probabilidad de que exactamente 3 de ellas contengas 4 o más larvas?

12. La administración de una universidad le asegura a un matemático que él tiene sólo una posibilidad en 10000 de encontrarse atrapado en un catastrófico ascensor en el edificio donde se encuentra el departamento de mátematicas. Si él va a trabajar 5 días a la semana, 52 semanas al año, durante 10 años y siempre toma el ascensor. ¿Cuál es la probabilidad de que nunca quede encerrado en el ascensor al subir? ¿Cuál es la probabilidad de que se quede encerrado una vez al subir? ¿Dos veces?. Suponer que los resultados de cada día son mutuamente independientes.

13. En un juego de tiro al blanco, la distancia al centro (en cm.) que obtiene Juan se considera una variable aleatoria X con la siguiente función de densidad

$$f_X(t) = \begin{cases} \frac{t}{72} & \text{si } 0 \leq t \leq 12 \\ \\ 0 & \text{en otro caso.} \end{cases}$$

   (a) Hallar la probabilidad de que un tiro de Juan diste menos de 1 cm. del blanco.
   (b) Hallar F_X.
   (c) Hallar $\mathbb{E}(X)$ y $Var(X)$.
   (d) Hallar el percentil o cuantil 0.90 de la distribución X.
   (e) En el pub se organiza un juego que otorga un premio de $\$120 - 10X$ para cada lanzamiento al blanco, donde X es la distancia conseguida. Si cada vez que se desea participar de este juego de debe pagar \$45, ¿cuál es la esperanza y varianza de la ganancia neta para Juan?
   (f) ¿Cuál es la probabilidad de que la gananacia neta sea mayor que la ganancia neta esperada?
   (g) Juan tira 12 veces al blanco, ¿cuál es la probabilidad de que dos o menos de sus tiros disten menos de 1 cm. del blanco?

14. Una barra de 12 pulgadas sujeta por ambos extremos, debe someterse a una creciente cantidad de esfuerzo hasta que se rompa. Sea Y = distancia desde el extremo izquierdo hasta dónde ocurre la rotura. Supongamos que la densidad de Y es la siguiente

$$f_Y(y) = \begin{cases} ay\left(1 - \frac{y}{12}\right) & \text{si } 0 \leq y \leq 12 \\ \\ 0 & \text{en otro caso.} \end{cases}$$

(a) Hallar a.

(b) Calcular $P(Y \leq 4), P(6 < Y); P(4 \leq Y < 6)$.

(c) Hallar la esperanza y la varianza de Y.

(d) Calcular la probabilidad de que el punto de ruptura ocurra a más de 2 pulgadas del punto de ruptura esperado.

15. El diámetro D (expresado en dm) del tronco de cierta especie de árboles es una variable aleatoria con función de densidad

$$f_D(x) = kxI_{(0;10)}(x)$$

(a) Hallar el valor de la constante k.

(b) ¿Cuál es la probabilidad de que el diámetro de un árbol de esa especie elegido al azar mida entre 4 y 6 dm?

(c) Se elige un árbol de esa especie al azar. Se sabe que tiene un diámetro de más de 5 dm. ¿Cuál es la probabilidad de que el diámetro de dicho árbol mida entre 4 y 6 dm? Comparar con la respuesta anterior.

(d) En un área del bosque hay 3 árboles de esa especie. Calcular la probabilidad de que exactamente 2 de ellos tengan el diámetro entre 4 y 6 dm.

(e) ¿Cuántos árboles habría que muestrear en el bosque para que la probabilidad de encontrar al menos uno cuyo diámetro mida entre 4 y 6 dm, sea mayor o igual que 0.99?

16. La cantidad de tiempo, en minutos, que una persona debe esperar el colectivo de una cierta linea los días de semana por la mañana es una variable aleatoria con distribución uniforme en el intervalo $[0, 15]$.

(a) ¿Cuál es la probabilidad de que espere entre 5 y 10 minutos?

(b) ¿Cuál es el tiempo promedio que debe esperar?

(c) Aproximadamente el 80% de las veces espera menos de ..... minutos. Completar y justificar.

(d) Una persona debe tomar el colectivo a las 8:30 para llegar a tiempo a su trabajo.

i. ¿A qué hora debería llegar a la parada para tener un 80% de probabilidades de llegar a tiempo?

ii. Supongamos que la persona llega a la parada todas las mañanas a la hora calculada en el item anterior.¿Cuál es la probabilidad de que no llegue tarde ningún día de una semana? ¿Cuál es la probabilidad de que llegue tarde exactamente dos días de una semana?

17. La biblioteca de una facultad dispone de una red de computadoras al alcance de los estudiantes. El tiempo medido en minutos que un estudiante destina a búsqueda bibliográfica semanalmente es una variable aleatoria T exponencial. Además, se sabe que el 49.9% de los estudiantes destina más de 20 minutos semanales a la búsqueda bibliográfica.

(a) Hallar la esperanza de la variable aleatoria T.

(b) Calcular la probabilidad de que un alumno destine más de 10 minutos a la búsqueda bibliográfica en una semana dada.

(c) Sabiendo que un alumno destinó esta semana más de 20 minutos a la búsqueda bibliográfica calcular la probabilidad de que destine más de 30. Comparar con la probabilidad calculada en el ítem anterior.

(d) Supongamos que de acuerdo al tiempo destinado a la búsqueda bibliográfica el usuario (siempre estudiante) es clasificado en una de tres categorías: I si $T < 25$, II si $25 \leq T \leq 50$ y III si $T > 50$. Hallar la esperanza de la variable aleatoria $W =$ categoría asignada al usuario.

18. La medida en centímetros de la longitud de la cintura de los hombres en Buenos Aires sigue una distribución normal con media 75 y varianza 25. Se sabe que todos los hombres de menos de 70 cm. de cintura usan cinturón de talle 1, mientras que los de cintura entre 70 y 81 cm. usan talle 2 y los restantes talle 3.

(a) ¿Qué proporción de hombres usa cintos de talle 2?

(b) ¿Cuál debería ser la longitud máxima de cintura del talle 1 si se quiere que el 30% de los hombres use talle 1?

(c) Carolina sabe que la cintura de su novio mide más de 70 cm. ¿Cuál es la probabilidad de que use talle 2?

(d) Si en la tienda entran azarosamente hombres a comprar de a un cinturón, ¿cuál es la probabilidad de que los primeros tres cinturones que se vendan sean del mismo talle?

19. (Interpretación de covarianza) En una cierta población, se elige un trabajador mayor de 30 años. Sean

$X =$ cantidad de años de educación que recibió

$Y =$ salario que cobra (en miles de pesos)

Se sabe que la función de probabilidad puntual del vector aleatorio (X, Y) está dado por $p_{XY}(x, y)$

Y/X	7	12	18	24
4	0.14	0.23	0.02	0.01
10	0.06	0.16	0.25	0.03
15	0	0.01	0.03	0.06

(es decir, $0.23 = p_{XY}(12, 4)$).

(a) Hallar $p_X, p_Y, E(X), E(Y)$.

(b) Para esta población, ¿las variables X e Y están positivamente asociadas?

(c) ¿Son las variables X e Y independientes?

(d) Suponga que las variables X e Y fueran independientes, con las funciones de probabilidad que calculó en el ítem a). Halle la probabilidad conjunta en este caso y compárela con la que figura más arriba.

20. Se quiere medir el número de consumidores de un producto A en una cierta población. Sea p la proporción real de individuos de la población que consumen el producto A. Para estimar p se eligen n personas al azar de la población y se les pregunta si consumen o no el producto A. Sea, para i entre 1 y n,

$$X_i = \begin{cases} 1 & \text{si la iésima persona encuestada dice consumir el producto A} \\ 0 & \text{en otro caso.} \end{cases}$$

Asumimos que las X_i son v.a.i.i.d.

(a) ¿Cómo estimaría p?

(b) Se desea estudiar cuánto difiere el estimador propuesto en a) del verdadero valor de p. Acotar la probabilidad de que el estimador difiera de p en más que 0.05 para $n = 50$, $n = 100$, $n = 500$ y $n = 1000$. ¿Que pasa cuando n tiende a infinito?

(c) ¿A cuántas personas habría que encuestar si desea que el estimador difiera de p en menos de 0.05 con probabilidad mayor a 0.99?

(d) Rehacer ambos items anteriores aproximando las probabilidades en lugar de acotarlas.

21. El gerente de un banco quiere estudiar el comportamiento de las extracciones del cajero automático de su sucursal para saber, entre otras cosas, cuánto dinero debe poner al final de cada día en el cajero para satisfacer las demandas de los clientes. Se sabe que "la cantidad de dinero requerida por un cliente para extracción en una operación del cajero automático de la sucursal de este gerente (en $)" es una variable aleatoria X con esperanza 1500 y varianza 22500. Se puede asumir que los distintos requerimientos realizados son variables aleatorias independientes. Suponga que una noche van 132 clientes a hacer extracciones:

(a) Hallar la esperanza y la varianza del total del dinero que requieren para extraer 132 clientes.

(b) Al finalizar la jornada bancaria el cajero tenía $200.000 ¿cuál es la probabilidad (aproximada) de que al día siguiente, cuando el gerente abre la sucursal, no haya dinero restante en el cajero automático?

(c) ¿Cuánto dinero debería depositar el gerente al finalizar una jornada bancaria si desea que con probabilidad por lo menos 0.99 haya dinero restante en el cajero a la mañana siguiente al abrir la sucursal?

(d) Si se ponen $200.000 en el cajero al finalizar la jornada bancaria, ¿cuál es la máxima cantidad de clientes que el cajero puede recibir de manera tal que la probabilidad de quedarse sin dinero sea menor a 0.1?

22. Se compararon tres dietas respecto al control de azúcar en la sangre en pacientes diabéticos. En el archivo `dietas.txt` se encuentran los valores de glucosa para las tres dietas consideradas (A, B, C), que contienen las lecturas de glucosa en la sangre de los pacientes. Es deseable que el paciente tenga valores entre 80 y 110 mg/dl. Para este ejercicio, chequee en la página de la materia los siguiente archivos: `dietas.txt`, `instrucciones estadistica descriptiva.pdf`, `ejerciciodescriptiva.R`.

(a) Cargue los datos al R.

(b) Para cada una de las tres dietas calcule medidas de centralidad: la media muestral, la mediana muestral, la media α-podada para $\alpha = 0.1, 0.2$. Para cada dieta compare los valores obtenidos de las cuatro medidas de posición, si observa una notable diferencia ¿a qué podría deberse?

(c) Calcule medidas de dispersión: el desvío estándar muestral, la distancia intercuartil y la MAD en cada una de las dietas. Compare los valores de dispersión obtenidos, si observa una notable diferencia ¿a qué podría deberse? ¿Cuál de las dietas parece ser la más estable?

(d) Obtenga los percentiles (muestrales o empíricos) 10, 25, 50, 75 y 90. Compare los valores de los percentiles obtenidos entre las distintas dietas.

(e) Construya histogramas que permitan visualizar los valores de glucosa para cada dieta. Compare la distribución de glucosa. ¿Alguna de ellas parece bimodal? ¿En alguna de ellas parece haber valores alejados? ¿Las dietas mantienen a los pacientes en los valores deseados? ¿La distribución de glucosa es asimétrica en alguno de los grupos? ¿En algún caso el ajuste normal parece razonable?

(f) Grafique 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, discuta simetría, presencia de outliers y compare dispersiones nuevamente.

(g) Grafique los qqplots correspondientes. ¿En algún caso el ajuste normal parece razonable?

(h) En base al análisis anterior, ¿cuál le parece la dieta más aconsejable?

23. En este ejercicio estudiaremos la distribución del promedio de variables independientes e idénticamente distribuidas y a través de los histogramas correspondientes analizaremos el comportamiento de estas distribuciones a medida que promediamos un número creciente de variables aleatorias. Es decir, trataremos de validar empíricamente los resultados de la Ley de los Grandes Números y el Teorema Central del Límite. Acompaña el archivo con instrucciones: `instrucciones_tcl_enR.R`

    Para ello generaremos una muestra de variables aleatorias con una distribución dada y luego calcularemos el promedio de cada muestra. Replicaremos ésto mil veces, es decir, generaremos una muestra aleatoria de la variable $\overline{X}$ de tamaño 1000. Observe que, en principio, desconocemos la distribución de $\overline{X}$. A partir de todas las replicaciones realizaremos un histograma para los promedios generados para obtener una aproximación de la densidad o la función de probabilidad de $\overline{X}$.

    (a) Comencemos por tomar un primer conjunto de datos de variables aleatorias $X_1, \ldots, X_{1000}$ independientes con distribución $U(0,1)$. Le pedimos al R que nos genere una muestra de ellas y luego hacemos un histograma. ¿A qué densidad se parece el histograma obtenido?

    (b) Considerar dos variables aleatorias X_1 y X_2 independientes con distribución $U(0,1)$ y el promedio de ambas, es decir,

    $$\overline{X} = \frac{X_1 + X_2}{2}\,.$$

    Generando una muestra de dos variables aleatorias con distribución $U(0,1)$ computar la variable promedio. Replicar 1000 veces y a partir de los valores replicados realizar un histograma. ¿Qué características tiene este histograma?

    (c) Aumentemos a cinco las variables promediadas. Considerar ahora 5 variables aleatorias uniformes independientes,es decir $X_1, X_2, \ldots, X_5$ i.i.d. con $X_i \sim U(0,1)$ y definir

    $$\overline{X} = \frac{1}{5}\sum_{i=1}^{5} X_i\,.$$

    Generando muestras de cinco variables aleatorias con distribución $U(0,1)$ computar la variable promedio. Repetir 1000 veces y realizar un histograma para los valores obtenidos. Comparar con el histograma anterior. ¿Qué se observa?

    (d) Aumentemos aún más la cantidad de variables promediadas. Generando muestras de 30 variables aleatorias con distribución $U(0,1)$ repetir el ítem anterior. ¿Qué se observa?

    (e) Ídem anterior generando muestras de 500 variables aleatorias. ¿Qué pasa si se aumenta el tamaño de la muestra? Observar que para poder comparar los histogramas de los distintos conjuntos de datos será necesario tenerlos dibujados en la misma escala tanto para el eje horizontal como para el vertical. Por eso, en general es más cómodo hacer boxplots para comparar distintos conjuntos de datos.

    (f) Finalmente hacerlo también para 1200, y hacer un boxplot de los 6 conjuntos de datos en el mismo gráfico. En este gráfico se verá que a medida que aumenta el n los valores de los promedios tienden a concentrarse, ¿alrededor de qué valor? Calcule media y varianza muestral para cada conjunto de datos. ¿Puede dar los valores teóricos a los que deberían parecerse? Realice un qqplot para cada uno de los 6 conjuntos de datos. ¿Son esperables los resultados?

    (g) El teorema central del límite nos dice que cuando hacemos la siguiente transformación con los promedios, $\frac{\overline{X}_n - E(X_1)}{\sqrt{\frac{Var(X_1)}{n}}}$, la distribución de estas variables aleatorias se aproxima a la de la normal estándar, cuando n es suficientemente grande. Para comprobarlo empíricamente, hagamos esta transformación en los 6 conjuntos de datos (es razonable hacerlo para valores de n suficientemente grandes, lo realizaremos en todos los casos para comparar) y luego comparemos los datos transformados mediante histogramas y boxplots.

(h) Repetir los ítems anteriores generando ahora variables con distribución $\mathcal{C}(0,1)$. Comparar los resultados obtenidos. Recordar que la densidad de una Cauchy es

$$f_X(x) = \frac{1}{\pi(1+x^2)},$$

que es una densidad simétrica alrededor del cero, con colas que acumulan más probabilidad que la normal estándar, y que no tiene esperanza ni varianza finitas.

24. Se hacen análisis de sangre a 25 personas elegidas al azar de la población obteniendo los siguientes índices de colesterol:

```
1.53 1.65 1.72 1.83 1.62 1.75 1.72 1.68 1.65 1.61
1.70 1.60 1.73 1.61 1.52 1.81 1.72 1.50 1.51 1.65
1.58 1.82 1.65 1.72 1.65
```

(a) Calcular estimadores para la esperanza y la varianza poblacional del índice de colesterol.

(b) Supongamos ahora que la distribución del índice de colesterol de la población es $N(\mu, \sigma^2)$

i. Si $\sigma^2 = 0.01$. Hallar un intervalo de confianza de nivel 0.95 para μ. Calcular la longitud del intervalo obtenido. ¿A cuántas personas debería realizarse el estudio si se quiere que la longitud sea menor que 0.05?

ii. Si σ^2 es desconocido, hallar un intervalo de confianza para μ de nivel 0.95

iii. Si σ^2 es desconocido, hallar un intervalo de confianza para σ^2 de nivel 0.95

25. Se quiere aumentar la producción anual de almendras en una zona de la provincia de Mendoza. Se observó el rendimiento de 140 parcelas elegidas al azar, plantadas con almendros, provistas de un nuevo sistema de riego obteniéndose una media muestral de 525 kilogramos por hectárea y un desvío muestral de 18 kilogramos por hectárea.

(a) En base a esta muestra hallar un intervalo de confianza de nivel asintótico 0.95 para el rendimiento medio de los campos con el nuevo sistema de riego.

(b) El rendimiento medio de los campos con el sistema de riego tradicional es de 512 kilogramos por hectárea. ¿Hay razones para afirmar que el nuevo sistema de riego produce un rendimiento mayor?

26. Una muestra de 1000 votantes es encuestada respecto a cierta propuesta política. Como resultado se obtiene que 200 están de acuerdo con la propuesta, 600 se oponen y 200 están indecisos.

(a) Hallar un intervalo de confianza de nivel asintótico 90% para la proporción poblacional de votantes que se oponen a la propuesta. ¿Es exacto o asintótico? ¿Por qué? Justifique su respuesta.

(b) ¿Cuántos votantes deberían encuestarse para que la longitud del intervalo obtenido sea menor o igual a 0.02?

27. El peso medio de calcio en un cemento estándar es de $94g/kg$. Se tomaron 16 muestras de cemento contaminado con plomo obteniéndose, en las 16 determinaciones de calcio, un peso promedio de $87g/kg$. Suponiendo que las mediciones de calcio siguen una distribución normal con desvío estándar $\sigma = 13g/kg$. Se quiere saber si la presencia de plomo afecta el peso medio de calcio en el cemento.

(a) ¿Qué hipótesis se deben testear para responder a esta pregunta?

(b) Si se tuviera información adicional respecto del efecto que la presencia del plomo en el cemento tiene sobre el contenido de calcio, en el sentido que este último sólo puede disminuir si hay plomo en el cemento, ¿cuál sería el test correcto para proponer para responder a la pregunta del enunciado, a nivel 5%? ¿Cuánto vale el p-valor en este caso?

(c) Repita lo realizado en (b) a nivel 1% y 10%.

(d) Hallar la potencia del test hallado en (b) si el verdadero peso medio del calcio del cemento contaminado es de $90g/kg$. Hallar la función de potencia del test.

(e) Si se quiere que la potencia del test sea de 0.90 cuando el peso medio del calcio del cemento contaminado es de $90g/kg$, hallar el tamaño de muestra necesario para lograrlo.

(f) Responder a la pregunta del ítem (a) con un test de hipótesis de nivel del 1%, definiendo claramente las variables aleatorias, los parámetros de interés y las hipótesis en cuestión. Escriba su conclusión en los términos del problema. Repita lo realizado a nivel 5%.

(g) Calcular el p-valor para el test del ítem anterior.

(h) Hallar el intervalo de confianza para el peso medio del calcio del cemento contaminado de nivel 99%. ¿Qué relación guarda con lo realizado en (f)? Si quisiéramos testear si el peso medio del calcio del cemento contaminado es de $92g/kg$ a nivel 1%, ¿podríamos sacar una conclusión sin hacer ninguna cuenta más? ¿Qué relación tiene este intervalo con lo realizado en (c)?

28. Una asociación de consumidores, preocupada por la cantidad de grasas contenida en una marca de hamburguesas, envía a un laboratorio independiente una muestra aleatoria de 12 hamburguesas para su análisis. El porcentaje de grasa en cada una de las hamburguesas de la muestra es:

21 18 19 16 18 24 22 19 24 14 18 15

El fabricante afirma que el contenido medio de grasa de este tipo de hamburguesas es menor al 18%. Basándose en la salida de R que figura más abajo, resuelva los siguientes items.

(a) La asociación de consumidores quiere saber si tiene motivos para decir que la afirmación del fabricante es falsa. Asumiendo que el contenido de grasa de cada hamburguesa de esta marca es una v.a con distribución normal y varianza conocida $\sigma^2 = 9$, proponer un test para resolver este problema. Escribir las hipótesis a testear, definir las variables aleatorias y los parámetros involucrados en el test, escribir el estadístico del test y su distribución bajo la hipótesis nula y dar la región de rechazo. ¿Qué le informaría al representante de la asociación de consumidores como conclusión del test?

(b) Dar un intervalo de confianza de nivel 0.95 para el verdadero contenido medio de grasa de este tipo de hamburguesas.

(c) Calcular el p-valor para el test del ítem (a). ¿Rechazaría la hipótesis nula a nivel 0.08?

(d) Si el verdadero contenido de grasa de las hamburguesas fuera del 20% ¿cuál sería la potencia de este test? Si se quiere que esta potencia sea de al menos 0.85, ¿cuántas hamburguesas habría que tomar en la muestra?

```
hamburguesas<-scan()
21 18 19 16 18 24 22 19 24 14 18 15
c(mean(hamburguesas),var(hamburguesas),sd(hamburguesas))
[1] 19.000000 10.545455  3.247377
```

29. Una empresa que vende gaseosas dice que la concentración media de azúcar en la gaseosa que vende es de 100 g/litro. Nosotros, que siempre dudamos de lo que nos dicen, queremos hacer un test de hipótesis para ver si hay evidencia suficiente para contradecir a la empresa.

(a) Se toman 14 muestras independientes de la gaseosa que venden y se les calcula la concentración de azúcar. Los datos obtenidos (en g/litro) son los siguientes:

102 102 93 95 102 103 99 97 108 94 102 102 99 98

(a) ¿Es razonable suponer que estos datos provienen de una distribucíón normal?

(b) Teniendo en cuenta el ítem anterior y suponiendo que la varianza poblacional es $\sigma^2 = 16$, realizar un test para ver si hay evidencias de que la concentración media sea distinta a 100 g/litro. Tomar una decisión a nivel $\alpha = 0.05$.

(c) Rehacer el ítem anterior suponiendo que ahora no sabemos cuál es la varianza poblacional.

(d) Rehacer el ítem anterior, si ahora sospechamos que la concentración media es menor a 100 g/litro.

30. Otro laboratorio tiene las mismas sospechas contra la empresa. Los datos que obtuvieron son los siguientes:

98 97 102 103 103 98 96 102 104 95 102 101 96 97

(a) ¿Es razonable suponer que estos datos provienen de una distribución normal?

(b) Realizar un histograma. Es razonable pensar que los datos provienen de una distribución simétrica?

(c) Teniendo en cuenta el ítem anterior, realizar un test para ver si hay evidencias a nivel $\alpha = 0.1$ de que la concentración media sea distinta a 100 g/litro.

(d) Realizar otro test que sea aplicable para este caso. Comparar con el ítem anterior.

31. Supongamos que queremos comparar las concentraciones de las gaseosas A y B. Se toman 14 muestras de cada una y se obtienen los siguientes resultados:

Gaseosa A: 94 90 100 104 100 100 96 98 100 105 99 108 101 99

Gaseosa B: 105 109 93 113 113 106 108 117 113 102 95 108 117 105

(a) Verificar que es razonable suponer que estos datos provienen de distribuciones normales.

(b) ¿Es razonable suponer que las varianzas de las dos poblaciones son iguales?

(c) Realizar un test para ver si hay evidencias de que la concentración media de la gaseosa A es menor a la de la gaseosa B. Tomar una decisión a nivel $\alpha = 0.1$.

(d) Realizar el test de Mann - Whitney para este caso. ¿Qué hipótesis se están testeando? Basándonos en este último test, ¿tiene sentido concluir que las medianas de los dos grupos son distintas?

(e) Hacer el mismo test en el caso hipotético en que las muestras fueran apareadas.

32. Nos cansamos de analizar gaseosas y ahora empezamos a jugar con un dado. Supongamos que lo tiramos 200 veces y que 60 veces sale el número 5. Dedicir si hay evidencia a nivel $\alpha = 0.001$ de que el dado esté cargado (es decir, de que la probabilidad de que salga el número 5 no sea 1/6).

33. Se mide el grado de impurezas de un producto químico. El método de medición está afectado por un error que se supone $N(0, \sigma^2)$, σ^2 desconocido. Además los errores correspondientes a diferentes mediciones son independientes entre sí. Se hicieron 12 observaciones obteniendo que el promedio es 0.85 con un desvío estándar muestral de 0.05. A partir de estos datos, ¿hay evidencia significativa para decir que el grado de impurezas del producto es distinto de 0.7 a nivel 0.05?

34. Se supone que 1 de cada 10 fumadores prefiere la marca A. Después de una campaña publicitaria en cierta región de ventas, se entrevistó a 200 fumadores para determinar la efectividad de la campaña. El resultado de esta encuesta mostró que 26 personas preferían la marca A.

(a) ¿Indican estos datos, a nivel aproximado 0.05, un aumento en la preferencia por la marca A?

(b) Calcular el valor p (o p-valor).

(c) ¿Cuál es la probabilidad aproximada de decidir que la campañaa publicitaria no fue efectiva, cuando en realidad la proporción de preferencia por la marca A después de la campaña es 0.15?

(d) ¿Qué tamaño de muestra debería tomarse para que la probabilidad de c) fuese a lo sumo 0.05?

35. El jefe de un Centro de gestión y participación (CGP) afirma que al menos el 92% de los documentos allí tramitados se entregan en un plazo estipulado. El defensor del pueblo, dudando de ese porcentaje, decide realizar un test de hipótesis en base a una muestra aleatoria de 130 documentos tramitados en ese CGP.

(a) Plantear el test de hipótesis, si el defensor del pueblo quiere tener una probabilidad aproximada de 0.05 de contradecir a jefe del CGP cuando su afirmación es correcta. Dar las hipótesis nula y alternativa, indicar el estadístico utilizado, su distribución aproximada bajo H_0 y dar la región de rechazo para el nivel propuesto.

(b) De los 130 documentos analizados 112 se entregaron a tiempo. ¿Cuál es su conclusión al nivel propuesto? ¿Es cierta la afirmación del jefe del CGP?

(c) Si en realidad se entregan a tiempo el 89% de los documentos ¿qué probabilidad tiene el defensor del pueblo de descubrir que la afirmación del jefe del CGP es falsa?

(d) ¿Cuántos de los 130 documentos deben ser entregados a tiempo para rechazar la hipótesis nula a nivel 0.0246?

36. Se mide el consumo diario de energía (MJ/día) en dos grupos de mujeres elegidas al azar: delgadas y obesas. Los resultados son

Delgadas	6.02	7.4	7.88	8.39	8.7	8.76	9.09	9.27	9.3	9.8	9.84	10.03	10.27
Obesas	8.42	9.16	9.69	10.21	10.4	10.48	10.93	11.14	11.14	11.81			

Se desea saber si las medias de consumo de energía de las poblaciones de las que provienen ambos conjuntos de datos coinciden o no. Responder mediante un test de hipótesis apropiado de nivel 0.05. Asuma que los consumos diarios de cada mujer tienen distribución normal, y que las varianzas poblacionales de ambos grupos de mediciones son iguales.

```
> delgadas<-scan()
1:  6.02 7.4 7.88    8.39    8.7    8.76    9.09    9.27    9.3    9.8    9.84    10.03

14:
Read 13 items
> mean(delgadas)
[1] 8.826923
> var(delgadas)
[1] 1.417956
> obe<- scan()
1:  8.42 9.16    9.69    10.21    10.4    10.48    10.93    11.14    11.14    11.81
11:
Read 10 items
> mean(obe)
[1] 10.338
> var(obe)
[1] 1.036707
```

37. Este ejercicio se ocupa de simular datos cuya distribución es conocida y luego evaluar con ellos el funcionamiento de los intervalos de confianza (vea el archivo de instrucciones de R: `clasedetestsenR.R` linkeado en la página para este ejercicio y los dos siguientes).

    (a) Simule 3 datos con distribución $N\left(\mu = 40, \sigma^2 = 4\right)$.

    (b) Simule otros 3 datos con distribución $N\left(40, 4\right)$. Observe que difieren de los anteriores. Guárdelos en un vector. Estime a μ y a σ con estos datos. Construya un intervalo de confianza para μ de nivel 0.95 basado en estos datos con R, sin usar el valor conocido de la varianza σ^2. Para hacerlo, construya el intervalo con los percentiles adecuados empleando la fórmula vista en clase. ¿Los percentiles de qué distribución debe usar? Luego use la instrucción automática para hacerlo en R. ¿El verdadero valor de μ (que es 40) pertenece a dicho intervalo?

    (c) Repita 100 veces lo siguiente:

        i. Genere 3 datos con distribución $N\left(\mu = 40, \sigma^2 = 4\right)$.

        ii. Con los tres datos recién generados, construya un intervalo de confianza para μ de nivel 0.90.

        iii. Observe si el verdadero valor de μ (que es 40) pertenece a dicho intervalo. Guarde esta información, anotando un 0 si no pertence y un 1 si sí lo hace.

        Observe que el output de este ejercicio debe ser un vector de longitud 100 de ceros y unos. ¿Cuántos unos espera tener en su vector? ¿Cuántos realmente tiene?

    (d) Repita el ítem anterior pero utilizando los percentiles de la distribución normal, en vez de los percentiles de la distribución t. ¿Qué proporción de los intervalos construidos en este ítem cubre al verdadero valor de μ?

38. Queremos comparar los datos de la ingesta media diaria (durante 10 días) de un grupo de individuos (kJ) que siguen una determinada dieta alimentaria, con la ingesta recomendada de 7752 kJ. Es decir, cada dato corresponde al promedio de ingesta de energía durante 10 días para un individuo en particular.

    5260 5470 5640 6180 6390 6515 6805 7515 7515 8230 8770

    (a) Se busca responder con un test de hipótesis de nivel del 5%, definiendo claramente las variables aleatorias, los parámetros de interés y las hipótesis en cuestión, así como el estadístico utilizado y su distribución bajo la hipótesis nula, y los supuestos realizados y verificados. Escriba su conclusión en los términos del problema. Los datos se presentan ordenados por comodidad. ($\overline{x} = 6753.6 \quad s = 1142.1$)

    (b) Construya intervalos de confianza de nivel 0.90, 0.95 y 0.99 para la ingesta media diaria (en kJ) para los que siguen esta dieta específica. ¿Cuál es más largo?

39. Hagamos un test sobre datos generados. Genere 50 datos con distribución Exponencial, con $\lambda = 1/4$. Luego aplique un test apropiado para chequear si la esperanza de la población de la que proviene la muestra es igual a 4.

40. Un diseñador de productos está interesado en reducir el tiempo de secado de una pintura. Se prueban dos fórmulas de pintura; la fórmula 1 tiene el contenido químico estándar y la fórmula 2 tiene un nuevo ingrediente secante que tiende a reducir el tiempo de secado. Se sabe que el tiempo de secado es una variable aleatoria con distribución normal. Se pintan 16 placas con la fórmula 1 y otras 16 con la fórmula 2. Los dos tiempos promedio de secado obtenidos fueron 118 y 112 minutos respectivamente, mientras que los desvíos estándar fueron de 10.5 y 7.

    (a) Defina cuáles son las hipótesis a testear y qué test se debe utilizar.

    (b) Realice el test de hipótesis correspondiente a nivel 0.05 e indique claramente la conclusión.

(c) Calcule el p-valor. ¿Qué decisión se hubiera tomado a nivel 0.01?

41. Se consideran dos fórmulas químicas A y B para un nuevo producto que se utilizará para teñir telas. La empresa productora está interesada en telas especialmente resistentes a perder color tras la exposición al sol. Diez piezas de diferentes tejidos se cortan en dos mitades y a cada una se le aplica uno de los dos tintes. Los 20 trozos de tela se exponen al sol durante un periodo de tiempo, al cabo del cual se mide la intensidad del color, obteniéndose lo siguiente (bajos resultados indican menos intensidad, es decir, mayor pérdida de color):

	Telas									
Fórmulas	1	2	3	4	5	6	7	8	9	10
A	7.2	4.3	5.8	6.5	4.9	6.8	6.3	7.0	6.5	6.2
B	5.1	4.1	5.5	4.1	5.0	5.1	5.3	7.3	4.8	5.8

$\overline{x}_A$	$\overline{x}_B$	$\overline{x}_{A-B}$	s_A	s_B	s_{A-B}
6.15	5.21	0.94	0.92	0.91	0.97

Se supone que la diferencia entre la intensidad de color con la fórmula A y la intensidad de color con la fórmula B sigue una distribución $N\left(\mu_A - \mu_B, \sigma^2\right)$. Se desea saber si existen diferencias entre las medias de ambas fórmulas.

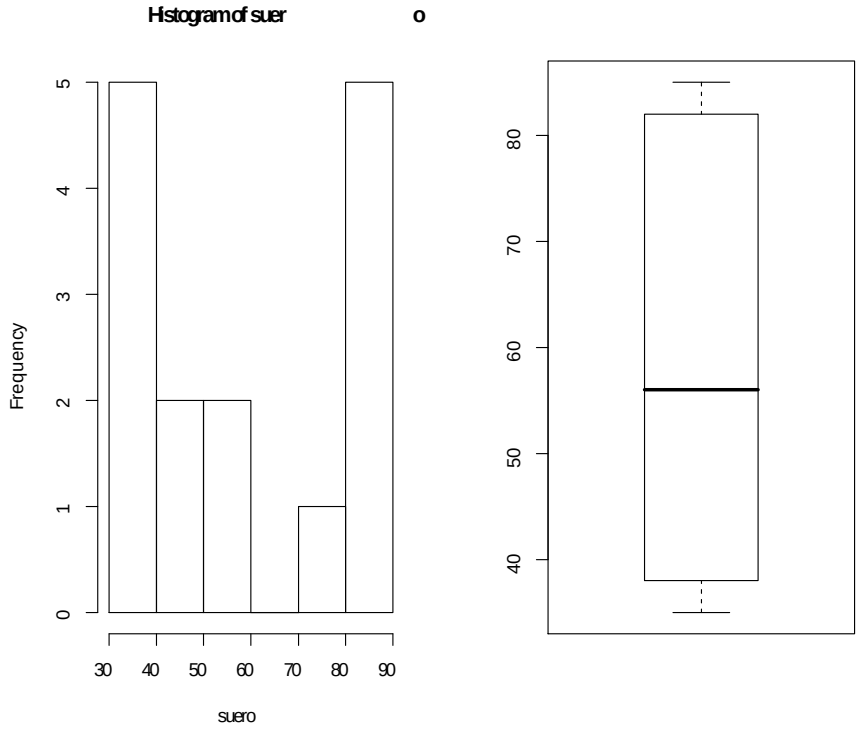
(a) Plantee las hipótesis correspondientes y proponga un test de nivel 0.05 para las hipótesis planteadas.

(b) Construir un intervalo de confianza de nivel 0.95 para la diferencia de intensidad media entre las fórmulas. ¿Tiene razones la empresa productora para sospechar que la formula A es mejor que la fórmula B?

(c) Suponiendo que la varianza poblacional es $\sigma^2 = 1$, calcular la probabilidad de cometer error de tipo II cuando la diferencia de intensidad es de 0.97.

42. Los datos siguientes corresponden al contenido de cierto compuesto en el suero de pacientes que padecen una enfermedad. Interesa testear la hipótesis de que la mediana de la población de la cual provienen los pacientes es $\widetilde{\mu} = 40$. Los datos se presentan ordenados.

Suero	35	36	37	37	39	44	48	56	60	76	81	83	83	84	85
$D_i = X_i - m_0$															
$\lvert D_i \rvert$															
Rango															



(a) Realice el histograma de estos datos. ¿Le parece que los datos tienen una distribución normal?

(b) Aplique el test de rangos signados de Wilcoxon para la mediana de una población a estos datos y dé la conclusión a nivel 0.05. ¿Son válidos los supuestos para realizar el test? Hágalo con el R y también a mano, completando la tabla anterior.

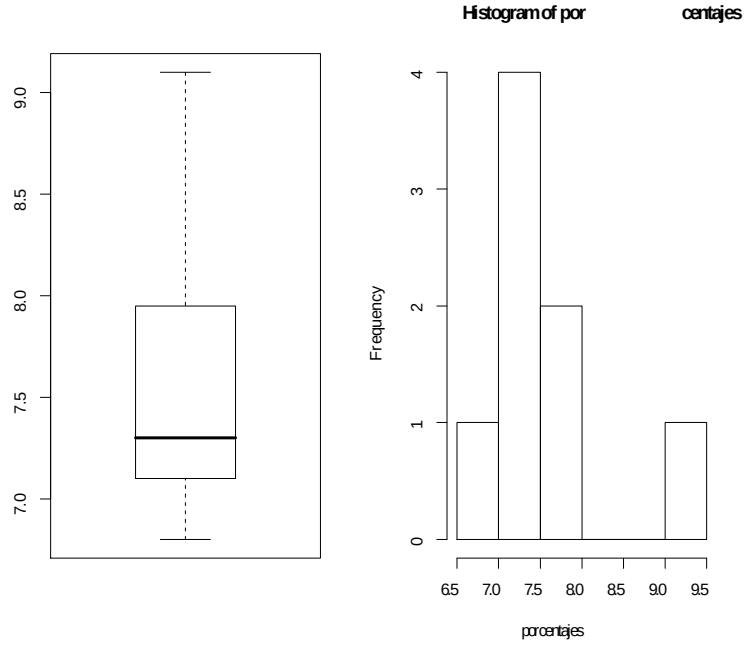
```
> wilcox.test(suero,mu=40)

        Wilcoxon signed rank test with continuity correction

data:  suero

V = 103.5, p-value = 0.01454

alternative hypothesis: true location is not equal to 40
```

43. Un farmacéutico prepara un medicamento y afirma que contiene un 8 por ciento de un componente en particular. De su preparación obtiene una gran cantidad de frascos. Se analizaron 8 de ellos elegidos al azar y se halló que contenían 7.3, 7.1, 7.9, 9.1, 8.0, 7.1, 6.8 y 7.3 por ciento del componente. Se quiere saber si estos resultados contradicen la afirmación del farmacéutico.

(a) Definir las variables aleatorias y los parámetros involucrados en el problema. Escribir las hipótesis a testear.

(b) ¿Cuál es el test apropiado para este problema? Escribir el estadístico del test y su distribución bajo

la hipótesis nula. Hallar el p-valor. Escribir la conclusión del test.



44. Una muestra de desechos fotográficos fue analizada midiéndose el contenido de plata por espectrometría de absorción atómica. Se hicieron 5 mediciones obteniendose 9.8, 10.2, 10.7, 9.5 y 10.5 g/ml. Después de un tratamiento químico, la muestra fue analizada de nuevo obteniéndose los valores 7.7, 9.7, 8.1, 9.9 and 9.0 g/ml. ¿Hay evidencias de que el tratamiento produjo un cambio en los niveles de plata? Asumir que ambas muestras provienen de poblaciones con la misma distribucion y que si hay una diferencia entre ellas se debe sólo a la posición de la distribución. Calcular el valor del estadístico del test para estas muestras y compararlo con el obtenido con el R.

```
> wilcox.test(muestra1,muestra2)

        Wilcoxon rank sum test

data:  muestra1 and muestra2

W = 22, p-value = 0.05556

alternative hypothesis: true location shift is not equal to 0
```

45. Se midió el tiempo (en segundos) que demora la concentración de un compuesto en reducirse a la mitad durante una reacción. Se realizaron 28 repeticiones de la reacción en condiciones independientes e idénticas. Los datos se exhiben ordenados. Ver el archivo `conc.txt`, donde están los datos, y el archivo `instrucciones popurri de Tests.txt`, donde están las instrucciones.

617.2	617.2	617.3	617.4
617.4	617.5	617.6	617.6
617.6	617.7	617.7	617.7
617.7	617.8	617.8	617.8
618.0	618.0	618.2	618.5
619.9	621.9	623.7	626.7
628.1	632.6	648.0	652.7

(a) Graficar el boxplot y el qqplot (o gráfico de probabilidad normal) de los datos.

(b) Nos interesa testear las hipótesis $H_0 : \widetilde{\mu} = 620$ versus la alternativa $H_0 : \widetilde{\mu} < 620$ donde $\widetilde{\mu}$ es la mediana de la distribución del tiempo (en segundos) que demora la concentración de un compuesto en reducirse a la mitad durante una reacción. Hacerlo usando el test apropiado. ¿Se puede usar el test de t?

46. En un culto suplemento de cultura se afirma que el precio mediano de los libros vendidos en Ciudad Gótica es de 120 pesos. Para validar dicha afirmación, se eligen al azar 20 facturas correspondientes a las ventas de un libro en distintas librerías de la ciudad elegidas al azar. Los precios de dichos libros aparecen listados más abajo. ¿A nivel 0.05 permiten sostener la afirmación hecha en el suplemento? ¿Qué tests podría aplicar? ¿Qué test conviene aplicar?

21.92	25.75	28.41	33.00	45.00	49.79	51.47
53.55	62.91	63.90	75.01	96.99	102.52	124.33
129.44	135.76	143.42	144.93	156.99	159.99	

47. Se desean comparar dos métodos para determinar el verdadero nivel medio de paracetamol presente en las tabletas. El objetivo del estudio es determinar si ambos métodos de medición miden lo mismo. Se decide medir el nivel de concentración de paracetamol en 10 tabletas con cada método.

(a) A continuación se describen dos propuestas para llevar a cabo el experimento:

i. Seleccionar 20 tabletas, y al azar elegir 10 de ellas para medir con el método que determina la concentración midiendo rayos UV (lo llamaremos procedimiento UV) y a otras 10 medirles la concentración de paracetamol midiendo rayos infrarrojo (lo llamaremos procedimiento infrarrojo).

ii. Seleccionar 10 tabletas, partirlas al medio y sortear cada una de las dos mitades para determinar en una de ellas la concentración con el procedimiento UV y en la otra con el procedimiento infrarrojo.

¿Qué procedimiento convendría usar? ¿Por qué?

(b) En el archivo `paracetamol.txt` figuran los datos, que están copiados en la tabla siguiente, y que fueran obtenidos con el mejor método de los dos anteriormente descriptos.

UV	Infrarrojo
84.63	83.15
84.38	83.72
84.08	83.84
84.41	84.20
83.82	83.92
83.55	84.16
83.92	84.02
83.69	83.60
84.06	84.13
84.03	84.24

¿Sugiere esta información una diferencia signicativa en los valores obtenidos con los dos métodos de medición? Plantee las hipótesis correspondientes y proponga un test de nivel 0.05 para las hipótesis planteadas. Valide las hipótesis necesarias para aplicarlo.

(c) Construir un intervalo de confianza de nivel 0.95 para la diferencia de concentración media de paracetamol entre los dos procedimientos.

48. Se mide el consumo diario de energía (MJ/día) en dos grupos de mujeres elegidas al azar: delgadas y obesas, presentados en el ejercicio 36. Los datos se encuentran el archivo `obesas y delgadas.txt`. En ese ejercicio se deseaba saber si las medias de consumo de energía de las poblaciones de las que provienen ambos conjuntos de datos coinciden o no. Se respondió mediante un test de hipótesis, para el que se hicieron ciertos supuestos. Validar los supuestos realizados.

49. Una compañía pretende decidir si basado solamente en el aspecto del envase será posible modificar el precio de un perfume. Para ello selecciona 15 clientes al azar les hace "probar" el perfume presentado en un envase de aspecto tradicional y les solicita que indiquen cual es el máximo valor que estarían dispuestos a pagar por el mismo. Selecciona otros 15 clientes al azar y repite la prueba pero usando el envase moderno. El precio máximo reportado por cada uno de los 30 clientes se muestra en la tabla siguiente. Ver el archivo `perfumes.txt`

Envase	Máximo precio que pagaría														
Moderno	20	25	40	44	43	13	32	34	35	11	12	46	13	17	47
Tradicional	5	7	10	11	12	17	21	28	33	35	40	40	41	44	45

(a) Realice gráficos descriptivos de ambos conjuntos de datos: histogramas, boxplot, qqplots.

(b) Decida qué test realizar para definir si el precio máximo que están dispuestos a pagar los clientes depende del envase. Indique en base a qué gráficos hace su elección.

(c) Realice el test de hipótesis correspondiente a nivel 0.05 e indique claramente la conclusión.

50. Se quiere estudiar si existen diferencias en la resistencia de dos tipos de materiales utilizados para la fabricación de calzado, a los que denominaremos A y B. Para ello, se arman pares de zapatos en los cuales uno de los zapatos –elegido al azar- se realiza enteramente con el material A y el otro con el material B. Se eligieron 12 chicos al azar, los que usaron durante 2 meses el calzado y al final de este período se midió el estado de deterioro de los zapatos. La tabla muestra los valores de una medida de deterioro (mayor deterioro implica menor resistencia). Están en el archivo `zapatos.txt`.

Chico	A	B
1	8.14	18.05
2	22.84	21.79
3	6.17	10.16
4	11.88	23.23
5	22.93	33.39
6	14.79	13.35
7	42.84	37.09
8	31.72	42.05
9	7.42	12.50
10	9.52	16.31
11	5.32	15.43
12	3.92	18.52

(a) Realice boxplots y qqplots de los datos de A, B y de la diferencia ($DIF = A - B$), los boxplots realícelos en la misma escala.

(b) Decida qué test realizar para decidir si la resistencia del calzado depende del material. Indique en base a qué gráficos hace su elección.

(c) Defina claramente las variables aleatorias involucradas y los parámetros en cuestión. Realice el test de hipótesis correspondiente a nivel 0.01, escriba las hipótesis e indique claramente la conclusión.

51. La respiración del suelo es una medida de la actividad microbiana en el suelo, que afecta el crecimiento de las plantas. En un estudio se extrajeron muestras de tierra de dos ubicaciones geográficas en un bosque: en un claro del follaje y en una ubicación cercana bajo una densa arboleda. Se midió la cantidad de dióxido de carbono emitida por cada muestra de suelo (mol CO_2/g suelo/hora). Los resultados figuran a continuación

claro	22	29	13	16	15	18	14	6
denso	17	20	170	315	22	190	64	

El objetivo del investigador es comparar los niveles de respiración de ambos suelos, y concluir a nivel 0.05. Realice el test apropiado para dar una conclusión. Justifique la utilización del test.

52. Se contaron los errores cometidos por 10 individuos en la traducción de dos párrafos de igual longitud del inglés al francés. El párrafo A está escrito por un autor inglés, el párrafo B por un autor estadounidense. Se quiere evaluar si la dificultad involucrada en la traducción de ambos párrafos es comparable. Aplique el test conveniente para decidir si hay diferencia significativa en la dificultad entrañada por ambos tipos de texto. Justifique la elección del test.

Sujeto	Errores en el párrafo A	Errores en el párrafo B
1	8	10
2	7	6
3	4	4
4	2	5
5	4	7
6	10	11
7	17	15
8	3	6
9	2	3
10	11	14

53. Un investigador registró las transpiración de plantas de tabaco de 5 variedades distintas. Para cada una de las variedades registró la transpiración de 12 plantas elegidas al azar. Se decide aplicar un modelo de análisis de la varianza (ANOVA) para analizar si existen diferencias significativas entre las variedades. Los datos observados son

Tabaco1	Tabaco2	Tabaco3	Tabaco4	Tabaco5
9.080729	8.798094	8.619636	9.003297	9.948916
8.898094	8.980729	9.403858	8.919114	9.509691
8.70729	8.577438	8.819636	8.785475	9.462271
8.577438	9.027343	9.103858	9.339848	9.56095
9.127343	8.607059	9.02778	9.226791	9.116161
8.607059	9.234845	9.443001	8.559244	9.727411
8.934845	8.459148	8.919614	8.425281	9.757861
8.759148	8.668539	9.263839	9.208764	9.087074
8.768539	8.828109	9.092951	8.680994	9.392484
8.428109	8.993441	8.869939	8.779978	9.53709
9.23441	8.798094	9.17763	8.978667	9.284733
8.98094	8.980729	8.817533	9.003297	9.432191

En el archivo `salida ejercicio anova.pdf` figuran las salidas de correr los procedimientos de ANOVA del R a estos datos. Los datos figuran en el archivo `tabac.txt`.

(a) Plantee un modelo para este problema definiendo claramente las variables aleatorias y los parámetros involucrados en este problema. Indique cuáles son los supuestos necesarios para aplicar el ANOVA. Escriba las hipótesis que se desea testear.

(b) Analizar la validez de los supuestos.

(c) Testee las hipótesis planteadas en a) con nivel de significación 0.05. ¿Qué test/salida emplea para tomar esta decisión? Plantee el p-valor. Escriba la conclusión del test.

(d) ¿Qué pares de variedades difieren significativamente entre sí con un nivel de significación simultáneo del 5%? Explique en qué método/s y en qué salida/s basa su conclusión.

(e) Usando algún método para calcular todos los intervalos de confianza de nivel simultáneo 0.95 para la diferencia de medias entre las variedades, construya un intervalo para la diferencia entre la variedad 1 y la 5.

54. La Compañía Toluca produce equipos de refrigeración así como diversos repuestos para estos. En el pasado, una de estas piezas de repuesto se ha producido periódicamente en lotes de diversos tamaños. Se desea determinar el tamaño óptimo del lote para producir este repuesto. La producción de este repuesto involucra modificar el proceso de producción (lo cual debe hacerse no importa el tamaño de lote que se desee producir) y también involucra operaciones de ensamblaje y movimiento de maquinaria. Una cuestión importante a la hora de determinar el tamaño de lote óptimo consiste en estudiar la relación entre el tamaño de lote y la cantidad de horas trabajadas requeridas para producir dicho lote. Para determinar tal relación se utilizaron los datos de 25 corridas recientes de producción. Sean:

X_i : tamaño del lote de la corrida iésima

Y_i : cantidad de horas trabajadas en la producción de la corrida iésima

Las condiciones de producción fueron estables durante el periodo de seis meses en el que las 25 corridas se realizaron, y se esperaba que continuaran así durante los siguientes 3 años. A continuación los datos, que también se encuentran en el archivo `toluca.txt`. En el archivo `salida ejercicio regresion.pdf` figuran las salidas de correr los procedimientos de ANOVA del R a estos datos.

Tamaño	Horas	Tamaño	Horas	Tamaño	Horas
20	113	60	224	90	377
30	121	70	252	90	376
30	212	70	361	100	353
30	273	70	323	100	420
40	160	80	399	110	435
40	244	80	342	110	421
50	221	80	352	120	546
50	157	90	389		
50	268	90	468		

(a) Hacer un scatter plot (gráfico de Y versus X) para ver si tiene sentido ajustar una regresión lineal a los datos.

(b) Ajuste un modelo lineal a los datos, con `tamanio` como variable independiente y `horas` como variable dependiente. ¿Cuál es la recta estimada?

(c) Decida si la relación lineal es estadísticamente significativa.¿Qué supuestos debe realizar para sacar esta conclusión? Verifique que los datos los soporten. Indique qué gráficos y tests empleó para concluir.

(d) Prediga la cantidad de horas que involucra la producción de un lote de tamaño 90, con el modelo lineal. Lo mismo para un lote de tamaño 50. Calcule el residuo correspondiente a la séptima, a la octava observación, y a las observaciones 18 y 19.

(e) Para los tamaños de lote 20, 50 100 y 120 dé los intervalos de confianza y de predicción de nivel 0.95 para la cantidad de horas trabajadas. ¿Cuál es el más largo? ¿Sería correcto predecir la cantidad de horas necesarias para producir un lote de tamaño 150 con el modelo ajustado?

55. Se quiere calibrar una termocupla. Se conoce la temperatura verdadera x, que supondremos sin error. Los correspondientes valores de y se observan en la termocupla. Las observaciones se muestran en la tabla de abajo. Supongamos que se hace una nueva observación $Y_0 = 200°C$. Estimar la verdadera temperatura x_0 y dar un intervalo de confianza.

	tempV	tempObs
1	100	88.80
2	120	108.70
3	140	129.80
4	160	146.20
5	180	161.60
6	200	179.90
7	220	202.40
8	240	224.50
9	260	245.10
10	280	257.70
11	300	277.00
12	320	298.10
13	340	318.80
14	360	334.60
15	380	355.20
16	400	377.00

56. Un investigador médico estudia un método nuevo, rápido para medir bajas concentraciones de galactosa (azúcar) en sangre. Para ello, utiliza 12 muestras de sangre de las que conoce las concentraciones de galactosa (X), con 3 muestras de cada uno de 4 niveles distintos: 1, 4, 7 y 10. La concentración de galactosa medida con el nuevo método (Y) se observó para cada muestra. Se ajustó el modelo lineal

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

obteniéndose los siguientes valores

$$\begin{aligned} n = 12, \quad \widehat{\beta}_0 = -0.1, \quad \widehat{\beta}_1 = 1.017, \\ \overline{X} = 5.5, \quad \overline{Y} = 5.492, \quad \sum \left(X_i - \overline{X}\right)^2 = 135, \\ \widehat{\sigma}_{\widehat{\beta}_1} = 0.0142. \end{aligned}$$

(a) Testee a nivel 0.05 si hay asociación lineal entre las variables X e Y.

(b) Se cuenta con un nuevo paciente, para el cual se quiere conocer la concentración de galactosa. Se le mide la concentración de galactosa con el procedimiento rápido, lo cual da 6.52. Dé un estimador puntual y un intervalo de predicción de nivel 0.95 para dicho valor.