Estadística (Química)
Práctica 7 - Análisis de la varianza

Comentario: En todos los ejercicios propuestos

a) defina las variables aleatorias y los parámetros involucrados.

b) de ser posible indique:

i. la distribución de las variables aleatorias

ii. el significado intuitivo de los parámetros.

(a) plantee las hipótesis nula y alternativa, e indique el nivel que usará para el test. Elija un test, calcule el valor del estadístico, calcule o acote el $p-$valor e indique la conclusión del test. Si el nivel del test no se especifica en el enunciado, tome por default 0.05. Dé las conclusiones en los términos del problema.

(b) compare los resultados de hacer las cuentas a mano con las salidas obtenidas con el R, de manera de chequear las primeras y aprender a usar las segundas, en aquellos ejercicios en los que ambas cosas sean posibles.

1. Se analizaron 6 muestras de cada uno de tres tipos de cereal producidos en cierta región para determinar el contenido de tiamina. Los resultados fueron los siguientes:

	Trigo	Maíz	Avena
	5.2	6.5	9.3
	4.5	8.0	7.1
	6.0	6.1	8.8
	6.1	7.5	8.0
	6.7	5.9	6.5
	5.8	5.6	8.2
media	5.72	6.60	7.98
desvío	0.77	0.95	1.04

(a) Suponga que se verifican los supuestos del modelo del Análisis de la Varianza. Construya la tabla y aplique el test F para decidir si existen diferencias en las medias del contenido de tiamina de los tres cereales a nivel 0.05. Defina las variables aleatorias y los parámetros involucrados, escriba el modelo bajo el cual vale el test F, establezca claramente las hipótesis de dicho test, dé el p-valor y escriba la conclusión.

(b) Encuentre intervalos de confianza de nivel simultáneo 95% para todas las diferencias de medias. Utilice el método que prefiera. ¿Cuántas comparaciones de a pares pueden hacerse? Si en el inciso (a) halló diferencias significativas, detecte a partir de los intervalos recién hallados cuáles son los cereales que difieren en sus contenidos medios de tiamina, con nivel simultáneo 5%.

(c) Los precios de los cereales en el mercado son los siguientes

Cereal	Precio (US$/Tonelada)
Avena	313
Maíz	223
Trigo	198

Se dispone de un presupuesto para comprar una gran cantidad de cereal. Se desea comprar aquel cereal que tenga el mayor contenido medio poblacional de tiamina, que se pueda pagar con dicho presupuesto. Basándose en el resultado del ítem anterior, decida qué cereal compraría en cada una de las siguientes situaciones:

i. Se puede gastar hasta US$ 240 por tonelada, ¿cuál cereal prefiere?

ii. Se puede gastar hasta US$ 350 por tonelada, ¿cuál cereal prefiere?

2. En un experimento se midió la pérdida de humedad de 6 variedades de sorgo sometidas a un cierto tratamiento. Sea Y_{ij} la pérdida de humedad de la semilla j de la variedad i. Considere el siguiente modelo:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \qquad 1 \leq i \leq 6,\ 1 \leq j \leq 8$$

donde ε_{ij} son variables aleatorias independientes e idénticamente distribuidas y μ_i es la esperanza de la pérdida de humedad para la variedad i. Los datos obtenidos son los siguientes:

Var. 1	Var. 2	Var. 3	Var. 4	Var. 5	Var. 6
11.55	10.12	9.53	11.28	10.38	9.77
11.52	9.34	9.51	11.22	10.40	10.56
11.61	9.34	9.95	11.05	10.18	10.36
11.61	10.15	9.43	11.05	10.40	10.60
11.89	9.48	9.99	11.08	10.07	10.56
11.70	9.27	9.48	11.02	10.35	10.46
11.61	10.18	9.96	11.97	10.35	10.13
11.49	9.68	9.25	11.15	10.54	10.34

(a) Calcule las medias y los desvíos muestrales de cada variedad y analice mediante técnicas gráficas si existen diferencias entre las distintas variedades.

(b) Suponga que se verifican los supuestos del modelo del Análisis de la Varianza. Construya la tabla y aplique el test F para decidir si existen diferencias entre las medias de pérdida de humedad de las distintas variedades a nivel 0.05.

(c) Encuentre intervalos de confianza de nivel simultáneo 95% para todas las diferencias de medias. ¿Cuántas intervalos (comparaciones de a pares) se pueden hacer? Utilice el método que prefiera. Si en el inciso anterior halló diferencias significativas, detecte a partir de los intervalos recién hallados cuáles son las variedades que difieren en sus pérdidas medias de humedad, con nivel simultáneo 5%.

(d) Calcule a mano el intervalo de confianza para la diferencia de medias de las variedades 2 y 1.

(e) Mirando los boxplots, ¿le parece válido el supuesto de homogeneidad de varianzas? Aplicando el test de Levene, ¿cuál es su conclusión respecto a la homogeneidad de varianzas a nivel 20%? ¿Encuentra alguna contradicción entre la conclusión del boxplot y la del test? ¿Es válido, entonces, usar el modelo de anova para estudiar a estos datos?

(f) Analice si es válido el supuesto de normalidad haciendo un histograma del conjunto de todos los residuos. Aplicando el test de Shapiro Wilk a los residuos, ¿a qué conclusión llega? Con esta nueva información, utilice alguna herramienta que le permita confirmar (o dudar de) su conclusión del ítem anterior respecto de la homoscedasticidad.

3. Un experimento comenzó dividiendo un grupo de ratas de 20 días de edad en tres grupos al azar. Un grupo recibió ATRO (atropina) solamente, el segundo grupo recibió SPI (spiroperidol) solamente y el tercer grupo recibió COMB (una combinación de ambos). Una hora después que la droga fuera suministrada se midió el tiempo en segundos de reacción de cada rata ante un estímulo. Se obtuvieron los siguientes

tiempos de reacción en segundos:

	ATRO	COMB	SPI
	10.5	16.0	35.8
	0.8	5.9	10.5
	0.7	11.5	10.5
	0.7	4.4	5.2
	0.3	17.7	20.9
	0.7	13.5	44.2
	0.3	60	19.6
		2.3	20.7
media	2.000	16.413	20.925
desvío	3.7537	18.471	13.253

(a) Construya boxplots para los datos y describa las características observadas.

(b) ¿Es razonable suponer el modelo del Análisis de la Varianza a un factor para estos datos?

(c) Intente aplicar el Análisis de la Varianza y aplique el test de Shapiro-Wilk a los residuos, ¿cuál es la conclusión?

(d) Sólo para comparar, calcule el valor p del test F para estudiar la hipótesis de igualdad de medias (aunque no es correcto, ¿verdad?).

(e) Aplique una transformación logarítmica a los datos (calcule log decimal para que todos lleguemos a los mismos resultados, pero sería equivalente calcular ln ya que difieren en una constante multiplicativa).

(f) Repita (a), (b) y (c) pero con los datos transformados.
De acá en adelante continúe el análisis estadístico con los datos originales o transformados, según le parezca más conveniente.

(g) Aplique el test F para comparar las medias de los tres tratamientos.

(h) En el caso de rechazar H_0 con el test F, detecte para cuáles de las drogas las respuestas difieren significativamente.

4. Se midieron las concentraciones de plasma (en nanogramos por mililitro) de 10 perros sometidos a 3 tratamientos distintos. Las mediciones se presentan en la siguiente tabla:

Perro	1	2	3	4	5	6	7	8	9	10
Tratamiento 1	0.28	0.51	1.00	0.39	0.29	0.36	0.32	0.69	0.17	0.33
Tratamiento 2	0.30	0.39	0.63	0.68	0.38	0.21	0.88	0.39	0.51	0.32
Tratamiento 3	1.07	1.35	0.69	0.28	1.24	1.53	0.49	0.56	1.02	0.30

Uno de los supuestos del modelo de ANOVA no es válido en este caso. Sin hacer cuentas, responda cuál y por qué.

5. En un ensayo de colaboración se envió una muestra de una sustancia que contiene olaquindox a tres laboratorios. Cada laboratorio realizó mediciones repetidas e independientes utilizando un detector ultravioleta. Los resultados obtenidos se presentan en la siguiente tabla:

	Lab 1	Lab 2	Lab 3
	21.0	26.5	21.2
	23.8	27.1	21.4
	23.0	25.9	22.6
	22.1	26.2	23.7
	22.8	25.6	21.9
media	22.54	26.26	22.16
desvío	1.05	0.58	1.02

(a) Mediante técnicas gráficas compare las mediciones obtenidas por los tres laboratorios y describa lo que ve.

(b) Plantee un modelo para analizar las diferencias entre laboratorios. ¿Qué supuestos debe hacer? ¿Puede analizar la validez de los mismos? En caso de que su respuesta sea afirmativa, analícelos.

(c) Escriba la tabla del Análisis de la Varianza y aplique un test de la hipótesis de que los tres laboratorios miden con igual media, a un nivel del 5%.

(d) ¿Se puede decir que la detección media de olaquindox de algún laboratorio es mayor o menos que la de los otros dos? Explique qué método utilizó para llegar a la conclusión.

(e) Calcule los intervalos de confianza para las diferencias de medias entre todos los pares de laboratorios.

6. Sea $Y_{ij} = \mu_i + \varepsilon_{ij}$ con $1 \leq i \leq I = 3$, $1 \leq j \leq J$, $\varepsilon_{ij} \sim N(0, \sigma^2)$ independientes. Considere los eventos

$$A_1 = \{\mu_1 - \mu_2 \in [a_1, b_1]\}$$
$$A_2 = \{\mu_1 - \mu_3 \in [a_2, b_2]\}$$
$$A_3 = \{\mu_2 - \mu_3 \in [a_3, b_3]\}$$

(a) Interprete los eventos A_1, A_2 y A_3 en el contexto de intervalos de nivel simultáneo. Y también interprete el evento $A_1 \cap A_2 \cap A_3$.

(b) Hallar la probabilidad de $A_1 \cap A_2 \cap A_3$ en el caso en el que A_1, A_2 y A_3 sean eventos independientes.

(c) Usando que

$$P\left(A_1^c \cup A_2^c \cup A_3^c\right) \leq P\left(A_1^c\right) + P\left(A_2^c\right) + P\left(A_3^c\right),$$

encontrar una cota inferior para $P\left(A_1 \cap A_2 \cap A_3\right)$ sin asumir independencia de los eventos A_i.

(d) ¿Cómo se generalizaría si I (la cantidad de tratamientos o grupos a comparar) fuera mayor a 3? Pruebe que

$$P\left(\bigcap_{i=1}^{I} A_i\right) \geq 1 - \sum_{i=1}^{I} P\left(A_i^c\right)$$

e interprételo.

7. Se ha realizado un experimento para determinar si la densidad de un tipo de ladrillo se ve afectada por la temperatura a la que es horneado. Para ello, se cocieron 19 ladrillos a distintas temperaturas, obteniéndose los siguientes resultados:

Temperatura	Densidad				
Muy baja	21.274	20.822	20.452	21.291	19.897
Baja	22.309	23.999	23.304	21.532	22.783
Media	24.010	23.132	23.848	22.300	23.153
Alta	25.445	24.660	24.229	22.895	

Por el modo en que fue llevado a cabo el experimento, se puede suponer independencia entre las observaciones.

(a) Se quiere aplicar el modelo de análisis de la varianza (ANOVA) para decidir si existen diferencias en la densidad media de los distintos grupos. El modelo de ANOVA para estos datos es

. . . . . .

donde las variables aleatorias y los parámetros involucrados son (defínalos con palabras)

. . . . . .

.....

.....

Aquí $1 \leq i \leq k =$ ..... y los n_i son .....

Las hipótesis que testea el ANOVA son

.....

.....

(b) Los supuestos necesarios para aplicar dicho modelo son

.....

.....

.....

(c) En base a los gráficos y salidas del R que se dan a continuación, decidir si se cumplen los supuestos del modelo de análisis de la varianza. Indique qué gráfico o salida utiliza para cada conclusión.

.....

.....

.....

(d) En base a las salidas del R que se dan a continuación, completar la siguiente tabla de ANOVA, en los espacios recuadrados. Dar las estimaciones de **todos** los parámetros del modelo que surjan del ajuste de ANOVA.

```
> salida<-aov(densidad~temp.f)
> summary(salida)
             Df       Sum Sq     Mean Sq    F value  Pr(>F)
temp.f       [ ]      31.169     [      ]    15.276   7.894e-05 ***
Residuals    15       [      ]    0.6801
```

.....

.....

.....

(e) El estadístico utilizado para este test es (completar con una fórmula, explicando cada uno de los términos involucrados en ella que no haya definido anteriormente)

.....

.....

Su valor observado es [ ] y su pvalor resulta ser [ ] .

Dar la conclusión del test cuyas hipótesis exhibió en (a), a nivel 0.05.

.....

.....

(f) Se desean hacer comparaciones múltiples con el método de Bonferroni, para identificar cuáles son los grupos cuyas densidades medias difieren, a nivel global 0.95. Indique cuántas comparaciones debe realizar [ ] . Los grados de libertad de la distribución t involucrada en el cálculo del

punto crítico son [ ] . Dicho punto crítico corresponde al percentil [ ] de la t. El valor crítico de la t resulta ser 3.036. La fórmula de dichos intervalos de confianza es

.....

Calcule los intervalos correspondientes a las comparaciones entre los grupos

"alta" y "muy baja" .....

"alta" y "baja" .....

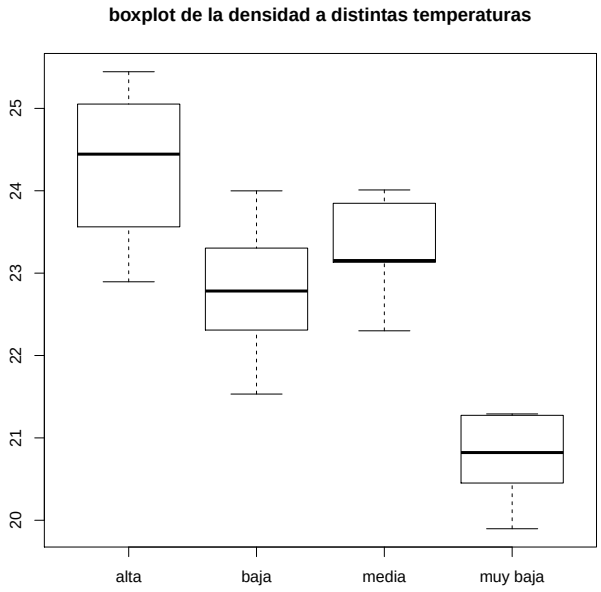
(g) De las dos comparaciones entre grupos realizadas más arriba, ¿alguna de ellas permite concluir que las medias poblacionales de los grupos correspondientes difieren entre sí a nivel conjunto 0.95? ¿Cuál/es? Justifique.

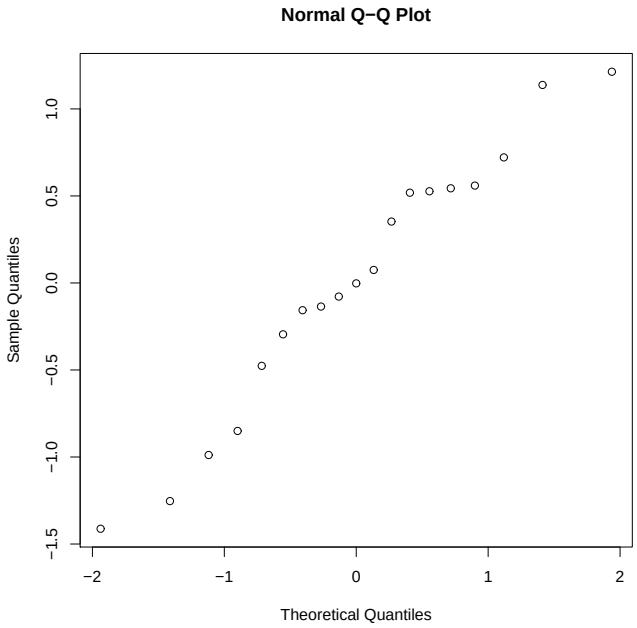
.....

.....

```
> shapiro.test(residuals(salida))

 Shapiro-Wilk normality test

data:  residuals(salida)

W = 0.9621, p-value = 0.6143

> bartlett.test(densidad,temp.f)

 Bartlett test of homogeneity of variances

data:  densidad and temp.f

Bartlett's K-squared = 1.4572, df = 3, p-value = 0.6922

> levene.test(densidad,temp.f)

 modified robust Brown-Forsythe Levene-type test based on the absolute

 deviations from the median

data:  densidad

Test Statistic = 0.4292, p-value = 0.735

> tapply(densidad, temp.f,mean)

 alta     baja      media     muy baja

24.30725  22.78540  23.28860   20.74720

> tapply(densidad, temp.f,var)

 alta      baja         media     muy baja

1.1398269 0.8849363  0.4633968 0.3472037
```

Los dos gráficos que siguen son los boxplots de la densidad a distintas temperaturas, y el qq-plot de los residuos del ajuste ANOVA.


boxplot de la densidad a distintas temperaturas
20
21
22
23
24
25
alta
baja
media
muy baja


Normal Q−Q Plot
Sample Quantiles
−1.5
−1.0
−0.5
0.0
0.5
1.0
−2
−1
0
1
2
Theoretical Quantiles