Estadística Aplicada 2018

Examen de regularidad.

NOMBRE Y APELLIDO:

Plazo máximo de entrega: Martes 13 de Noviembre, 8:00 hs.

Ejercicio 1 (puntos) Un científico de alimentos diseñó un estudio para evaluar el efecto de la edad de un consumidor sobre cómo califican la amargura de las cuatro marcas más populares de queso cheddar. Se utilizaron dos muestras de cada uno de los cuatro tipos de queso cheddar (T). Diez evaluadores de gustos profesionales (R) son seleccionados al azar de cada uno de tres grupos de edad (A): 20-29, 30-49 y 50-70 años. Cada uno de los 30 evaluadores probará ocho porciones de queso presentadas en orden aleatorio. Las ocho porciones consisten en una porción de cada uno de las 8 muestras (los evaluadores desconocen el tipo de queso dentro de cada porción). Cada evaluador registra un puntaje de amargura después de cada bocado de queso: Y_{ijkm}. Hay suficiente tiempo de espera entre las 8 calificaciones para que las calificaciones no estén sesgadas por una calificación previa del gusto. El científico de alimentos está interesado en estimar si la diferencia en el puntaje promedio de amargor para los cuatro tipos de quesos es constante en los tres grupos de edad. También le interesa determinar el tamaño de la variación en las calificaciones dentro de los tres grupos de edad.

a) Complete la siguiente tabla ANOVA ingresando los valores de los grados de libertad y los cuadrados medios esperados (*teóricos*).

Fuente de variabilidad	Df	MS	$\mathbb{E}$(MS)
T		3.79	
A		13.27	
T$*$A		2.78	
R(A)		2.58	
T$*$R(A)		1.06	
Error		0.91	

b) Se ajustó el siguiente modelo para los datos donde Y_{ijkm} es la calificación de amargor de la m-ésima porción de queso tipo i del k-ésimo evaluador en el grupo de edad j:

$$Y_{ijkm} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + C_{k(j)} + D_{ik(j)} + \epsilon_{ijkm}$$

Establezca todas las condiciones que se deben asumir sobre los términos del modelo para poder llevar a cabo los procedimientos de análisis de varianza.

c) A nivel $\alpha = 0.05$, evalúe el efecto de la *Edad del evaluador* en el *promedio de la clasificación de amargor del queso*. Tenga en cuenta que los números que figuran en la tabla de ANOVA son los cuadrados medios (MS) NO la suma de cuadrados (SS).

d) Estime el error estándar de la diferencia estimada en las calificaciones promedio entre los *Grupos de edad* 20-29 y 50-70.

e) ¿Considera que, en promedio, las calificaciones otorgadas por los evaluadores de más de 50 años difiere de las otorgadas por los evaluadores más jóvenes?

Ejercicio 2 (puntos) Algunas especies de hongos pueden causar grandes pérdidas económicas por producir enfermedades en los cultivos o el deterioro de los alimentos. Un microbiólogo diseñó un estudio para determinar qué condiciones favorecen el crecimiento de una especie de hongo en particular. Los factores seleccionados para el estudio son tres: el medio de cultivo (M1, M2); nivel de humedad (45 %, 60 %, 85 %); y tiempo de cultivo (25, 50, 75 horas). Dos muestras del hongo estudiado fueron asignados aleatoriamente a cada uno de los 18 tratamientos al inicio del experimento. La respuesta estudiada es el crecimiento del hongo. Los datos están en el archivo *hongos.txt*, aunque la mayor parte del ejercicio no los requiere.

a) Escriba un modelo para el análisis de estos datos. Indique el significado de cada parámetro que usa en el contexto del problema, valores de los subíndices y todas las suposiciones realizadas.

b) Explique por qué decidió incluir (o no) un término correspondiente a la interacción triple entre los tres factores en el modelo propuesto.

c) ¿Qué gráficos utilizaría para dar una respuesta exploratoria al problema planteado? Explique qué precauciones tendría al realizarlo y por qué.

d) El ajuste del modelo a los datos arrojó la siguiente tabla anova:

```
Df Sum Sq Mean Sq F value   Pr(>F)
fTiempo                   2   86.3    43.1   5.609  0.01278 *
fHumedad                  2  540.6   270.3  35.144 6.09e-07 ***
fMedio                    1 1009.1  1009.1 131.206 1.06e-09 ***
fTiempo:fHumedad          4   34.9     8.7   1.135  0.37155
fTiempo:fMedio            2  123.8    61.9   8.049  0.00318 **
fHumedad:fMedio           2  132.6    66.3   8.622  0.00236 **
fTiempo:fHumedad:fMedio   4   29.4     7.4   0.957  0.45459
Residuals                18  138.4     7.7
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

En función de estos resultados, escriba un nuevo modelo que le parezca adecuado para los datos.

e) A partir del último modelo, ¿qué gráficos exploratorios podría realizar ahora para estudiar la dependencia entre la respuesta y los factores considerados? ¿En qué se diferencia esta respuesta de la dada en el item *c)*?

f) Prosiguiendo con su análisis, el investigador obtuvo la siguiente tabla:

```
Call:
lm(formula = crecimiento ~ fTiempo * fHumedad * fMedio - (fTiempo:fHumedad:fMedio +
fTiempo:fHumedad))

Residuals:
```

```
Min      1Q  Median      3Q     Max
-5.0111 -1.3486 -0.0333  1.1403  6.0722

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)           8.2889     0.6583  12.592 1.43e-12 ***
fTiempo.L            -0.4950     1.1402  -0.434 0.667782
fTiempo.Q            -0.2586     1.1402  -0.227 0.822377
fHumedad.L            2.6045     1.1402   2.284 0.030762 *
fHumedad.Q            5.2732     1.1402   4.625 9.05e-05 ***
fMedioM2             10.5889     0.9310  11.374 1.36e-11 ***
fTiempo.L:fMedioM2    6.3050     1.6125   3.910 0.000591 ***
fTiempo.Q:fMedioM2    1.2315     1.6125   0.764 0.451879
fHumedad.L:fMedioM2   5.9161     1.6125   3.669 0.001101 **
fHumedad.Q:fMedioM2  -3.0346     1.6125  -1.882 0.071075 .
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 2.793 on 26 degrees of freedom
Multiple R-squared:  0.9032,Adjusted R-squared:  0.8697
F-statistic: 26.96 on 9 and 26 DF,  p-value: 5.566e-11
```

¿A qué análisis corresponde este resumen? Qué representa cada uno de los términos?

g) En función del reporte presentado en la tabla anterior, escriba un modelo genérico (sin el valor de los parámetros) adecuado para describir la (superficie de) respuesta del crecimiento de la especie de hongo estudiada en función de los factores considerados.

h) Explique cómo procedería para determinar cuál es la combinación de factores que más favorece el cremiento de hongos. ¿Encuentra alguna limitación o dificultad en el procedimiento que propone?

i) Indique cuál es el crecimiento medio del hongo que predice el modelo si se cultiva en el medio M2 durante 50 horas y con una humedad del 60 %.

j) Todo el análisis efectuado hasta aquí supuso que las mediciones son independientes. ¿Considera que es adecuada esa suposición? Justifique su respuesta y, en caso de estar en desacuerdo con la metodología empleada, indique qué cambios introduciría en el procedimiento de análisis (no necesita hacer nada en R).

Ejercicio 3 Elija la mejor respuesta a cada pregunta entre las opciones disponibles. Debe elegir sólo UNA opción para cada pregunta. JUSTIFIQUE brevemente su elección.

a) Un veterinario quiere investigar $t = 5$ tratamientos para controlar parásitos en mascotas. Para ello determina que necesitará $r = 9$ réplicas por tratamiento. Existe una variabilidad notoria en la efectividad de los tratamientos, de modo que el veterinario decide usar grupos homogéneos de mascotas, destinando 9 camadas de cahorros hermanos como variable bloque, con 5 cachorros por camada. En el experimento obtiene $MS_{bloque} = 34.2$ y $MS_{error} = 11.4$. El veterinario quiere realizar un segundo experimento pero sin usar bloques. ¿Cuál es el mínimo número de réplicas

que necesita por tratamiento para igualar la precisión en la estimación de la media de cada tratamiento que obtenía en el experimento con bloques?

1) 5
2) 9
3) 13
4) 62
5) No se puede calcular con la información provista.

b) Si desea utilizar un procedimiento de comparaciones múltiples para testear 25 contrastes ortogonales referidos a 40 medias de tratamientos, qué valor de α_{PC} (probabilidad de error de tipo I de cada comparación individual) usaría a fin de obtener una probabilidad global de error de tipo I de nivel $\alpha_F = 0.05$ para la familia de tests simultáneos?

1) 0.05000
2) 0.00125
3) 0.00128
4) 0.00200
5) 0.00205

c) En un estudio para medir diferencias en el contenido medio de potasio de variedades de banana, un nutricionista decide seleccionar en forma aleatoria una misma cantidad de bananas de cada una de las variedades incluidas en el estudio. ¿Por qué considera que es una elección acertada utilizar igual cantidad de réplicas para cada tratamiento?

1) Para obtener una distribución F exacta para el estadístico del test.
2) Para simplificar la forma de los cuadrados medios esperados (E(MS)).
3) Aumentar la potencia en caso que los tratamientos tengan varianza distinta.
4) Mantener control de la probabilididad de error de tipo I en caso que los tratamientos tengan varianzas distintas.
5) Todas las anteriores.

d) En un estudio para determinar si existe diferencia en la resistencia media de las fibras de algodón producidas por fabricantes de Estados Unidos. Para ello, el investigador selecciona 10 fabricantes al azar y de cada uno de ellos toma 20 ejemplares de fibra, escogidos al azar de sus depósitos. Luego de medir la resistencia a la tensión de cada ejemplar de fibra, el experimentador decide utilizar el procedimiento de comparaciones múltiples de Tukey para estudiar diferencias significativas entre fabricantes. ¿Cuál es la mayor crítica que usted haría a este estudio?

1) Tukey HSD es demasiado conservativo para detectar diferencias de medias pequeñas.
2) El uso de Tukey HSD podría incrementar la probabilidad de error tipo II.
3) El uso de Tukey HSD podría incrementar la probabilidad de error tipo I.
4) Tukey HSD debería usarse cuando los tratamientos tienen efectos fijos.
5) Ninguna de las anteriores. Con la información provista, Tukey HSD parece adecuado para el experimento.

e) En un diseño completamente aleatorizado con un factor cuantitativo F_1 con 4 niveles y un factor cualitativo F_2 con tres niveles, el experimentador desea saber si existe una tendencia lineal en las

respuestas medias respecto a los niveles de F_1. La tabla ANOVA indica que la interacción $F_1 : F_2$ no es significativa. ¿Cuál de los siguientes contrastes le permitiría responder a su pregunta de interés?

1) $L = -3\mu_{11} - \mu_{21} + \mu_{31} + 3\mu_{41} - 3\mu_{12} - \mu_{22} + \mu_{32} + 3\mu_{42}$
2) $L = 3\mu_{11} + \mu_{21} - \mu_{31} - 3\mu_{41} - 3\mu_{12} - \mu_{22} + \mu_{32} + 3\mu_{42}$
3) $L = \mu_{11} + \mu_{21} + \mu_{31} + \mu_{41} - \mu_{12} - \mu_{22} - \mu_{32} - \mu_{42}$
4) $L = \mu_{11} - \mu_{21} - \mu_{31} + \mu_{41} - \mu_{12} + \mu_{22} + \mu_{32} - \mu_{42}$
5) Ninguno de los anteriores.

f) Un experimentador está diseñando un experimento completamente aleatorizado que incluye dos factores: A con 2 niveles fijos y B con 3 niveles fijos. Piensa inicialmente en utilizar 10 réplicas por tratamiento. Desea determinar si esta cantidad será suficiente para tener una potencia de al menos 0.80 para detectar una diferencia de medias de 19 unidades usando un test de nivel $\alpha = 0.05$. Por experiencia sabe que $\sigma^2 = 100$ es un estimador razonable de la variabilidad experimental. ¿En qué intervalo se encuentra la potencia del test F para este experimento?

- (0,0.5)
- (0.5,0.65)
- (0.65,0.8)
- (0.8,1.0)
- No se puede calcular con la información suministrada.