Examen de regularidad
Estadística aplicada 2018

Alumno: Emiliano P. López
Docentes: Diego Tomassi / Antonella Gieco

November 12, 2018

Ejercicio 1

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

Se completaron los E(MS) utilizando el modelo con restricciones (como los arroja R)

Fuente de variabilidad	Df	MS	E(MS)
T	(t-1) = (4-1) = 3	3.79	$\sigma^2 + r\sigma^2_{T*R(A)} + acr\theta^2_T$
A	(a-1) = (3-1) = 2	13.27	$\sigma^2 + rt\sigma^2_{R(A)} + tcr\theta^2_A$
T*A	(t-1)(a-1) = (4-1)(3-1) = 6	2.78	$\sigma^2 + r\sigma^2_{TR(A)} + cr\theta^2_{TA}$
R(A)	a(c-1) = 3(10-1) = 27	2.58	$\sigma^2 + rt\sigma^2_{R(A)}$
T*R(A)	(t-1)a(c-1) = (4-1)3(10-1) = 81	1.06	$\sigma^2 + r\sigma^2_{TR(A)}$
Error	*tac(r-1) = 4*3*10*(2-1) = 120	0.91	σ^2

Donde,

- t es la cantidad de niveles del factor T (tipos de quesos cheddar)
- a es la cantidad de niveles del factor A (grupos de edad)
- c es la cantidad de niveles del factor R (profesionales)
- r es la cantidad de réplicas

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.

El modelo previo es de efectos mixtos, esto es, de términos fijos y aleatorios, donde para los efectos aleatorios es de interés estudiar la varianza de dichos efectos mientras que para los fijos es de interés probar la hipótesis directamente sobre sus medias.

Las suposiciones sobre el error aleatorio es que sigue una distribución normal con media cero y varianza constante, $N(0, \sigma^2)$, y son independientes entre sí. A continuación se detalla cada término del modelo:

- μ es la media global
- α_i efecto fijo, i niveles para el factor tipos de queso (T)
- β_j efecto fijo, j niveles para el factor grupos de edad (A)
- $(\alpha\beta)_{ij}$ efecto fijo interacción entre T y A
- $C_{k(j)}$ efecto aleatorio, k niveles de profesionales (R) anidados a los grupos de edad
- $D_{ik(j)}$, efecto aleatorio interacción entre los tipos de queso y los profesionales anidados a los grupos de edad
- ϵ_{ijkl} : es el error experimental, con $l = 1 \cdots 120$

Los términos $C_{k(j)}$, $D_{ik(j)}$ y ϵ_{ijkm} son variables aleatorias independientes, normales, con media cero y varianzas σ^2_C, $\sigma^2_{\alpha C}$ y σ^2 respectivamente, por lo tanto, en la varianza total del modelo los factores α y β no aportan dado que son fijos, pero sí se mantiene el componente de interacción D por ser C aleatorio. Esto puede ser escrito del siguiente modo:

$$var(Y_{ijkm}) = \sigma^2_C + \sigma^2_{\alpha C} + \sigma^2$$

c)

- H0: $\beta_j = 0$, para $j = 1 \cdots a$
- H1: $\beta_j \neq 0$

$$F = \frac{MSA}{MSR(A)} = \frac{13.27}{2.58} = 5.1434$$

Como $pvalor = 1 - pf(5.1434, 2, 27) = 0.0128 < 0.05$, por lo tanto rechazamos H0 y concluímos que el efecto de la edad del evaluador es significativa.

Ejercicio 2

```
In [11]: # lectura de datos
         datos = read.csv("hongos.txt", sep = "\t")
         head(datos)
         attach(datos)
```

time	humidity	grow.media	growth
25	45	M1	8.1
25	45	M1	8.9
25	60	M1	2.2
25	60	M1	1.1
25	85	M1	18.6
25	85	M1	12.3

```
In [12]: timef = as.factor(time) # conversion a factor
         humidityf = as.factor(humidity)
         growmediaf = as.factor(grow.media)
```

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.

El modelo es:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta\gamma)_{ijk} + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \epsilon_{ijkl}$$

A continuación se describen los parámetros del modelo y el rango de los índices:

- μ: es la media global
- α_i: es el tiempo, con $i = 1 \cdots 3$
- β_j: es la humedad, con $j = 1 \cdots 3$
- γ_k: es el medio de crecimiento, con $k = 1 \cdots 2$
- $(\alpha\beta\gamma)_{ijk}$: efecto de la interacción entre el tiempo, humedad y crecimiento
- $(\alpha\beta)_{ij}$: efecto de la interacción entre tiempo y humedad
- $(\alpha\gamma)_{ik}$: efecto de la interacción entre tiempo y medio de crecimiento
- $(\beta\gamma)_{jk}$: efecto de la interacción entre la humedad y medio de crecimiento
- ϵ_{ijkl} : es el error experimental, con $l = 1 \cdots 36$

Las suposiciones del modelo son:

- $\epsilon_{ijkl} \sim N(0, \sigma^2)$, independientes idénticamente distribuídos

Las restricciones del modelo son:

- $\alpha_1 = 0$
- $\beta_1 = 0$
- $\gamma_1 = 0$
- $(\alpha\beta)_{1j} = 0$
- $(\alpha\beta)_{i1} = 0$

- $(\alpha\gamma)_{1k} = 0$
- $(\alpha\gamma)_{i1} = 0$
- $(\beta\gamma)_{1k} = 0$
- $(\beta\gamma)_{j1} = 0$
- $(\alpha\beta\gamma)_{1jk} = 0$
- $(\alpha\beta\gamma)_{i1k} = 0$
- $(\alpha\beta\gamma)_{ij1} = 0$

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

El objetivo de un diseño factorial es estudiar el efecto de varios factores sobre una o varias respuestas, cuando se tiene el mismo interés sobre todos los factores, como es el caso del ejercicio. De manera que se incluyó un término de interacción triple en el modelo ya que se pretende estudiar tanto los efectos individuales como de las interacciones de varios factores sobre una o varias respuestas.

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

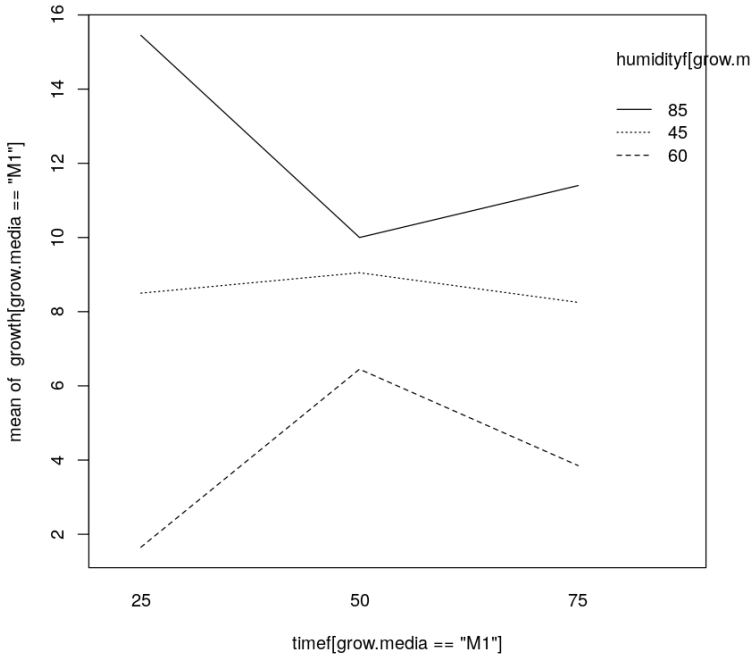
Debido a que tenemos tres factores fijos que interactúan entre sí, no tendría sentido ver el boxplot, ya que la respuesta dependerá de uno o varios factores individuales e interactuando entre sí. Idealmente sería útil analizar gráficos de interacción, dejando fijo uno de los factores y observando la respuesta de la interacción entre los dos restantes.

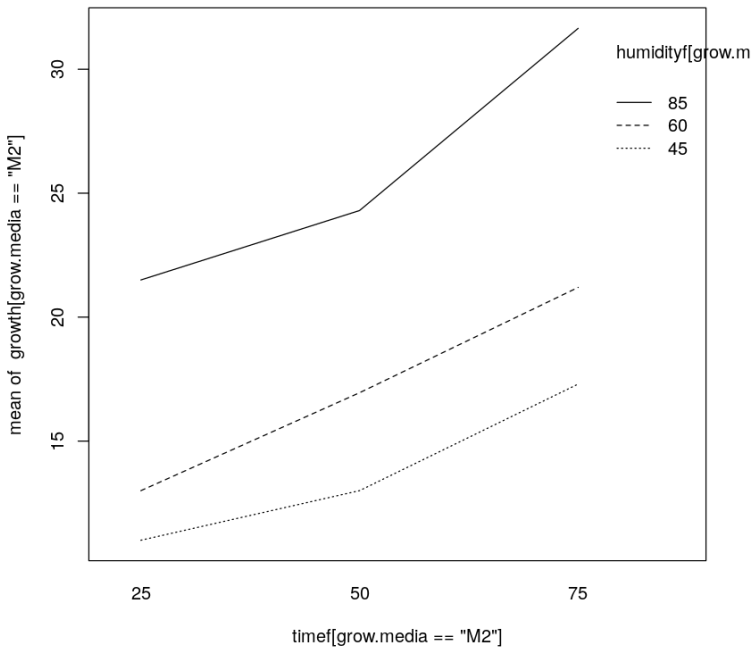
En este caso particular uno podría pensar visualizar la interacción entre la humedad y el tiempo fijando cada medio de cultivo (M1 y M2). Para tener una idea del comportamiento, lo más razonable sería dejar el tiempo en el eje de abscisas y ver para las distintas humedades la correspondiente respuesta (crecimiento del hongo).

Debemos tener la precaución en estos gráficos de interacción que las conclusiones no son definitivas, sino que deben comprobarse numéricamente ya que la escala en el eje de respuesta puede darnos una idea falsa de interacción o falta de ella.

A continuación observamos uno de los tipos de gráficos de interacción mencionados:

```
In [17]: interaction.plot(timef[grow.media=="M1"], humidityf[grow.media=="M1"],
                          growth[grow.media=="M1"])
         interaction.plot(timef[grow.media=="M2"], humidityf[grow.media=="M2"],
                          growth[grow.media=="M2"])
```


mean of growth[grow.media == "M1"]
timef[grow.media == "M1"]
humidityf[grow.m
85
45
60
2
4
6
8
10
12
14
16
25
50
75


mean of growth[grow.media == "M2"]
timef[grow.media == "M2"]
humidityf[grow.m
85
60
45
15
20
25
30
25
50
75

d) El ajuste del modelo a los datos arrojo la siguiente tabla anova... En función de estos resultados, escriba un nuevo modelo que le parezca adecuado para los datos.

```
In [14]: modelo = aov(growth~timef*humidityf*growmediaf)
         summary(modelo)
```

```
                           Df Sum Sq Mean Sq F value   Pr(>F)
timef                       2   86.3    43.1   5.609  0.01278 *
humidityf                   2  540.6   270.3  35.144 6.09e-07 ***
growmediaf                  1 1009.1  1009.1 131.206 1.06e-09 ***
timef:humidityf             4   34.9     8.7   1.135  0.37155
timef:growmediaf            2  123.8    61.9   8.049  0.00318 **
humidityf:growmediaf        2  132.6    66.3   8.622  0.00236 **
timef:humidityf:growmediaf  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
```

De la tabla ANOVA previa podemos descartar aquellos factores o interacción de factores cuyo *p-valor* sea mayor a la significancia $\alpha = 0.05$, ya que sus efectos no son significativos, de este modo se excluye del modelo los siguientes términos:

- timef:humidityf, $(\alpha\beta)$, ya que $0.37155 > 0.05$
- timef:humidityf:growmediaf, $(\alpha\beta\gamma)$, ya que $0.45459 > 0.05$

Finalmente el modelo reducido nos queda:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \epsilon_{ijkl}$$

e) A partir del ultimo modelo, ¿que gráficos exploratorios podria realizar ahora para estudiar la dependencia entre la respuesta y los factores considerados? ¿En que se diferencia esta respuesta de la dada en el ítem c)?

Luego de descartar el término de interacción doble `timef:humidityf` y triple `timef:humidityf:growmediaf` realizaría dos gráficos de interacción para los términos cuya interacción sea significativa, es decir:

- `timef:growmediaf`: dejando fijo la humedad (β)
- `humidityf:growmediaf`: dejando fijo el tiempo (α)

En el *item c)*, sin tener conocimiento sobre las interacciones, se fijó el `medio de crecimiento` y se observaron las interacciones entre el tiempo y la humedad, sin embargo ahora vemos que este factor interactúa con cada uno de los otros dos factores en forma separada, por lo que es relevante analizar su comportamiento sin ser éste el factor fijado.

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

```
In [16]: # genera tabla del enunciado
         contrasts(humidityf) = contr.poly(3)
         modelo2 = lm(growth~timef*humidityf*growmediaf -
                      (timef:humidityf:growmediaf + timef:humidityf))
         summary(modelo2)
```

```
Call:
lm(formula = growth ~ timef * humidityf * growmediaf - (timef:humidityf:growmediaf +
    timef:humidityf))

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)               8.53333    1.14018   7.484 6.03e-08 ***
timef50                  -0.03333    1.61246  -0.021 0.983665
timef75                  -0.70000    1.61246  -0.434 0.667782
humidityf.L               2.60451    1.14018   2.284 0.030762 *
humidityf.Q               5.27321    1.14018   4.625 9.05e-05 ***
growmediafM2              6.63333    1.61246   4.114 0.000347 ***
timef50:growmediafM2      2.95000    2.28036   1.294 0.207159
timef75:growmediafM2      8.91667    2.28036   3.910 0.000591 ***
humidityf.L:growmediafM2  5.91613    1.61246   3.669 0.001101 **
humidityf.Q:growmediafM2 -3.03465    1.61246  -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
```

OBSERVACIÓN: La tabla del enunciado contiene un error ya que supuso que la humedad se encontraba equispaciada cuando en realidad no es así, a continuación se usa la misma tabla para que los resultados sean consistentes a los del enunciado.

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

La tabla previa se corresponde con un análisis de tendencia para estudiar si la variable respuesta se incrementa o decrementa cuando varía la *Humedad* usando una estimación lineal o cuadrática. Cuando los niveles de un factor son cuantitativos y presentan un orden es interesante realizar un análisis de tendencias a partir de comparaciones múltiples usando contrastes ortogonales.

Para comprender cuál es la tendencia es necesario observar los *p-valores* comparando cada uno con un nuevo α_{PC} que determine la probabilidad de cometer al menor un error de tipo I para

una familia de contrastes ortogonales. Para el cálculo se requiere un valor deseado de $\alpha_0 = 0.05$ utilizando la siguiente ecuación:

$$\alpha_{PC} = 1 - (1 - \alpha_0)^{1/9} = 0.0056$$

De manera que se cumplen aquellas tendencias donde el p-valor < 0.0056. En la tabla, los factores o interacciones de factores que terminan en `.Q` (o `.50` en este caso) refieren a la tendencia cuadrática, mientras que los que finalizan con `.L` (o `.75` en este caso) a la lineal.

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.

Sea x_1, x_2 y x_3 tiempo, humedad y medio de crecimiento respectivamente, y ϵ el error aleatorio, según la tabla previa el modelo cuadrático genérico tiene la siguiente forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_3 + \beta_6 x_1 x_3 + \beta_7 x_1^2 x_3 + \beta_8 x_2 x_3 + \beta_9 x_2^2 x_3 + \epsilon$$

h) Explique como procedería para determinar cual es la combinación de factores que mas favorece el crecimiento de hongos. ¿Encuentra alguna limitación o dificultad en el procedimiento que propone?

Para determinar la mejor combinación de los factores sería conveniente realizar comparaciones múltiples usando Dunnet, en este caso previamente habría que optimizar para elegir el máximo, para eso es posible usar la función `maxHSU` ya que queremos saber aquellas combinaciones que favorecen el crecimiento. Se recomienda que el tamaño de muestra del tratamiento control sea grande, a fin de estimar su media con mayor precisión.

La limitación que podríamos encontrar es que al contar con pocas réplicas (2 por hongo) tenemos baja potencia

```
In [7]: source("mymultcomp.R")
        fABC = factor(paste(timef,humidityf,growmediaf))
        maxHSU(growth, fABC, alpha=0.05, mse=7.7, dof=18)
```

```
[1] "WARNING: esta funcion considera que todos los ni son iguales"
[1] "50 85 M2"
[1] "75 85 M2"
```

1. NA 2. '50 85 M2' 3. '75 85 M2'

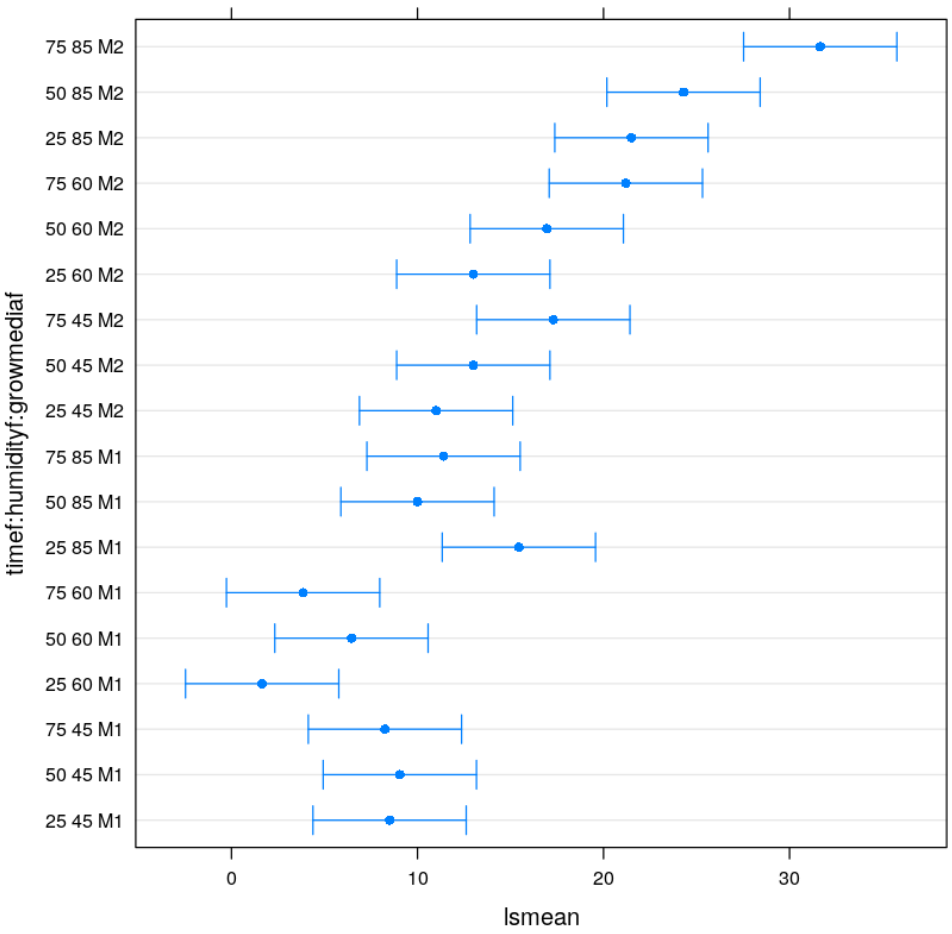
i) Indique cual es el crecimiento medio del hongo que predice el modelo si se cultiva en el medio M2 durante 50 horas y con una humedad del 60%.

A partir de la tabla de medias vemos que para el medio M2, durante 50 horas a una humedad del 60%, tenemos una media de 16.95.

```
In [8]: library(lsmeans)
        lsmeans(modelo, ~timef*humidityf*growmediaf)
        plot(lsmeans(modelo, ~timef*humidityf*growmediaf))
```

```
 timef humidityf growmediaf lsmean       SE df   lower.CL  upper.CL
 25    45         M1           8.50 1.961009 18  4.3800734 12.619927
 50    45         M1           9.05 1.961009 18  4.9300734 13.169927
 75    45         M1           8.25 1.961009 18  4.1300734 12.369927
 25    60         M1           1.65 1.961009 18 -2.4699266  5.769927
 50    60         M1           6.45 1.961009 18  2.3300734 10.569927
 75    60         M1           3.85 1.961009 18 -0.2699266  7.969927
 25    85         M1          15.45 1.961009 18 11.3300734 19.569927
 50    85         M1          10.00 1.961009 18  5.8800734 14.119927
 75    85         M1          11.40 1.961009 18  7.2800734 15.519927
 25    45         M2          11.00 1.961009 18  6.8800734 15.119927
 50    45         M2          13.00 1.961009 18  8.8800734 17.119927
 75    45         M2          17.30 1.961009 18 13.1800734 21.419927
 25    60         M2          13.00 1.961009 18  8.8800734 17.119927
 50    60         M2          16.95 1.961009 18 12.8300734 21.069927
 75    60         M2          21.20 1.961009 18 17.0800734 25.319927
 25    85         M2          21.50 1.961009 18 17.3800734 25.619927
 50    85         M2          24.30 1.961009 18 20.1800734 28.419927
 75    85         M2          31.65 1.961009 18 27.5300734 35.769927

Confidence level used: 0.95
```



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 que cambios introduciría en el procedimiento de análisis (no necesita hacer nada en R).

Para corroborar la suposición de que las mediciones son independientes habría que graficar el orden en que se colectó un dato contra el residuo correspondiente, de esta manera, si al graficar en el eje horizontal el tiempo y en el eje vertical los residuos, se detecta una tendencia o patrón no aleatorio claramente definido, esto es evidencia de que existe una correlación entre los errores y, por lo tanto, el supuesto de independencia no se cumple.

Para el caso puntual, sería razonable dudar que los factores sean independiente del tiempo, por lo que había que utilizar un modelo estadístico que incluya al resto de los factores como función del tiempo.

Ejercicio 3

a) Opción elegida:

- No se puede calcular con la información provista.

Para calcular el número de réplicas por tratamiento sin tener en cuenta los bloques es necesario calcular la potencia iterando para distintos N y detener el ciclo cuando se alcance un valor cercano al buscado. Para esto es necesario calcular el parámetro de no centralidad λ para la distribución F, del siguiente modo:

$$\lambda = \frac{rD^2}{2MSE}$$

que a su ver requiere conocer la diferencia D , valor no especificado en el enunciado. De todos modos, en este caso podríamos ir variando incrementalmente para graficar la potencia, o sea que podríamos salvar este inconveniente, sin embargo, también es necesario contar con el MSE del experimento sin bloques, y en este caso no es posible suponerlo ni calcularlo con la información provista.

b) Opción elegida:

- 5) 0.00205

Como las comparaciones son con contrastes ortogonales recalculamos la probabilidad de error de tipo I individual haciendo $\alpha_{PC} = (1 - \alpha_F)^{1/N_c}$, donde N_c es la cantidad de contrastes. El cálculo nos arroja:

```
In [9]: alphaF = 0.05
        Nc = 25
        (alpha_pc = 1 - (1- alphaF)^(1/Nc))
```

0.0020496284126208

c) Opción elegida:

- 5) Todas las anteriores.

Cuando el tamaño de observaciones es diferente, el estadístico $F = \frac{MS_{TRT}}{MSE}$ no tiene siempre una distribución F exacta como sí lo es en el caso de que el tamaño de muestras sean iguales. Para el caso de diferente n deben modificarse la suma de cuadrados MS y por ende se dificulta el cálculo de la $E(MS)$. Respecto a la potencia, esta se maximiza cuando las muestras tienen el mismo tamaño.

d) Opción elegida:

- 4) Tukey HSD debería usarse cuando los tratamientos tienen efectos fijos.

Tukey es un método para comparar medias de tratamientos y, según el enunciado el único factor que existe es aleatorio, por lo que no sería razonable utilizarlo.

e) Opción elegida

- 1) $L = -3\mu_{11} - 1\mu_{21} + 1\mu_{31} + 3\mu_{41} - 3\mu_{12} - 1\mu_{22} + 1\mu_{32} + 3\mu_{42}$

Los coeficientes para un contraste ortogonal con un polinomio de tendencia lineal con 4 niveles en la variable cuantitativa son (-3, -1, 1, 3), lo que se corresponde para el caso del enunciado donde desea ver la tendencia para dos niveles del factor F2.

f) Opción elegida:

- (0.8,1.0)

Haciendo el cálculo de la potencia obtenemos un valor de 0.967. A continuación el método de cálculo

```
In [10]: D = 19 # diferencia de medias
         b = 3  # 3 niveles del factor B
         a = 2  # 2 niveles del factor A
         r = 10
         n = a*b*r
         sigma2 = 100
         lambda = r*D^2/(2*sigma2)
         alpha = 0.05
         Q = qf(1-alpha, (a-1)*(b-1), n-a*b)
         (potencia = 1-pf(Q, (a-1)*(b-1), n-a*b, lambda))
```

0.967181638979737