Natalie Julian

Ejercicio 1: Modelo One Way efecto aleatorio

En una compañía productora de tabacos se desea evaluar la calidad de los cigarrillos producidos por operario. Se desea evidenciar si el número de cigarrillos de buena calidad producidos por minuto difere según operario, para ello se se escojen operarios al azar y se realizan experimentos de un minuto, se replica el experimento 10 veces, midiendo la cantidad de cigarrillos producidos que cumplen con el estándar de calidad de la empresa. La compañía busca evaluar si el desempeño por operario difiere, es decir, si se observa una variabilidad considerable en los cigarros producidos por operario. Los datos se encuentran en la base de datos `operario`.

a) Usted como analista, ¿qué modelo le sugeriría considerar a la compañía productora? Defina explícitamente los supuestos.

Respuesta

La variable respuesta es la cantidad de cigarrillos producidos por minuto y que cumplen con un estándar de calidad. El factor es operador, el cual es aleatorio pues existe esta nueva fuente de variabilidad asociada a la selección aleatoria de operadores. Este factor cuenta con 3 niveles.
Además, nos encontramos en un caso balanceado, pues se repiten 10 veces los experimentos por operador. El modelo entonces, puede plantearse como sigue:

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad i = 1, 2, 3 \quad j = 1, ..., 10$$

$$\mu_i \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2_\mu)$$

$$\epsilon_{ij} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$

b) Plantee el modelo propuesto en a). ¿Se podría asumir que la producción de cigarrios de calidad es invariante por operador? ¿por qué?

```
print(operario)
# A tibble: 30 x 2
   respuesta operario
       <dbl>    <dbl>
 1       700        1
 2       850        1
 3       820        1
 4       640        1
 5       920        1
 6       480        2
 7       460        2
 8       500        2
 9       570        2
10       580        2
# ... with 20 more rows

table(operario$operario)  #caso balanceado

 1  2  3
10 10 10

operarios<-as.factor(operario$operario)
cigarros<-operario$respuesta

modelo<-aov(cigarros~operarios)
```

Para analizar la significancia en las diferencias de productividad por operador, realizamos el test de hipótesis:

$$H_0 : \sigma_\mu^2 = 0 \qquad H_1 : \sigma_\mu^2 \neq 0$$

El estadístico asociado es:

$$F_{Trat} = \frac{MCTrat}{MCE} \sim F_{(r-1),r(n-1)}$$

Si $F_{Trat} > F^{1-\alpha}_{(r-1),r(n-1)}$ se rechaza la hipótesis nula de variabilidad nula.

```
anova(modelo)[,-c(4,5)]
           Df Sum Sq Mean Sq
operarios   2 503215  251608
Residuals  27 134880    4996

(Ftrat<-anova(modelo)[1,3]/anova(modelo)[2,3])
[1] 50.36626

(r<-length(levels(operarios))) #Niveles de factor operarios
[1] 3
(n<-unique(table(operarios)))  #Cantidad de observaciones por nivel del factor
[1] 10
alpha<-0.05

Ftrat>qf(p=(1-alpha),df1=(r-1),df2=r*(n-1))
[1] TRUE

#Se rechaza la hipotesis de variabilidad nula de mu con un 95% de confianza
```

No rechazar $\sigma_\mu^2 = 0$ significa que $\sigma_\mu^2 > 0$, i.e no podríamos asumir homogeneidad de medias por nivel del factor: $\mu_i = \mu_j \quad \forall i, j = 1, 2, 3$. Existen diferencias significativas inducidas por operario.

c) La compañía quiere establecer un rango de valores para el promedio de cigarros de calidad que se producen en un minuto, para así medir la productividad media de cigarros de calidad por minuto. ¿Entre qué rango podría establecer que varía la cantidad de cigarros de calidad producidos por minuto?

Respuesta
Sabemos que $\mu_i \sim N(\mu, \sigma_\mu^2)$. Si queremos conocer un rango de valores para la cantidad media de cigarros (global, no por operador), debemos realizar inferencias respecto a μ. Calculamos el intervalo de confianza como sigue:

$$IC(\mu) = \left(\bar{Y}_{\cdot\cdot} \pm t^{1-\alpha/2}_{(r-1)} \sqrt{\frac{MCTrat}{nr}} \right)$$

```
(Ybar<-mean(cigarros))
[1] 646.4333
(MCtrat<-anova(modelo)[1,3])
[1] 251607.6

(IC<-c(Ybar+c(-1,1)*qt(1-alpha/2,r-1)*sqrt(MCtrat/(n*r))))
[1]  252.3958 1040.4709
```

Es decir, con un 95% de confianza, se esperaría un promedio mínimo de 252 cigarros de calidad por minuto, mientras que como máximo, se esperaría un promedio de 1040 cigarros de calidad.

d) Suponga que se tomó otra muestra aleatoria de operadores y observó la mayor homogeneidad posible en términos de producción de cigarros de calidad, ¿qué porcentaje de variabilidad explicada por el factor operario esperaría en ese caso en base a los datos? Comente.

Respuesta

Sabemos que el total de variabilidad es $\sigma^2 + \sigma_\mu^2$ por lo tanto, la tasa o porcentaje de variabilidad explicada por el factor operario respecto al total es $\frac{\sigma_\mu^2}{\sigma^2+\sigma_\mu^2}$ y sabemos que un intervalo de confianza es:

$$\frac{L}{1+L} < \frac{\sigma_\mu^2}{\sigma_\mu^2+\sigma^2} < \frac{U}{1+U}$$

Con $L = \frac{1}{n}[\frac{MCTrat}{MCE}\frac{1}{F^{1-\frac{\alpha}{2}}_{r-1,r(n-1)}} - 1]$ y $U = \frac{1}{n}[\frac{MCTrat}{MCE}\frac{1}{F^{\frac{\alpha}{2}}_{r-1,r(n-1)}} - 1]$.

```
(MCE<-anova(modelo)[2,3])
[1] 4995.559

L=(1/n)*(MCtrat/(MCE*qf(p=(1-alpha/2),df1=(r-1),df2=r*(n-1)))-1)
U=(1/n)*(MCtrat/(MCE*qf(p=(alpha/2),df1=(r-1),df2=r*(n-1)))-1)

(InC<-c(L/(L+1),U/(U+1)))
[1] 0.5209115 0.9949912
```

En el peor de los casos, se lograría explicar el 52% de variabilidad con el factor operario y en el mejor de los casos, se podría explicar prácticamente toda la variabilidad total.

Ejercicio 2: Modelo Two Way efectos aleatorios Fuente base de datos Complemento

La base de dados `suicide` contiene información sobre la cantidad de suicidios registrados en el año 2015. Los investigadores buscan determinar si por continente se observan diferencias en la cantidad de suicidios. Para ello, por continente se eligieron 4 países aleatorios como representantes. Además, otro tema de interés es determinar si existen diferencias en la cantidad de suicidios por edad, puesto que se cree que hay rangos etarios más predispuestos a suicidarse por factores sociales, laborales y emocionales. Para este análisis, considere que para cada rango etario se tomó una muestra aleatoria (del mismo tamaño para cada rango etario) de los fallecidos el 2015 y se realizó un conteo de quiénes tuvieron como causa de muerte asociada al sucidio.

a) Realice algunas inferencias descriptivas por continente. ¿Observa diferencias? Comente.

Respuesta

```
print(suicide)
# A tibble: 96 x 4
   Pais      Continente Etario      Suicidios
   <chr>     <chr>      <chr>           <dbl>
 1 Armenia   Asia       15-24 years         5
 2 Armenia   Asia       25-34 years        12
 3 Armenia   Asia       35-54 years        19
 4 Armenia   Asia       5-14 years          0
 5 Armenia   Asia       55-74 years        21
```

```
 6 Armenia Asia        75+ years          17
 7 Austria Europa      15-24 years        80
 8 Austria Europa      25-34 years        99
 9 Austria Europa      35-54 years       392
10 Austria Europa      5-14 years          2
# ... with 86 more rows
```

```
table(suicide$Continente,suicide$Etario)  #Balanceado n=4 observaciones por celda
```

```
                15-24 years 25-34 years 35-54 years 5-14 years 55-74 years
  Asia                    4           4           4          4           4
  Europa                  4           4           4          4           4
  Latinoamdrica           4           4           4          4           4
  Norte America           4           4           4          4           4

                75+ years
  Asia                  4
  Europa                4
  Latinoamerica         4
  Norte America         4
```

```
library(dplyr)
```

```
suicide%>%group_by(Continente)%>%summarise(Media=mean(Suicidios), Minimo=min(Suicidios),
Maximo=max(Suicidios), Mediana=median(Suicidios), n=n())
# A tibble: 4 x 6
  Continente     Media Minimo Maximo Mediana     n
  <chr>          <dbl>  <dbl>  <dbl>   <dbl> <int>
1 Asia            52.6      0    335    18       24
2 Europa         380.       2   1921   204.      24
3 Latinoamerica  681.       8   4148   325       24
4 Norte America 2101.       0  15687    88.5     24
```

En términos de media de suicidios, se observan diferencias relevantes por Continente. Sin embargo, al mirar la mediana ocurre algo interesante respecto al Norteamérica, pues la media de suicidios da bastante alta pero la mediana no. Seguramente existe alguna observación dentro del continente que infla el promedio. También es importante mencionar que estos valores se verán muy influenciados por los países que se eligieron al azar.

b) ¿Qué modelo le sugeriría estudiar a los investigadores?. Sea explícito con los supuestos.

Respuesta
El factor Continente es un factor aleatorio pues en él incluye otras fuentes de variabilidad, dado que para representar cada continente se eligen al azar 4 países). Es aleatorio pues se eligieron países arbitrarios para realizar el estudio.
El factor Etario es un factor aleatorio pues para recopilar la información en cada rango etario se realizaron muestras aleatorias para realizar el conteo de fallecidos por suicidio.
El modelo propuesto sería:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \text{con} \quad i = 1, ..., a = 4, \quad j = 1, ..., b = 6 \ \text{ y } \ k = 1, ..., 4$$

$$\alpha_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_\alpha)$$

$$\beta_j \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_\beta)$$

$$(\alpha\beta)_{ij} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_{\alpha\beta})$$

$$\epsilon_{ijk} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$

c) En base al modelo propuesto, estime la variabilidad asociada a cada componente. Comente.

Respuesta
Es de interés analizar la variabilidad de las componentes del modelo pues es posible establecer

que:

$$Var(Y_{ijk}) = \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_{\alpha\beta} + \sigma^2$$

Así, también es posible observar cuánta variabilidad del total es explicada por cada componente. Observando la tabla ANOVA:

```
anova(model)[,c(1,3)]
                   Df   Mean Sq
Continente          3  19535116
Edad                5   5340776
Continente:Edad    15   1719294
Residuals          72   5012338
```

Y utilizamos como estimadores de σ^2, σ^2_α, σ^2_β y σ^2_α los siguientes:

$$\hat{\sigma}^2 = MCE = 5012338$$
$$\hat{\sigma}^2_\alpha = \frac{MCA - MCAB}{nb} = \frac{19535116 - 1719294}{4 \cdot 6} = 742325.9$$
$$\hat{\sigma}^2_\beta = \frac{MCB - MCAB}{na} = \frac{5340776 - 1719294}{4 \cdot 4} = 3621482$$
$$\hat{\sigma}^2_{\alpha\beta} = \frac{MCAB - MCE}{n} = \frac{1719294 - 5012338}{4} = -823261$$

Interpretación

- $\hat{\sigma}^2 = 5012338$ variabilidad asociada a los valores muestreados de cantidad de suicidios
- $\hat{\sigma}^2_\alpha = 742325.9$ es la variación de la cantidad de suicidios asociada a la información muestreada por continente
- $\hat{\sigma}^2_\beta = 3621482$ es la variación asociada a los distintos muestreos que se realizaron por rango etario y también otras predisposiciones psicológicas, emocionales o sociales respecto al sucidio por rango etario
- $\hat{\sigma}^2_{\alpha\beta} = -823261$ Cuando la estimación de una componente de variabilidad es negativa, esto es un gran indicio de que el modelo utilizado no es el correcto. En este caso, observe que la componente estimada negativa corresponde a la variabilidad de la interacción, lo que pudiera significar que incorporar la interacción en este modelo no logra aportar variabilidad al modelo. Es necesario realizar un test de significancia de los componentes del modelo.

d) ¿Son los factores Continente y Edad significativos a la hora de modelar la cantidad de suicidios? Utilice 90% de confianza. ¿Usted propondría otro modelo? Comente.

Respuesta

En el caso de factores aleatorios se tienen los siguientes tests de interés:
$H_0 : \sigma^2_\alpha = 0$, $H_1 : \sigma^2_\alpha > 0$:

$$F_\alpha = \frac{MCA}{MCAB} \sim F^{1-\alpha}_{(a-1),\ (a-1)(b-1)}$$

```
 anova(model)[,c(1,3)]
                Df  Mean Sq
Continente       3 19535116
Edad             5  5340776
Continente:Edad 15  1719294
Residuals       72  5012338

(Fcontinente=anova(model)[1,3]/anova(model)[3,3])
[1] 11.36229

(a<-length(levels(Continente)))
[1] 4
(b<-length(levels(Edad)))
[1] 6

alpha=0.1

Fcontinente>qf(p=(1-alpha),df1=(a-1),df2=(a-1)*(b-1))
[1] TRUE


#Se rechaza la hipotesis de variabilidad nula de sigma_alpha con un 90% de confianza
```

$H_0 : \sigma^2_\beta = 0$, $H_1 : \sigma^2_\beta > 0$:

$$F_\beta = \frac{MCB}{MCAB} \sim F^{1-\alpha}_{(b-1),\ (a-1)(b-1)}$$

```
(Fedad=anova(model)[2,3]/anova(model)[3,3])
[1] 3.106377

Fedad>qf(p=(1-alpha),df1=(b-1),df2=(a-1)*(b-1))
[1] TRUE


#Se rechaza la hipotesis de variabilidad nula de sigma_beta con un 90% de confianza
```

Y por último, la interacción $H_0 : \sigma^2_{\alpha\beta} = 0$, $H_1 : \sigma^2_{\alpha\beta} > 0$:

$$F_{\alpha\beta} = \frac{MCAB}{MCE} \sim F^{1-\alpha}_{(a-1)(b-1),ab(n-1)}$$

```
(n<-unique(table(Continente,Edad)))
[1] 4

(Finter=anova(model)[3,3]/anova(model)[4,3])
[1] 0.3430124

Finter>qf(p=(1-alpha),df1=(a-1)*(b-1),df2=a*b*(n-1))
[1] FALSE

#La interaccion no resulta significativa al 90% de confianza
```

La interacción al no ser significativa, se sugiere utilizar un modelo aditivo.

e) Plantee el modelo propuesto en c), ¿cuáles son los rangos etarios más propensos al suicidio?

Respuesta

```
aditivo<-aov(Suicidios~Continente+Edad)

coef(aditivo)
      (Intercept)   ContinenteEuropa ContinenteLatinoamerica
        -52.71875          327.70833              628.12500
ContinenteNorte America    Edad25-34 years    Edad35-54 years
```

```
          2048.54167            126.81250            983.25000
       Edad5-14 years       Edad55-74 years       Edad75+ years
          -641.62500            486.56250           -322.93750
```

```
levels(Continente)
[1] "Asia"          "Europa"        "Latinoamerica" "Norte America"

levels(Edad)
[1] "15-24 years" "25-34 years" "35-54 years" "5-14 years"  "55-74 years"
[6] "75+ years
```

Los rangos etarios con mayor incremento promedio respecto al intercepto son entre 35 y 54 años y 55 a 64. Pareciera que efectivamente al cruzar la barrera de adulto joven a adulto, la cantidad promedio de suicidios se ve considerablemente incrementada pero después de los 75 años el efecto es negativo, por lo que se espera una disminución respecto al intercepto. Respecto a continentes es relevante analizar el caso de Norte América pues tiene un incremento de 2048 en el número promedio de suicidios, bastante alto respecto a los demás continentes.

f) ¿Cuál sería la cantidad de suicidios esperada en el año 2016 para un país en el continente Europeo en el rango etario 35-54 años?

Respuesta

```
new<-rbind(c("Europa","35-54 years"))

colnames(new)<-c("Continente","Edad")

new

predict(aditivo,newdata = data.frame(new))
       1
1258.24
```

Lo que es equivalente a: `Intercepto+ContinenteEuropa+Edad 35-54 years` $= -52.71875 + 327.70833 + 983.25 = 1258.19$.

Ejercicio 3: Determinar naturaleza de los efectos

En los siguientes enunciados, determine si los factores involucrados son de tipo fijo o de tipo aleatorio.

- Un investigador de mercado está interesado en evaluar el efecto de los honorarios (alto, medio, bajo), forma de trabajo (solo desde casa, algún trabajo en terreno) y el tipo de supervisión (supervisores locales, supervisores itinerantes) en la calidad del trabajo realizado por agencias de investigación del trabajo. La calidad del trabajo fue medida con un índice de diversos factores de calidad. Cuatro agencias fueron elegidas por combinación de tratamientos.

  Respuesta

  Factor A: Honorarios (efecto fijo de 3 niveles)
  Factor B: Forma de trabajo (factor fijo de 2 niveles)
  Factor C: Tipo de supervisión (factor fijo de 2 niveles)

- Los agrónomos se esfuerzan por desarrollar nuevas variedades de maíz que sean más resistentes a las enfermedades, menos sensibles as las condiciones climáticas y más productivas. Se teme que estas variedades se planten en cualquier sitio y sin cuidado alguno. Se propone el siguiente experimento: Cuatro nuevas variedades de maíz escodidas al azar son

probados en tres subsuelos aleatorios. Se mide la cantidad de cosechas obtenidas.

Respuesta

Factor A: Variedad de maíz (efecto aleatorio de 4 niveles)
Factor B: Subsuelos (factor aleatorio de 3 niveles)

- Un grupo de investigación quiere estudiar la eficacia de de tres tipos de programas de formación (Gestión de conflictos, la psicología y la negociación) destinados a agentes del FBI y mejorar su manejo con jóvenes relacionados con ataques terroristas. En todo el país, las oficinas del FBI han llevado a cabo uno o más de estos programas de formación. El grupo de investigación decide seleccionar aleatoriamente 4 oficinas donde se hayan aplicado estos programas de formación y luego a cada persona de las 4 oficinas seleccionadas se le registró el programa de formación que realizó y se le dio una prueba para evaluar sus habilidades en el trato de los jóvenes en riesgo de involucrarse con el terrorismo.

  Respuesta

  Factor A: programas de formación (efecto fijo de 3 niveles)
  Factor B: oficinas (factor aleatorio de 4 niveles)