EP05_2024

B García, JL Jara

2024-04-02, 24-04-14

Enunciado

En una emocionante competencia de cubos Rubik, participantes de Chile, Argentina, Colombia, Uruguay, Perú y Ecuador demostraron su destreza en resolver tres tipos de cubos: 2x2x2, 3x3x3 y Megaminx.

Después del torneo, un grupo de investigadores de la Asociación Mundial del Cubo, interesado en los tiempos que hicieron las jugadoras y los jugadores en la competencia, decidieron estudiar si el país y el tipo de cubo usado en cada prueba tienen influencia en los segundos que se tardan en resolverlos. Para ello usaron una muestra aleatoria de los datos de la competencia, en la cual participaron más de 2.000 personas, con las siguientes variables:

Variable	Descripción
id	Identificador único de cada participante.
pais	País de procedencia de la persona (Argentina, Chile, Colombia, Ecuador, Perú, Uruguay).
tipo	Tipo de cubo usado en la prueba (2x2x2, 3x3x3 y Megaminx).
tiempo	Tiempo necesitado por cada participante en resolver el cubo de la prueba (en segundos).

¿Existen diferencias en el tiempo de resolución de cubos 3x3x3 entre participantes de Chile, Uruguay y Colombia?

En esta pregunta se pide inferir acerca de las medias de una variable numérica (tiempo) medidas en grupos independientes formados por un factor con tres niveles (país). Luego se requiere usar un **procedimiento ANOVA para muestras independeintes**.

Las hipótesis serían:

H$_0$: los tiempos promedio requeridos para resolver un cubo de tipo 3x3x3 por las y los participantes de Chile (μ_{CL}), Uruguay (μ_{UY}) y Colombia (μ_{CO}) son iguales; es decir ($\mu_{CL} = \mu_{UY} = \mu_{CO}$).
H$_A$: el tiempo requerido resolver un cubo de tipo 3x3x3 es diferente para las y los participantes de al menos uno de estos países ($\exists \mathrm{i}, \mathrm{j} \in \{CL, UY, CO\} : \mu_\mathrm{i} \neq \mu_\mathrm{j}$).

Obtengamos la muestra de datos que debemos utilizar.

```r
# Lectura de datos
src_dir <- "~/Downloads"
src_basename <- "EP05 Datos.csv"
src_file <- file.path(src_dir, src_basename)
datos <- read.csv2(src_file, stringsAsFactors = TRUE)
datos[["id"]] <- factor(datos[["id"]])
```

```r
library(dplyr)

# Seleccionamos datos de interés
datos_largos <- datos %>%
  filter(tipo == "3x3x3") %>%
  filter(pais == "Chile" | pais == "Uruguay" | pais == "Colombia") %>%
  select(id, pais, tiempo) %>%
  droplevels()
datos_largos[["id"]] <- factor(datos_largos[["id"]])
head(datos_largos)
```

```
   id     pais tiempo
1  31    Chile  15.36
2  40  Uruguay  16.84
3  45 Colombia  16.48
4  76  Uruguay  16.64
5  94  Uruguay  16.14
6 142  Uruguay  16.43
```

Ahora verifiquemos las condiciones para asegurar que podemos aplicar el procedimiento con validez.

La variable dependiente corresponde a tiempo, que sabemos tiene escala de razón, y por lo tanto una escala continua de intervalos iguales, por ser una medida física.

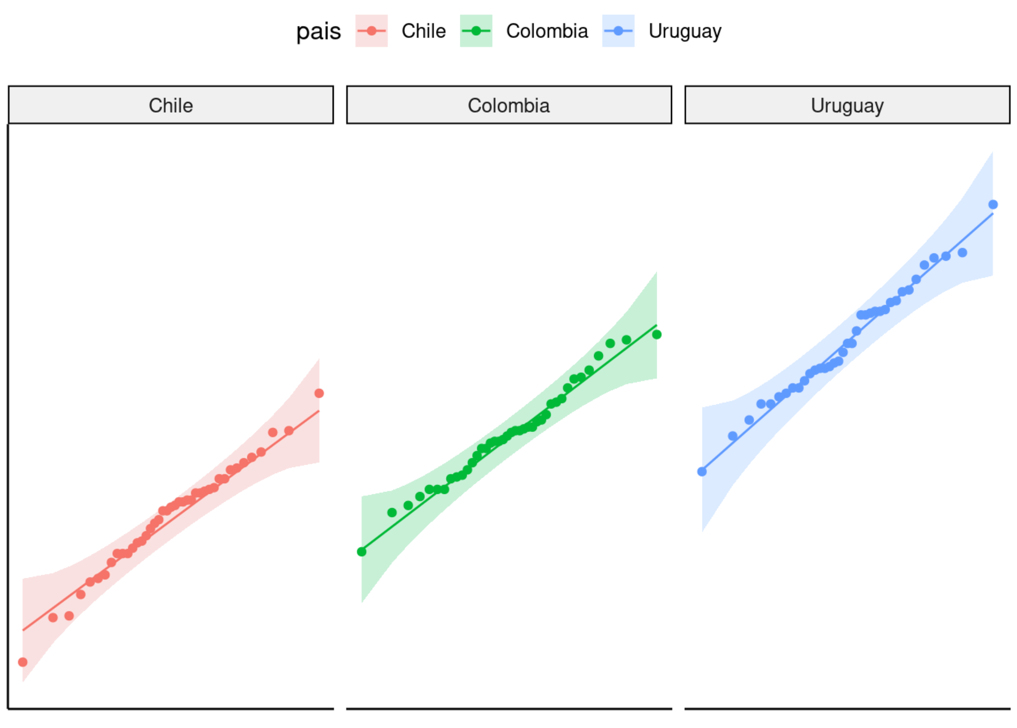
Por otro lado, el enunciado indica que las observaciones son independientes entre sí, pues provienen de personas diferentes.

Revisemos ahora la condición de normalidad por medio de un gráfico Q-Q.

```r
library(ggpubr)

g <- ggqqplot(datos_largos,
              x = "tiempo",
              y = "pais",
              color = "pais")

g <- g + facet_wrap(~ pais)
g <- g + rremove("x.ticks") + rremove("x.text")
g <- g + rremove("y.ticks") + rremove("y.text")
g <- g + rremove("axis.title")
print(g)
```



El gráfico generado muestra que la distribución de los datos de cada una de las muestras puede considerarse cercana a la normal pues, si bien no forman una recta, todos se encuentran dentro de la región aceptable del gráfico Q-Q y no se observan comportamientos extraños ni aleatorios.

De forma alternativa, podemos usar pruebas de normalidad para hacer esta verificación. Por el tamaño de las muestras disponibles aquí, sería apropiado aplicar la prueba de Shapiro-Wilk como muestra el siguiente código.

```r
# Realizar el test de Shapiro-test para cada país
tests_normalidad <- by(datos_largos[["tiempo"]],
                       datos_largos[["pais"]],
                       shapiro.test)
print(tests_normalidad)
```

```
datos_largos[["pais"]]: Chile

    Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.98584, p-value = 0.8816

------------------------------------------------------------
datos_largos[["pais"]]: Colombia

    Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.98293, p-value = 0.7961

------------------------------------------------------------
datos_largos[["pais"]]: Uruguay

    Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.98744, p-value = 0.9443
```

Vemos que estas pruebas, de forma consistente con los gráficos Q-Q, descartan que debamos sospechar que alguna de estas muestras provenga de una población que no siga una distribución normal.

En cuanto a la condición de homocedasticidad, se posterga su discusión hasta ver el resultado de la prueba de Levene efectuada por `ezAnova()`.

Puesto que hasta ahora no tenemos motivos que indiquen que los datos podrían incumplir alguna de las condiciones, podemos proceder con el procedimiento ANOVA para muestras independientes considerando un nivel de significación de 0,05.

```r
library(ez)

alfa <- 0.05

omnibus <- ezANOVA(
  data = datos_largos,
  dv = tiempo,
  between = pais,
  wid = id,
  return_aov = TRUE
)
```

```
Warning: Data is unbalanced (unequal N per group). Make sure you specified a
well-considered value for the type argument to ezANOVA().
```

```
Coefficient covariances computed by hccm()
```

Notemos los mensajes que nos presenta esta función. La primera nos advierte usar un “buen” valor para el argumento `type` debido a que las muestras tienen tamaños diferentes. En este caso no hemos cambiado el valor por omisión (`type=2`) que, como se dijo en el apunte, funciona para la mayoría de los casos (al menos al trabajar con un solo factor).

El segundo es menos importante y solamente nos informa de la función que está usando internamente para calcular la matriz de covarianzas.

Veamos el resultado del procedimiento por pantalla.

```r
print(omnibus)
```

```
$ANOVA
  Effect DFn DFd        F            p p<.05       ges
1   pais   2 115 103.5329 1.920038e-26     * 0.6429301

$`Levene's Test for Homogeneity of Variance`
  DFn DFd        SSn      SSd         F         p p<.05
1   2 115 0.05277397 4.024616 0.7539857 0.4727989

$aov
Call:
   aov(formula = formula(aov_formula), data = data)

Terms:
                    pais Residuals
Sum of Squares  19.14243  10.63130
Deg. of Freedom        2       115

Residual standard error: 0.3040495
Estimated effects may be unbalanced
```

```r
print(summary(omnibus[["aov"]]))
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
pais          2  19.14   9.571   103.5 <2e-16 ***
Residuals   115  10.63   0.092
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Podemos ver que la prueba de homocedasticidad de Levene resulta no significativa con 95% de confianza ($p = 0{,}473$), por lo que se falla en rechazar la hipótesis nula de esta prueba, y debemos concluir que **no hay suficiente evidencia para descartar** que se cumple la condición de homocedasticidad en estos datos.

Interpretemos este resultado ómnibus.

El procedimiento ANOVA resultó significativo ($p < 0,001$). En consecuencia, con 95% de confianza, rechazamos la hipótesis nula en favor de la hipótesis alternativa y concluimos que las y los participantes de al menos un país (Chile, Uruguay o Colombia) resolvieron en una cantidad de tiempo diferente los cubos de 3x3x3.

Puesto que el procedimiento ómnibus encuentra diferencias estadísticamente significativas, es necesario realizar un procedimiento post-hoc. Puesto que no requerimos hacer contrastes adicionales, usaremos la prueba HSD de Tukey, más poderosa que los factores de corrección no paramétricos (como Bonferroni y Holm), ya que no se ha descartado que los datos siguen distribuciones normales y con igual varianza.

```r
post_hoc <- TukeyHSD(omnibus[["aov"]], which = "pais", ordered = TRUE,
                     conf.level = 1 - alfa)
print(post_hoc)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = formula(aov_formula), data = data)

$pais
                      diff       lwr       upr p adj
Colombia-Chile   0.4646037 0.3041585 0.6250488     0
Uruguay-Chile    0.9920699 0.8283654 1.1557743     0
Uruguay-Colombia 0.5274662 0.3627939 0.6921385     0
```

Podemos ver una representación gráfica del efecto encontrado en este análisis (producido en la variable dependiente `tiempo` por la variable independiente `país`).

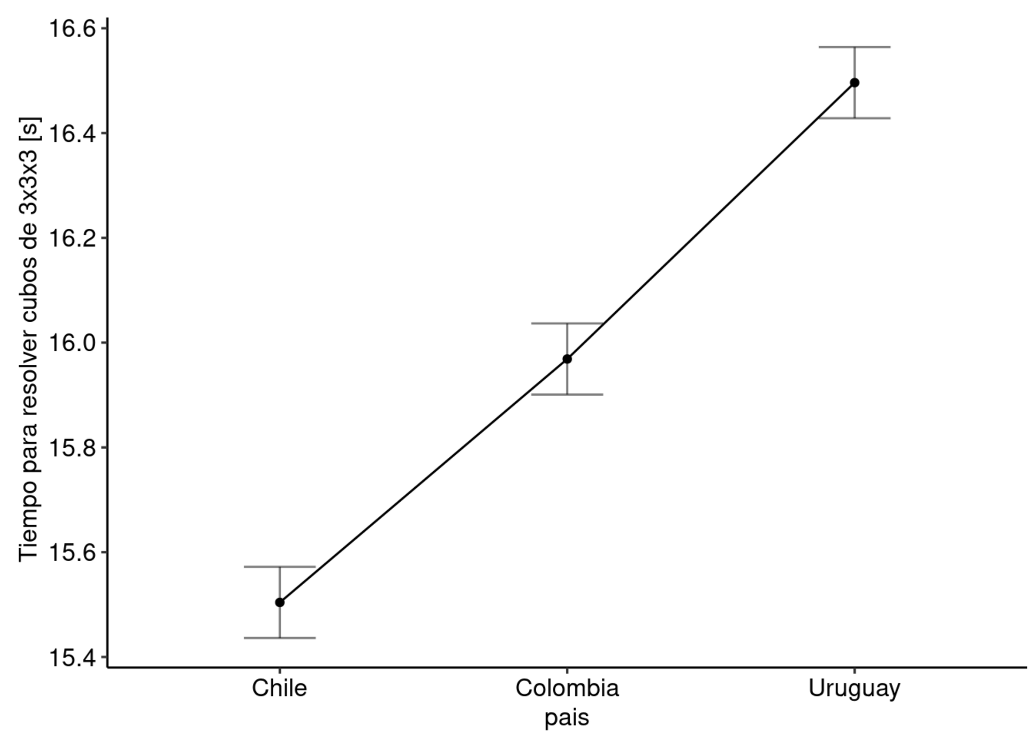
```r
efecto <- ezPlot(data = datos_largos, dv = tiempo, wid = id,
                 between = pais, x = pais,
                 y_lab = " Tiempo para resolver cubos de 3x3x3 [s]")
```

```
Warning: Data is unbalanced (unequal N per group). Make sure you specified a
well-considered value for the type argument to ezANOVA().
```

```
Coefficient covariances computed by hccm()
```

```
Warning in ezStats(data = data, dv = dv, wid = wid, within = within,
within_full = within_full, : Unbalanced groups. Mean N will be used in
computation of FLSD
```

```r
efecto <- efecto + theme_pubr()
print(efecto)
```



Vemos que el gráfico del efecto coincide con los resultados de la prueba post-hoc, mostrando con claridad diferencias entre participantes de los países estudiados. Concluyamos con todos estos resultados.

El análisis post-hoc indica que participantes provenientes de Uruguay son más lentos que quienes vienen de Colombia al resolver un cubo de 3x3x3 (entre 0,363 y 0,692 [s], $p < 0,001$), que a su vez son más lentos que participantes de Chile (entre 0,304 y 0,625 [s], $p < 0,001$).