12.1ANOVA

Paula Vargas Pellicer

17/03/2022

Otro ejemplo

Para profundizar con lo que vimos en la clase pasada, trabajaremos con un ejemplo a mano.

Tenemos un experimento en el que los rendimientos de los cultivos por unidad de área se midieron de 10 campos seleccionados al azar en tres tipos de suelo. Todos los campos fueron sembrados con la misma variedad de semilla y provistos de los mismos fertilizantes e insumos para el control de plagas. La pregunta es, si el tipo de suelo afecta significativamente el rendimiento del cultivo y, de ser así, en qué medida.

```
arena<- c(6,10,8,6,14,17,9,11,7,11)
arcilla<- c(17,15,3,11,14,12,12,8,10,13)
roca<- c(13,16,9,12,15,16,17,13,18,14)
resultados<- data.frame(arena,arcilla,roca)

#calcular las medias de cada tipo de suelo
sapply(list(arena,arcilla,roca),mean)

#pongamos los datos en solo vector
res.largos<- stack(resultados)
#cambiar nombre columnas
names(res.largos)<- c("produccion","suelo")
```

Checar si hay varianzas constantes

```
tapply(res.largos$produccion,res.largos$suelo,var)
```

Las varianzas difieren en más de un factor de 2. Pero, ¿es esto significativo? Probamos la heteroscedasticidad usando la prueba de homogeneidad de varianzas de Fligner-Killeen

```
fligner.test(res.largos$produccion~res.largos$suelo)
```

Debido a que la variable explicativa es categórica (tres niveles de tipo de suelo), la inspección inicial de datos implica un diagrama de caja contra el suelo

```
plot(res.largos$produccion~res.largos$suelo,col="green")
```

Podemos ahora hacer un ANOVA

```
 Modelo<-aov(res.largos$produccion~res.largos$suelo)
summary(Modelo)
```

Aquí puedes ver que la fila de error está etiquetada como Residuales. En las columnas siguientes se ven los grados de libertad para tratamiento y error (2 y 27), el tratamiento y sumas de cuadrados de error (99.2 y 315.5), el cuadrado medio del tratamiento de 49.6, la varianza de error s2 = 11.685, la relación entre F y el valor de p (etiquetado como Pr(>F)). El asterisco junto al valor p indica que la diferencia entre las medias del suelo es significativa al 5% (pero no al 1%, lo que hubiera merecido dos asteriscos).

Lo siguiente que haríamos es verificar los supuestos del modelo aov. Esto se hace usando las graficas siguientes:

```
plot(aov(res.largos$produccion~res.largos$suelo))
```

La primera gráfica comprueba la suposición más importante (constancia de la varianza); No debería haber patrón en los residuos contra los valores ajustados (las tres medias de tratamiento). La segunda gráfica prueba la suposición de normalidad de los errores: debe haber una relación de línea recta entre nuestros residuos estandarizados y los cuantiles teóricos derivados de una distribución normal. Los residuos se comportan bien (tercera grafica) y no hay valores muy influyentes que puedan estar distorsionando el parámetro de estimaciones (ultima grafica).

Efectos de tamano

La mejor manera de ver gráficamente los efectos de tamano es usar `plot.design` (que toma una fórmula en lugar de un objeto modelo), pero nuestro modelo actual con solo un factor es quizás demasiado simple para obtener el valor completo de esto (parcela.diseño(rendimiento~suelo)). Para ver los tamaños del efecto en forma tabular, usa `model.tables` (quetoma un objeto modelo como su argumento) así:

```
model.tables(Modelo, se=T)
```

Los efectos se muestran como desviaciones de la media general: el suelo arena tiene un rendimiento medio que está 2,0 por debajo del promedio general, y el suelo roca tiene un promedio que está 2.4 por encima del promedio general. El error estándar de los efectos es 1,081 en una réplica de n = 10 (este es el error estándar de una media).

Experimentos factoriales

El conjunto de datos Diet.csv contiene información sobre 78 personas que realizaron una de tres dietas. Hay información de fondo como la edad, el género (Mujer = 0, Hombre = 1) y altura. El objetivo del estudio era ver qué dieta era mejor para perder peso, pero también se pensó que las mejores dietas para hombres y mujeres pueden ser diferentes, por lo que las variables independientes son la dieta y el sexo.

```
dieta<- read.csv("~/Downloads/Diet_R.csv", header=T, sep=",")
```

Calcula el peso perdido por persona (diferencia de peso antes y después de la dieta) y sumar la variable al conjunto de datos

```
dieta$weightlost<-dieta$pre.weight-dieta$weight6weeks
```

Hagamos ANOVA

```
anova<-aov(weightlost~gender*Diet, data=dieta)
```

Revisemos los supuestos

```
fligner.test<- anova$residuals

#Produce un histograma
hist(anova$residuals,main="Histograma de
residuos",xlab="Residuales")
```

Puedes uar tambien la prueba de levene en el paquete `car`

```
library(car)

## Loading required package: carData

#leveneTest(weight6weeks~gender*Diet,data=dieta)
```

Los valores p no son significatibos por lo que podemos asumir equidad en las varianzas

Incluso podemos verlo con las graficas

```
plot(anova)
```

Interpreta cada grafica

Ahora si podemos ver los resultados

```
summary(anova)
```

Visualicemos los datos

```
library(ggplot2)
# Función para calcular la media y la desviación estándar para cada grupo
# datos: un marco de datos
# varname: el nombre de una columna que contiene la variable para ser resumid
o
# groupnames : vector de nombres de columna que se utilizará como variables d
e agrupación

data_summary <- function(data, varname, groupnames){
  require(plyr)
  summary_func <- function(x, col){
    c(mean = mean(x[[col]], na.rm=TRUE),
      sd = sd(x[[col]], na.rm=TRUE))
  }
  data_sum<-ddply(data, groupnames, .fun=summary_func,
                  varname)
  data_sum <- rename(data_sum, c("mean" = varname))
```

```
 return(data_sum)
}

# Crear la base de datos para graficar
dieta2<- data_summary(dieta, varname="weight6weeks",
                      groupnames=c("Diet", "gender"))

## Loading required package: plyr

# Grafica
ggplot(na.omit(dieta2), aes(x=Diet, y=weight6weeks, group=gender, color=gende
r)) +
  geom_line() +
  geom_point()+
  geom_errorbar(aes(ymin=weight6weeks-sd, ymax=weight6weeks+sd), width=.2,
                position=position_dodge(0.05))+
  labs(x="Dieta", y = "Peso")+
   theme_classic() +
   scale_color_manual(values=c('#999999','#E69F00'))+
  scale_fill_discrete(name = "Genero", labels = c("Hombres", "Mujeres"))

## Error: Continuous value supplied to discrete scale
```

Pruebas Post Hoc

El comando TukeyHSD(anova) producirá pruebas post hoc para los efectos principales y las interacciones. Solo debes interpretar las pruebas post hoc para los factores significativos del ANOVA.

```
TukeyHSD(anova)

## Warning in replications(paste("~", xx), data = mf): non-factors ignored: g
ender

## Warning in replications(paste("~", xx), data = mf): non-factors ignored: D
iet

## Warning in replications(paste("~", xx), data = mf): non-factors ignored: g
ender,
## Diet

## Error in TukeyHSD.aov(anova): no factors in the fitted model
```

La interacción fue significativa, entonces los efectos principales no son interpretado aquí; pero si tus datos no tienen una Interacción significativa, interprétalos de la misma manera que las pruebas post hoc en el recurso ANOVA unidireccional.

El siguiente resultado para las pruebas de interacciones post hoc se ha ajustado en Excel para que sea más fácil de leer.

	Groups being compared	Difference in means	Lower confidence interval	Upper confidence interval	Adjusted p-value
		diff	lwr	upr	p adj
comparisons with females on diet 1 (0:1)	1:1-0:1	0.60	-2.21	3.41	0.9888
	0:2-0:1	-0.44	-3.01	2.13	0.9958
	1:2-0:1	1.06	-1.68	3.80	0.8657
	0:3-0:1	2.83	0.31	5.35	0.0191
	1:3-0:1	1.18	-1.49	3.86	0.7855
comparisons with males on diet 1 (1:1)	0:2-1:1	-1.04	-3.86	1.77	0.8852
	1:2-1:1	0.46	-2.51	3.43	0.9975
	0:3-1:1	2.23	-0.54	5.00	0.1863
	1:3-1:1	0.58	-2.33	3.49	0.9916
Comparisons with females on diet 2 (0:2)	1:2-0:2	1.50	-1.24	4.24	0.5963
	0:3-0:2	3.27	0.75	5.80	0.0040
	1:3-0:2	1.63	-1.05	4.30	0.4833
Comparisons with males/ diet 2	0:3-1:2	1.77	-0.93	4.47	0.3965
	1:3-1:2	0.12	-2.71	2.96	1.0000
Male 3: Female 3	1:3-0:3	-1.65	-4.28	0.98	0.4514

Hay 6 combinaciones de dieta y género. La prueba post hoc de interacciones compara cada par de combinaciones. Esto muestra que las unicas diferencias significativas son para las mujeres y están entre las dietas 1 y 3 (p=0,0191) y las dietas 2 y 3 (p=0,004). Las mujeres con la dieta 3 perdieron en promedio 2.83kg más que las de la dieta 1 y 3,27 kg más que las de dieta 2.