Bioestadística con R

Roberto Bustillos, MVZ, M.Sc - Basado en un material de D. Conesa y X. Barber

Universidad Central del Ecuador

Facultad de Medicina Veterinaria y Zootecnia

Maestría en Epidemiología y Salud Pública Veterinaria, 2016-2018

Contenido

Análisis de la Varianza.

Motivación: Comparación entre grupos

Varios estudios en M. Veterinaria se basan en la idea de comparar la media de varios grupos: unos que han recibido tratamientos y otro que no (control). Hay que observar que tenemos una situación similar a la comparativa de la media de dos grupos (resolvíamos con un test t), salvo que ahora tenemos más de dos grupos.

Examples

Se pretende valorar la producción de leche de 5 vacas de alta producción, en concreto si el valor medio es similar en los cinco animales. Los datos siguientes son una parte de los obtenidos:

	Vaca1	Vaca2	Vaca3	Vaca4	Vaca5
	35	40	45	41	50

Análisis de la Varianza: ANOVA.

- Una variable respuesta se puede modelizar en función de un conjunto de variables explicativas contínuas o discretas.
- Las variables explicativas son categóricas y les llamaremos factores.
- Se podría considerar como un caso particular de regresión. Pero vamos a focalizar la forma de analizarlo en la descomposición de la suma de cuadrados de las variabilidades entre y dentro de los diferentes grupos analizados.
- El objetivo será determinar si las condiciones que marcan los factores que estamos analizando tienen efecto en la variable respuesta.

ANOVA de un factor de clasificación.

- La situación más básica es la que sólo tenemos una variable explicativa (o factor de clasificación).
- Nuestro objetivo será pues valorar si existen diferencias en los valores de la variable respuesta en las diferentes categorías del factor de clasificación.
- El factor puede ser de:
  - Efectos fijos: nos interesa el efecto de cada tratamiento en concreto en comparación con los otros.
  - Efectos aleatorios: nos interesa la variabilidad entre grupos.
- El modelo (o el diseño) es equilibrado cuando el número de observaciones por nivel del factor sea el mismo.

ANOVA de un factor de efectos fijos

$$Y_{ij} = \underbrace{\mu + \alpha_i}_{\text{Comp. Sist.}} + \underbrace{\varepsilon_{ij}}_{\text{Comp. Aleat.}} \quad , \quad i = 1, \ldots, a; \quad j = 1, \ldots, n_i$$

- $\sum_{i=1}^{a} \alpha_i = 0$ para evitar la sobreparametrización. Equivalente a $Y_{ij} = \mu_i + \varepsilon_{ij}$.
- μ representa la media global de la población
- α_i es la desviación de la media del grupo i de la media global
- ε_{ij}: desviación del individuo j de la media del grupo i; $\varepsilon_{ij} \sim N(0, \sigma^2)$ independientes

ANOVA de un factor de efectos aleatorios

$$Y_{ij} = \underbrace{\mu + A_i}_{\text{Comp. Sist.}} + \underbrace{\varepsilon_{ij}}_{\text{Comp. Aleat.}} \quad , \quad i = 1, \ldots, a; \quad j = 1, \ldots, n_i$$

- $A_i \sim N(0, \sigma_A^2)$
- μ representa la media global de la población
- A_i es el efecto aleatorio debido a ser del grupo i
- ε_{ij}: desviación del individuo j de la media del grupo i; $\varepsilon_{ij} \sim N(0, \sigma^2)$ independientes

Descomposición suma de cuadrados y Tabla ANOVA

- Si denotamos con $\overline{y}_{i.}$ a la media de cada grupo, y con $\overline{y}_{..}$ a la media global de todos los datos ($= \frac{\sum_{i,j} y_{ij}}{N} = \frac{\sum_i n_i \overline{y}_{i.}}{N}$ con $N = \sum_i n_i$), podemos descomponer cada dato de la siguiente manera:

$$y_{ij} - \overline{y}_{..} = (y_{ij} - \overline{y}_{i.}) + (\overline{y}_{i.} - \overline{y}_{..}), i = 1, \ldots, a; j = 1, \ldots, n_i$$

- Elevando al cuadrado y sumando para todos los valores de i y de j:

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{..})^2}_{\text{SS Total}} = \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i.})^2}_{\text{SS Intra o Error}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\overline{y}_{i.} - \overline{y}_{..})^2}_{\text{SS Entre}}$$

- En forma de Tabla de ANOVA (Análisis de la Varianza):

F. Variación	SS	gl	MS
Entre	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\overline{y}_{i.} - \overline{y}_{..})^2$	$a-1$	MS Entre = SS Entre/(a-1)
Error	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i.})^2$	$N-a$	MS Error = SS Error/(N-a)
Total	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{..})^2$	$N-1$	

- ¡Las sumas de cuadrados se pueden reescribir para que los cálculos sean más cómodos!

Inferencia sobre los parámetros I.

Inferencia sobre los parámetros ANOVA de un factor de efectos fijos

- Estimación parámetros (por máxima verosimilitud):
  1. $\hat{\mu} = \overline{y}_{..}$
  2. $\hat{\alpha}_i = \overline{y}_{i.} - \overline{y}_{..}$
  3. $\hat{\sigma}^2 = \text{MS Error}$
- Contraste de hipótesis: el cociente $F_s = \frac{\text{MS Entre}}{\text{MS Error}}$ es el estadístico de contraste de

$$\left.\begin{array}{ll} H_0: & \mu_1 = \mu_2 = \ldots = \mu_a \\ H_A: & \text{no } H_0 \end{array}\right\} \equiv \left.\begin{array}{ll} H_0: & \alpha_1 = \alpha_2 = \ldots = \alpha_a = 0 \\ H_A: & \text{no } H_0 \end{array}\right\}$$

  y su distribución en el muestreo bajo H_0 es una $F(a-1, N-a)$.
- Observar que si $a = 2$ el contraste resultante equivale a una comparación de dos muestras independientes. Se puede demostrar que ambos estadísticos de contraste (el F_s y el T_s) están relacionados (en concreto: $F_s = T_s^2$), y por tanto llegan a la misma conclusión.

Inferencia sobre los parámetros II.

ANOVA de un factor de efectos aleatorios

- Estimación parámetros (por máxima verosimilitud):
  1. $\hat{\mu} = \overline{y}_{..}$;
  2. $\hat{\sigma}^2 = \text{MS Error}$
  3. $\hat{\sigma}^2_A = \frac{\text{MS Entre - MS Error}}{n_0}$ (si esta cantidad es menor que 0, $\hat{\sigma}^2_A = 0$) con:
     - $n_0 = \frac{1}{N(a-1)}\left(N^2 - \sum_{i=1}^{a} n_i^2\right)$.
     - Si el diseño es equilibrado, $n_0 = n_i = n$.
- Contraste de hipótesis: el cociente $F_s = \frac{\text{MS Entre}}{\text{MS Error}}$ es el estadístico de contraste de

$$\left.\begin{array}{ll} H_0: & \sigma^2_A = 0 \\ H_A: & \sigma^2_A \geq 0 \end{array}\right\}$$

  y su distribución en el muestreo bajo H_0 es una $F(a-1, N-a)$.
- Así pues, mismo estadístico de contraste pero diferentes hipótesis, ya que el modelo es diferente y las conclusiones también son diferentes.

Validez del modelo: condiciones de aplicabilidad.

- La inferencia está basada en tres condiciones que vienen determinadas por los propios modelos.
- $\epsilon_{ij} \sim N(0, \sigma^2)$ independientes
  - Homocedasticidad: misma varianza en los grupos analizados.
  - Normalidad: de los datos de cada grupo.
  - Independencia: tanto de las muestras entre sí como de las observaciones en cada grupo.
- Tenemos dos opciones para comprobar estas condiciones:
  - Antes del ANOVA, tests de homogeneidad de varianzas (Barlett, Levene), tests de normalidad (Kolmogorov-Smirnov o Shapiro-Wilks) y tests de independencia (rachas).
  - Después del ANOVA, analizar los residuos del ajuste.
- Si no se cumplen las condiciones:
  - Podemos transformar los datos (p.e. Box-Cox).
  - Utilizar un método no paramétrico: el test de Kruskal-Wallis, una generalización del test de Mann-Whitney.

Tarea

Los siguientes datos provienen de un experimento realizado en la estación experimental de Rothamsted. El objetivo era medir la eficacia de tres insecticidas, el clorodinitrobenceno (CN), el carbón disulfido (CD) y un preparado propio denominado cymag (CM). Cada insecticida se aplicó a dosis normal (1) y doble (2). Por último se contó con un grupo control al que no se aplicó ningún insecticida. Los pesticidas se aplicaron antes de la siembra del trigo, y los datos recogidos muestran el incremento del número de gusanos encontrados en cada parcela después de la recolección del trigo.

	Insecticida					
Control	1CN	1CD	1CM	2CN	2CD	2CM
466	222	194	306	92	166	28
421	219	221	176	114	172	179
561	332	308	215	80	111	165
433	298	256	199	128	80	82

1. Especifica un modelo estadístico adecuado para analizar este experimento y explica el significado de sus parámetros.

Tarea

2. Comenta las hipótesis de aplicabilidad que deberían cumplirse para poder analizar estos datos con la técnica especificada en el apartado anterior.
3. ¿Existen diferencias estadísticamente significativas en el incremento de gusanos dependiendo del insecticida aplicado? En otras palabras, ¿hay efecto insecticida? Plantea y resuelve el contraste adecuado.
   Ayuda: comprueba para ello que con esos datos se obtiene la siguiente tabla ANOVA:

Fuentes de variación	SS	gl	MS	F
Insecticida	392447	6	65408	22.473
Residual	61121	21	2911	

4. ¿Cual es el alcance de las conclusiones que te aporta el contraste que has realizado?

ANOVA de un factor con R.

- Auque podemos utilizar el comando lm (R toma como referencia un nivel del factor y utiliza variables dummy para el resto de niveles del factor).
- Habitualmente se utiliza el comando aov

Algunos elementos importantes

- La función `summary` sobre un objeto tipo `aov` produce la tabla de ANOVA necesaria para realizar el contraste de comparación de las medias.
- `fitted.values`: Valores ajustados, $\hat{y}_i$
- `residuals`: Valores de los residuos (no tipificados)
- Con el comando `rstandard` accedemos a los residuos estandarizados.

Validación del ajuste realizado

- Podemos realizar la comprobación de las condiciones de aplicabilidad antes de realizar el ANOVA:
  1. *Homocesdasticidad*: para comprobar si las varianzas son homogéneas podemos utilizar el test de Bartlett (`bartlett.test()`), el de Fligner-Killeen (`fligner.test()`) o el de Levene (`leveneTest()` de la librería `car`).
  2. *Normalidad*: para comprobarla podemos utilizar el test de Kolmogorov Smirnov (`ks.test()`) o el de Shapiro Wilk (`shapiro.test()`).
  3. *Indepedencia*: comprobando si existen rachas (grupos de valores ininterrumpidos en la misma dirección) podemos valorar la independencia de los datos.
- Como en regresión con el comando `plot(objetoaov)` tenemos cuatro gráficas que nos permiten validar la adecuación del modelo.
- Si no tenemos condiciones de aplicabilidad podemos transformar los datos (la mejor transformación nos la da la función `boxcox`) o utilizar el test de Kruskal Wallis (`kruskal.test`) que contrasta la hipótesis:

$$H_0\text{: Mediana}_1 = \ldots = \text{Mediana}_p$$

¿Cambia la longitud de la sección de un guisante con la adición de azúcares?

Situación

Los siguientes datos provienen del ejemplo expuesto en el libro de Sokal y Rohlf (Biometría: principios y métodos estadísticos en la investigación biológica) y hace referencia a un estudio sobre la longitud de secciones de guisantes (en unidades oculares de 0.114 mm.) criados en cultivos con adición de distintos azúcares.

control	glucosa	fructosa	gluc+fruct	sacarosa
75	57	58	58	62
67	58	61	59	66
70	60	56	58	65
75	59	58	61	63
65	62	57	57	64
71	60	56	56	62
67	60	61	58	65
67	57	60	57	65
76	59	57	57	62
68	61	58	59	67

Lectura y descriptiva.

```
data <- read.table(file="guisantes.dat",header=T)
attach(data)
by(longitud,tratamiento,summary)
by(longitud,tratamiento,var)
boxplot(longitud ~ tratamiento,boxwex=0.75,ylim=c(50,80),
xlim=c(0.5,7),col=4)
boxplot(longitud,add=TRUE,boxwex=1.5,at=6.5,col=2)
text(6.5,52,"Total")
```

Validación condiciones de aplicabilidad antes.

```
bartlett.test(longitud ~ tratamiento)
fligner.test(longitud ~ tratamiento)
library("car")
leveneTest(longitud ~ tratamiento)
by(longitud, tratamiento,shapiro.test)
library(tseries)
runs.test(as.factor(longitud > median(longitud)))
```

Validación condiciones aplicabilidad después.

```
anova <- aov(longitud ~ tratamiento)
summary(anova)
model.tables(anova,type="means")
model.tables(anova,type="effects")
```

Test no paramétrico y transformaciones.

```
anova.ks <- kruskal.test(longitud,tratamiento)
anova.ks
#Transformaciones
library(MASS)
bc <- boxcox(anova,lambda=seq(-6,0,length=20))
lambda <- bc$x[which.max(bc$y)]
lambda
trans.longitud <- longitud^(-3)
#Luego de transformar, las condiciones de aplicabilidad se
cumplen.
bartlett.test(trans.longitud,tratamiento)
by(trans.longitud,tratamiento,shapiro.test)
anova.trans <- aov(trans.longitud~ tratamiento)
summary(anova.trans)
model.tables(anova.trans,type="means")
```

Comparaciones a posteriori.

- Cuando se rechaza la hipótesis nula en un contraste en el que el interés es las diferencias entre grupos, la conclusión es que al menos uno de los grupos tiene la media diferente.
- Pero no nos aclara ni que grupo es el que tiene la media diferente ni si es uno sólo.
- Para mejorar las conclusiones se puede utilizar comparaciones dos a dos de las medias de los grupos y en algunas de ellas hay que tener presente la protección del error global:
  - Método de Tukey
  - Método basado en las diferencias significativas utilizando la corrección de Bonferroni.
  - Otros más (Scheffé, Student-Newman-Keuls, Dunn-Sidak, etc.)

Comparaciones a posteriori: caso particular diseño equilibrado (n tamaño grupo)

1. Cálculo y ordenación de las medias de los grupos.
2. Cálculo de las diferencias dos a dos de las medias ordenadas y construcción de una tabla de diferencias de medias.
3. Obtención de las diferencias significativas (en base a cualquiera de los métodos) como aquellas que superan los valores:
   - Método Bonferroni. Si α' es el error global máximo a cometer:

$$LSD_\alpha = \sqrt{F_{\alpha(1,N-a)}}\sqrt{\frac{2}{n}MSError}$$

     con $\alpha = \frac{\alpha'}{k}$ siendo k el número de comparaciones dos a dos.
   - Método Tukey:

$$MSR_\alpha = Q_{\alpha,(a,N-a)}\sqrt{\frac{MSError}{n}}$$

     con $Q_{\alpha,(a,N-a)}$ valor crítico en la tabla de rangos estudentizados.
4. Obtención de subgrupos homogéneos: aquellos subgrupos con medias similares.

Tarea

1. Obtener los grupos homogéneos resultantes de aplicar el método de Tukey con los datos de los insecticidas. ¿De qué manera observas que se amplian las conclusiones que puedes aportar sobre el problema planteado?
2. Para controlar el posible impacto medioambiental que supondría el incendio de varias fábricas de tejidos próximas a un bosque, se determinó el tiempo (en segundos) que tardaban en arder 5 vestidos, elegidos al azar, realizados en cada una de ellas. Del análisis de los datos se obtuvieron los siguientes resultados:

Fábrica	1	2	3	4	5
Media	16.78	11.76	10.24	11.98	15.26
Desv. Típica	1.167	2.3298	1.1437	1.862	0.9182

   1. Si medimos la peligrosidad de una fábrica por el tiempo que tardan en arder sus vestidos, ¿hay evidencia para pensar que los tejidos de las fábricas influyen en su peligrosidad?
   2. Calcula los subgrupos homogéneos resultantes de aplicar tanto el método de la diferencia significativa con la corrección de Bonferroni como el método de Tukey. ¿Qué conclusiones puedes extraer de estos grupos homogéneos? ¿Cuáles son tus conclusiones globales sobre el análisis realizado?

Comparaciones a posteriori con R.

Comparaciones a posteriori.

```
ajuste.tukey <- TukeyHSD( anova.trans, ordered =T)
ajuste.tukey
plot ( ajuste.tukey )
library ("agricolae")
ajuste.LSD <- LSD.test( anova.trans,"tratamiento",p.adj
="bonferroni")
ajuste.LSD
ajuste.SNK <- SNK.test( anova.trans ,"tratamiento")
ajuste.SNK
ajuste.Scheffe <- scheffe.test ( anova.trans,"tratamiento")
ajuste.Scheffe
```