

IMM con R. Lab11: Efecto de la dispersión multivariada sobre ANOSIM, PERMANOVA y PERMANOVA modificado

Dra. Maite Mascaro y Dr. Edlin Guerra Castro

25/05/2022

Parte I: Desempeño de pruebas estadísticas en presencia de diferencias de dispersión.

Para evaluar el desempeño de las pruebas estadísticas multivariadas más populares usaremos simulación de datos garantizando igualdad en la posición de sus centroides grupales pero generando diferencias en dispersión. Luego someteremos a prueba la hipótesis nula (cierta) de igual posición de los centroides usando las pruebas ANOSIM (`anosim` de `vegan`), PERMANOVA (`adonis` de `vegan`) y una versión modificada de PERMANOVA. Lo primero que haremos será simular dos juegos de datos (30 variables cada uno) usando los siguientes comandos:

```
# Parámetros para simular las 30 variables
set.seed(42)
ave <- runif(30, min = 10, max = 40)
var <- sqrt(ave)

# Matriz para juego de datos 1, 20 observaciones
sim1 <- matrix(data = NA, nrow = 20, ncol = 30)

# Matriz para juego de datos 2, 10 observaciones
sim2 <- matrix(data = NA, nrow = 10, ncol = 30)

# Bucle para llenar matrices, noten como se duplica la varianza
# en el juego de datos 2
for (j in 1:30){
  sim1[,j] <- rnorm(20, mean = ave[j], sd=var[j])
  sim2[,j] <- rnorm(10, mean = ave[j], sd=2*var[j])
}

# Convertimos ambas matrices en data.frames, y se combinan
sim1 <- as.data.frame(sim1)
sim2 <- as.data.frame(sim2)
sim <- rbind(sim1, sim2)

# Generamos un vector que reconoce ambos grupos
grupos <- c(rep("a",20), rep("b",10))
```

El objeto de nombre `sim` alberga los dos juegos de datos. Podemos verificar las propiedades multivariadas de este juego de datos proyectando los 30 objetos en un MDS no métrico, y luego podemos medir la dispersión. Esto se logra con los siguientes códigos:

```

# centramos todas las variables con una normalización
sim.n <- decostand(sim, method = "standardize")

#Estimamos las distancias Euclidianas
euc <- dist(sim.n)

#Generamos un MDS con vegan y graficamos con ggplot2
mds1<-metaMDS(euc)
MDS1 <- as.data.frame(mds1$points)
MDS1$grp <- grupos

plot.MDS1 <- ggplot(data=MDS1 ,aes(x=MDS1, y=MDS2))+
  geom_point(aes(colour=grupos), size=3.5)+
  theme_bw(base_size=16)

plot.MDS1

```

Con base en este gráfico, responda las siguientes preguntas:

- ¿Puede apreciar las diferencias multivariadas en dispersión entre ambos grupos?
- ¿Puede apreciar las diferencias (de haber) en la posición de los centroides de ambos grupos?
- ¿Cómo puede confirmar que ambas diferencias en dispersión son poco probables de ocurrir por simple azar?

Para responder la última pregunta se sugiere aplicar una prueba estadística para dispersión multivariada. Una de las más exitosas es **PERMDISP**, que en **vegan** está disponible como la función **betadisper** y **permutest**. Le recomiendo aplicar estas funciones para responder la pregunta. Verificada las diferencias en dispersión, lo que sigue es someter a prueba la hipótesis nula de que ambos grupos tienen el mismo centroide. Recuerde que la hipótesis nula es cierta, por lo que cualquier resultado significativo será un error inferencial del tipo I. Para esto use **anosim**, **adonis** y el script de **PERMANOVA** acá compartido (corra todas las funciones para que tenga disponible **PERMANOVA2**):

```

## Funciones preliminares

# Sum of Squares using Theorem of Huygen
SS <- function (d) {
  n <- dim(as.matrix(d))[1]
  ss <- sum(d ^ 2) / n
  return(ss)
}

# Multivariate Dispersion
v = function (d) {
  n <- dim(as.matrix(d))[1]
  ss <- sum(d ^ 2) / n
  v <- ss / (n - 1)
  return(v)
}

# Modified Pseudo-F (Anderson et al., 2017)
pseudo.F <- function (x, factor, distancia) {
  d <- vegdist(x, method = distancia)
  TSS <- SS(d)
  group <- as.data.frame(factor)

```

```

x$grp <- factor
factor <- as.factor(factor)
lab <- names(table(group))
lev <- table(group)
CR <- c(1:nlevels(factor))
for (i in 1:nlevels(factor)) {
  CR[i] <-
    SS(vegdist(x[x$grp == lab[i], 1:length(x) - 1], method = distancia))
}
RSS <- sum(CR)
Var <- c(1:nlevels(factor))
d.res <-
  as.data.frame(matrix(nrow = length(levels(factor)), ncol = 3))
for (i in 1:nlevels(factor)) {
  Var[i] <-
    v(vegdist(x[x$grp == lab[i], 1:length(x) - 1], method = distancia))
  d.res[i, ] <- c(lev[i],
                  Var[i],
                  (1 - (lev[i] / sum(lev))) * Var[i])
}
den <- sum(d.res$V3)
ASS <- TSS - RSS
Fobs <- ASS / den
return(Fobs)
}

## PERMANOVA2

PERMANOVA2 <- function(x, factor, distancia, nperm = 999) {
  control <- how(nperm = nperm, within = Within(type = "free"))
  Fobs <- pseudo.F(x, factor, distancia = distancia)
  Nobs <- nobs(x)
  F.permu <- numeric(length = control$nperm) + 1
  F.permu[1] <- Fobs
  ## Generation of pseudo.F values for H0 using permutations without replacement
  for (i in seq_along(F.permu)) {
    ## return a permutation
    want <- permute(i, Nobs, control)
    ## calculate permuted F
    F.permu[i + 1] <-
      pseudo.F(x[want, ], factor, distancia = distancia)
  }
  ## probability for Fobs
  pval <- sum(abs(F.permu) >= abs(F.permu[1])) / (control$nperm + 1)
  ## Results
  return(data.frame("Pseudo-F" = F.permu[1], "p(perm)" = pval))
}## Permutation test based on Anderson et al. (2017)

```

Para efectos de este ejercicio, PERMANOVA2 correrá así:

```

PERMANOVA2(x = sim.n,
  factor = grupos,
  distancia = "euclidean",
  nperm = 999

```

)

- d. Con base en los resultados, interprete los resultados de **ANOSIM**.
- e. ¿Qué sugiere el resultado de **PERMANOVA**?
- f. En ambos casos ¿cuál fue la probabilidad del estadístico en pertenecer a la hipótesis nula?
- g. ¿qué indicó PERMANOVA2, modificado por Anderson et al 2027?
- h. ¿Cuál de las tres pruebas le permitió respaldar lo que se aprecia en el gráfico MDS arriba generado?

Parte II: Caso Ekofisk

En el *Laboratorio 6* se evaluó gráficamente los cambios en la macrofauna del bentos respecto a la distancia a la plataforma de perforación Ekofisk. Visualmente apreciamos patrones de diferencia según la hipótesis, pero formalmente no evaluamos esa hipótesis. Usemos para esto las funciones **anosim** y **adonis** para detectar diferencias entre los grupos de distancia. También evaluemos la dispersión multivariada con **betadisper** y **permutest**. Interprete los resultados y proponga una solución en caso de que se comprometa la interpretación de **ANOSIM** y **PERMANOVA**. Formalmente las preguntas a atender son:

- a. Importe la matriz de datos. Identifique la información de sus columnas y filas.
- b. Recuerde que se dividir los datos en dos objetos: una matriz/dataframe con las columnas con información de especies, y una dataframe con información que identifique a los grupos.
- c. Aplique el índice de similitud Bray-Curtis luego de transformar las abundancias a raíz cuarta.
- d. Genere el MDS y represente los grupos de distancia, usando isoMDS y ggplot
- e. Evalúe la dispersión. ¿qué interpreta? ¿cuál grupo presenta mayor dispersión?
- f. Ahora someta a prueba la hipótesis de cambio en composición de especies según la distancia. ¿cómo se puede interpretar el resultado, considerando la salida de betadisper?
- g. Considerando este set de datos ¿podría replantear un proyecto con un nuevo diseño de muestreo? Intente estimar el error estándar multivariado a distintos esfuerzos de muestreo a los datos de Ekofisk, usando para ello el paquete **SSP**.