Principios de Estadística

Leonardo Collado Torres y María Gutiérrez Arcelus
Licenciatura en Ciencias Genómicas, UNAM
www.lcg.unam.mx/~lcollado/index.php
www.lcg.unam.mx/~mgutierr/index.php

Cuernavaca, México
Febrero - Junio, 2009

ANOVA

Objetivos

- Hoy vamos a ver como resolver una ANOVA en R
- Terminaremos con unos problemas para que los resuelvan :)

Definiendo ANOVA

- Es un método para comparar medias basado en variaciones de la media.
- La sencilla, *one-way*, es una generalización de la prueba t para dos muestras independientes que nos permite comparar varias muestras independientes.
- Tenemos k poblaciones con una muestra de cada una, siendo las poblaciones independientes. Si la media de la población i es μ_i y la desviación estándar es σ_i [1], nuestro modelo estadísto es:

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

- donde los términos de error, ε_{ij}, son independientes con una distribución Normal (o, σ)

[1] Si son iguales usamos solo σ

Hipótesis en prueba

- Los modelos se van a hacer más complicados, pero por ahora nuestras hipótesis son las siguientes:
  1. $H0$: $\mu_1 = \mu_2 \ ... \ = \mu_k$
  2. HA: $\mu_i \neq \mu_j$ para al menos un par i y j.
- ¿Por qué? Simplemente porque estamos asumiendo que todas nuestras poblaciones se distribuyen normalmente. [2]

[2] En wiki viene como el "modelo de efectos fijos".

ANOVA como Fisher

- En sí una ANOVA es una prueba que utiliza la estadística F de Fisher. Para esto, tenemos los siguientes términos[3]:
  - Suma total de cuadrados, $\text{STC} = \sum_i \sum_j (x_{ij} - \bar{x})^2$
    - Mide la cantidad de variación desde el centro de todos los datos.
  - Suma de errores cuadrados, $\text{SEC} = \sum i \sum j (x_{ij} - \bar{x}_i)^2$
    - Mide la variación dentro del grupo i.
  - Suma de tratamientos cuadrados, $\text{STrC} = \sum_i n_i (\bar{x}_i - \bar{x})^2$
    - Compara la media de cada grupo con la media total.
- La estadística F como tal es así:

$$F = \frac{STrC/(k-1)}{SEC/(n-k)}$$

[3] SST, SSE y SSTr en inglés

oneway.test

- Todo el rollo de la ANOVA es que no sabemos si la variación que observamos está dada porque nuestra $H0$ es falsa o porque se deba a la variación entre las muestras.
- Es por eso que usamos la F, y bueno, ya conociendo nuestras hipótesis, la función más directa para este tipo de ANOVA es la oneway.test. Chequen la ayuda :)

  ```
  > `?`(oneway.test)
  ```

- Como ven, el objeto resultante es de clase htest.
- Fíjense bien que los datos se los pasamos en tipo "formula". Claro, si quieren siempre pueden hacerlo paso a paso con las fórmulas que les puse anteriormente :p

Example (Primera ANOVA)

Supongamos que medimos el tiempo (en segundos) que 15 personas toman para completar la misma entrevista de trabajo. Por cuestiones logísticas, los dividieron en grupos de 5 para entrevistarlos en 3 días diferentes y estos fueron sus tiempos:

1. 2166, 1568, 2233, 1882, 2019
2. 2279, 2075, 2131, 2009, 1793
3. 2226, 2154, 2583, 2010, 2190

Asumimos que nuestros datos se distribuyen normalmente con la misma varianza. Nuestras $H0$ y HA son iguales a las que acabamos de ver. Hagan una prueba de ANOVA y encuentren el valor p.

- Así lo podemos resolver:

```
> datos <- stack(list(dia1 = c(2166,
+     1568, 2233, 1882, 2019), dia2 = c(2279,
+     2075, 2131, 2009, 1793), dia3 = c(2226,
+     2154, 2583, 2010, 2190)))
> names(datos)

[1] "values" "ind"

> oneway.test(values ~ ind, data = datos,
+     var.equal = T)

One-way analysis of means


data:  values and ind
F = 1.7862, num df = 2, denom df =
12, p-value = 0.2094
```

- ¿Qué concluímos?
- Noten que usamos una nueva función, stack, para agrupar nuestros datos en un `data.frame` pero manteniendo la información de nuestros 3 días.
- Les recomiendo que luego chequen como se ve el objeto `datos` con y sin `stack`.

aov

- Existe otra función para hacer ANOVAs sencillas, *oneway*, aunque también sirve para otras más complicadas. Se llama aov.
- Si checan la ayuda se van a dar cuenta de que es mucho más complicada, así que mejor sigamos con nuestro ejemplo. Es que usa modelos lineales que no hemos visto, los `lm`.

```
> `?`(aov)
```

```
> dos <- aov(values ~ ind, data = datos)
```

Utilidad de aov

- ¿Para que usamos aov? Simplemente porque podemos imprimir más datos con ella. Podemos ver cierta info usando `print` o llamando el objeto. Además podemos obtener la tabla de resumen usando `summary`.

```
> dos

Call:
   aov(formula = values ~ ind, data = datos)

Terms:
                       ind Residuals
Sum of Squares  174664.1  586719.6
Deg. of Freedom         2        12
```

Utilidad de aov


```
Residual standard error: 221.1183
Estimated effects may be unbalanced
```

```
> summary(dos)
```

```
             Df Sum Sq Mean Sq F value
ind           2 174664   87332  1.7862
Residuals    12 586720   48893
             Pr(>F)
ind          0.2094
Residuals
```

- "Residuals" es lo mismo que "Error".

Problema 1

- Ahora quiero que resuelvan los siguientes ejercicios. Tienen que subir a la página de Cursos su script con comentarios[4]. Por problema, deben hacer un `boxplot` u otra gráfica antes para ver si pueden asumir varianzas iguales o no.
- Problema 1. El set de datos de `morley` contiene mediciones de la velocidad de la luz hechas por Michaelson y Morley. Hicieron 5 experimentos, cada uno con varias repeticiones. Hagan una ANOVA simple para ver si los 5 experimentos tienen la misma media poblacional.
- Les recomiendo que usen head y tail para explorar sus datos en cada problema.

```
> head(morley)
```

Problema 1

	Expt	Run	Speed
001	1	1	850
002	1	2	740
003	1	3	900
004	1	4	1070
005	1	5	930
006	1	6	850

[4] No olviden sus conclusiones!!

Problema 2

- Usando el set de datos `Cars93` del paquete `MASS`, hagan una ANOVA simple para las variables `MPG.highway` y `DriveTrain`. ¿Sus datos apoyan a la $H0$ de medias poblacionales iguales?
- Tienen que cargar la librería `MASS` con el siguiente comando para poder usar `Cars93`.

```
> library(MASS)
```

Problema 3

- Una compañía necesita de cierto químico como materia prima y está buscando donde mandarlo a hacer. Antes de tomar una decisión, le pide a 4 laboratorios que le hagan 5 muestras. Vemos los resultados en alguna métrica en la siguiente tabla.
- ¿Hay una diferencia entre las medias de las poblaciones?

Lab 1	4.13	4.07	4.04	4.07	4.05
Lab 2	3.86	3.85	4.08	4.11	4.08
Lab 3	4.00	4.02	4.01	4.01	4.04
Lab 4	3.88	3.89	3.91	3.96	3.92

Table 1: Producción de un químico