Sesion 4: Diseño de Experimentos

Carlos Ignacio Patiño (cpatinof@gmail.com)

Agosto 1, 2015

Agenda

- Diseño de Experimentos
- Análisis de Variables Categóricas
- Taller en Clase
- Discusión Capsula de Datos
- Break
- Sesión de trabajo en proyecto final (hasta el final de la sesión)

Diseñando Experimentos

Hasta ahora hemos venido haciendo inferencia con experimentos obsevacionales de muestreo y no con experimentos diseñados de muestreo.

Estudios Observacionales

- Poco control
- Simplemente se observa lo que se tiene

Experimentos Diseñados

- El analista intenta controlar los niveles de una o más variables, con el fin de determinar el efecto sobre la variable de interés
- Permiten inferir causa y efecto

Elementos de un diseño experimental

- Variable de respuesta (también conocida como variable dependiente)
- Factores: variables cuyo efecto sobre la respuesta es el interés fundamental del diseñador (también llamadas variables independientes)
- Niveles de los factores: valores del factor usado en el experimento
- Tratamientos: Combinaciones factor-nivel usadas en el experimento
- Unidad experimental: objeto sobre el cual se observan las respuestas

Experimento completamente aleatorio: un sólo factor

El diseño más simple consiste en la selección completamente aleatoria de las unidades experimentales que representan cada tratamiento.

El objetivo de este tipo de experimentos is comparar las medias de cada tratamiento:

$$H_o : \mu_1 = \mu_2 = ... = \mu_k$$

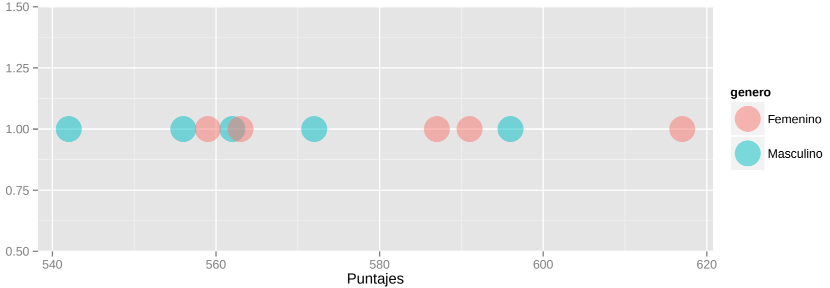
$$H_a : \text{Al menos dos de las k medias difieren}$$

En esta hipótesis, comparamos las medias de las muestras aleatorias independientes (cada tratamiento).

Ejemplo 1

scores	genero
556	Masculino
596	Masculino
542	Masculino
572	Masculino
562	Masculino
559	Femenino
591	Femenino
617	Femenino
587	Femenino
563	Femenino

Diferencia dominada por variabilidad de muestreo




```
[1] "Media puntajes hombres"

[1] 565.6

[1] "Media puntajes mujeres"

[1] 583.4
```

Calculemos algunas métricas de interés (1/2)

Suma de Cuadrados de los Tratamientos (SST), o variación entre las medias de cada tratamiento

Calculada como la diferencia entre cada media de cada tratamiento y la media global (todos las unidades experimentales), elevada al cuadrado, multiplicado por el número de medidas en la muestra y sumados para cada tratamiento.

$$SST = 5(565.6 - 574.5)^2 + 5(583.4 - 574.5)^2 = 792.1$$

Suma de Cuadrados para el Error (SSE), o variabilidad muestral dentro de los tratamientos

Calculada al sumar la distancia cuadrada entre cada respuesta medida y su correspondiente media de tratamiento y luego sumando las diferencias al cuadrado para todas las medidas en la muestra completa. SSE se puede simplificar como:

$$SSE = (n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2 + \ldots + (n_k - 1){s_k}^2$$

$$SSE = (5 - 1)(406.8) + (5 - 1)(552.8) = 3838.4$$

Calculemos el estadístico F

$$MST = \frac{SST}{k-1} = SST/(2-1) = 792.1$$

$$MSE = \frac{SSE}{n-k} = SSE/(10-2) = 479.8$$

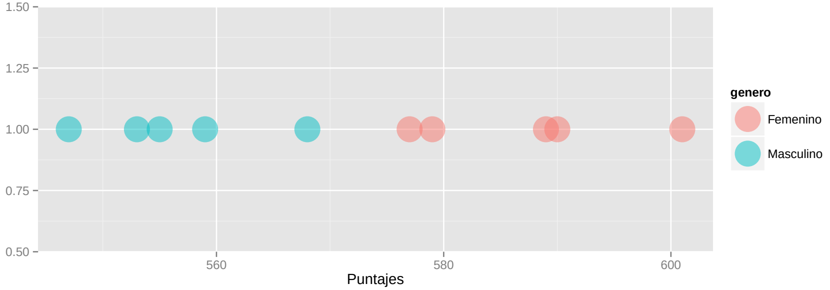
$$F = \frac{MST}{MSE} = 1.65$$

Se observa un estadístico F muy bajo, por lo que no es posible rechazar la H_0. El F crítico para este caso es 4.96 (ver tabla F)

Ejemplo 2

scores	genero
553	Masculino
568	Masculino
547	Masculino
559	Masculino
555	Masculino
577	Femenino
590	Femenino
601	Femenino
589	Femenino
579	Femenino

Diferencia elevada relativa a la variabilidad de muestreo




```
[1] "Media puntajes hombres"

[1] 556.4

[1] "Media puntajes mujeres"

[1] 587.2
```

Suma de Cuadrados de los Tratamientos (SST), o variación entre las medias de cada tratamiento

Calculada como la diferencia entre cada media de cada tratamiento y la media global (todos las unidades experimentales), elevada al cuadrado, multiplicado por el número de medidas en la muestra y sumados para cada tratamiento.

$$SST = 5(556.4 - 571.8)^2 + 5(587.2 - 571.8)^2 = 2371.6$$

Calculemos algunas métricas de interés (2/2)

Suma de Cuadrados para el Error (SSE), o variabilidad muestral dentro de los tratamientos

Calculada al sumar la distancia cuadrada entre cada respuesta medida y su correspondiente media de tratamiento y luego sumando las diferencias al cuadrado para todas las medidas en la muestra completa. SSE se puede simplificar como:

$$SSE = (n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2 + \ldots + (n_k - 1){s_k}^2$$

$$SSE = (5 - 1)(60) + (5 - 1)(93.2) = 616$$

Calculemos el estadístico F

$$MST = \frac{SST}{k-1} = SST/(2-1) = 2371.6$$

$$MSE = \frac{SSE}{n-k} = SSE/(10-2) = 77$$

$$F = \frac{MST}{MSE} = 30.8$$

Se observa un estadístico F superior al valor crítico de 4.96, por lo que **es posible rechazar** la H_0. Las medias son diferentes.

- Empleamos las funciones `lm()` y `anova()`

```
modelo <- lm(scores~genero,df)
anova(modelo)
```

```
Analysis of Variance Table

Response: scores
          Df Sum Sq Mean Sq F value    Pr(>F)
genero     1 2371.6  2371.6    30.8 0.0005411 ***
Residuals  8  616.0    77.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Otros diseños empleados

Diseño de bloques aleatorios:

- Agrupa las unidades experimentales en subconjuntos homogéneos (bloques) para luego aplicar (aleatoriamente) los tratamientos a cada subconjunto
- En el ejemplo de los estudiantes y sus puntajes, una posible agrupación (parejas de estudiantes) puede provenir de su colegio de origen y promedio acumulado.

Experimento Factorial Completo:

- En este diseño cada combinación factor-nivel es empleada, en otras palabras, el número de tratamientos en el experimento es igual al total de combinaciones factor-nivel

Variables categóricas

- Tabla de una vía que permite mirar las probabilidades de una categoría
- Tabla de dos vías permite analizar dos factores
- Pruebas *Chi Cuadrado* para pruebas de independencia entre variables

- Dos variables aleatorias son independientes si la distribución de probabilidad de una variable no se afecta por la presencia de la otra
- Para probar independencia (en variables categóricas) se emplea la prueba *Chi Cuadrado*
- Miremos un ejemplo (filas: fumador, columnas: ejercicio)

	Freq	None	Some
Heavy	7	1	3
Never	87	18	84
Occas	12	3	4
Regul	9	1	7

H_0 : Las dos clasificaciones son independientes
H_a : Las dos clasificaciones son dependientes

```
chisq.test(tbl)
```

```
    Pearson's Chi-squared test

data:  tbl
X-squared = 5.4885, df = 6, p-value = 0.4828
```