

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

Pruebas de hipótesis: lo básico

Principios de
Estadística

Clase htest

Ejemplos
prácticos

Problema de
la clase

1 Clase htest

2 Ejemplos prácticos

3 Problema de la clase

Objetivos

Principios de
Estadística

Clase htest

Ejemplos
prácticos

Problema de
la clase

- Después de un largo descanso...
- En la clase de hoy vamos a ver las bases de las funciones de pruebas de hipótesis en R .

Márgenes de una tabla

Principios de
Estadística

Clase htest

Ejemplos
prácticos

Problema de
la clase

- Antes de entrar en el tema, les quiero enseñar dos formas de obtener fácilmente los datos marginales de una tabla.

```
> x <- matrix(rnorm(9), 3, 3)  
> margin.table(x, 1)  
> margin.table(x, 2)  
> addmargins(x)
```
- Intuitivamente, ¿qué resultado nos da cada una de las `margin.table` ?
- Esto les debe ser útil en casos como la ji cuadrada, etc.

- Para ver que tan rápido corren podemos usar la función `system.time`¹ la cual nos regresa 3 medidas. Fíjense en la de elapsed.

```
> x <- matrix(rnorm(1e+06), 10000,  
+           100)
```

```
> system.time(margin.table(x, 1))
```

```
user  system elapsed  
0.17   0.00   0.17
```

```
> system.time(margin.table(x, 2))
```

```
user  system elapsed  
0.03   0.00   0.03
```

```
> system.time(addmargins(x))
```

Speedy

Principios de
Estadística

Clase htest

Ejemplos
prácticos

Problema de
la clase

```
user  system elapsed
0.70   0.06   0.77
```

¹Para una buena comparación tienen que hacer unas 500 o 1000 réplicas y comparar las medias

Clase de un objeto

- Si recuerdan, hay varios tipos de objeto que hemos usado a lo largo del semestre. Por ejemplo:

```
> x <- 1
> y <- "1"
> z <- list(reprobados = 5)
> class(x)
```

```
[1] "numeric"
```

```
> class(y)
```

```
[1] "character"
```

```
> class(z)
```

```
[1] "list"
```

- De los anteriores 3 objetos, ¿cuál tiene atributos?

Atributos de un objeto

Principios de
Estadística

Clase htest

Ejemplos
prácticos

Problema de
la clase

- En realidad no hemos aprovechado los atributos de los objetos mucho. Si a caso en los objetos `data.frame` o `list`.

```
> attributes(x)
```

```
NULL
```

```
> attributes(y)
```

```
NULL
```

```
> attributes(z)
```

```
$names
```

```
[1] "reprobados"
```


Usando los atributos

- En R, las funciones de prueba de hipótesis regresan objetos de clase `htest`. Estos varían dependiendo de cada función; usen la función `attributes` o chequen la ayuda.

- Todos regresan el atributo `names` con el cual pueden obtener algo que les interese, por ejemplo, el valor p.

```
> z$reprobados
```

```
[1] 5
```

```
> class(z$reprobados)
```

```
[1] "numeric"
```

- Noten que aunque `z` es de clase `list`, `z$reprobados` es de clase `numeric`.

- Tal vez una de las pruebas más fáciles es la de la binomial, que es la repetición de un experimento Bernoulli n veces.
> `?` (*binom.test*)
- Chequen la ayuda y ahora resolvamos el siguiente problema.
- Un examen dado tiene 70 preguntas donde cada una tiene 5 respuestas posibles. ¿Cuál es el valor p de que alguien saque x preguntas buenas?
 - 1 7
 - 2 68
 - 3 50
- ¿Cuál creen que será el que nos de el valor p más extremo?

Solución

- Para resolver este problema en realidad solo vamos a usar 3 argumentos de la función `binom.test`. ¿Cuáles?
- Como solo queremos el valor p de regreso, aprovechemos dicho atributo.

```
> binom.test(7, 70, 0.2)$p.value
```

```
[1] 0.03589705
```

```
> binom.test(68, 70, 0.2)$p.value
```

```
[1] 4.594981e-45
```

```
> binom.test(50, 70, 0.2)$p.value
```

```
[1] 2.327952e-20
```

- Si se dan cuenta, ya ni guardamos el resultado de la función en algún objeto, pues solo nos interesaban los 3 valores p.

Cuidado al concluir

Principios de
Estadística

Clase htest

Ejemplos
prácticos

Problema de
la clase

- Tengan **cuidado** para que no vayan a concluir algo erróneamente después de usar una prueba.
- En el ejemplo anterior, ¿cuál era la hipótesis alternativa? Simplemente utilicemos la función sin almacenarla en algún objeto.
- Corran el siguiente comando:

```
> binom.test(x = 35, n = 70, p = 0.2)
```
- De esta prueba, podemos rechazar la H_0 en favor de la H_A y nada más.

- Si recuerdan, vimos las gráficas tipo QQplot y en específico a la QQnorm.
- Existe una prueba de hipótesis, la cual es más fácil de interpretar, para checar si tus datos se distribuyen como una normal.
- Chequen la ayuda y corran el siguiente comando de **shapiro.test**:

```
> `?`(shapiro.test)
> shapiro.test(rnorm(10000))
```
- Como ven, cada función tiene sus peculiaridades. Por ahora no nos interesa mucho saber como funcionan, si no como utilizarlas. Aunque si alguien quiere, R te ofrece referencias por cada función.

```
> shapiro.test(rnorm(1000))$p.value
```

```
[1] 0.7557873
```

```
> shapiro.test(runif(1000, 0, 3))$p.value
```

```
[1] 1.936899e-17
```

- Con `shapiro.test` es evidente cual no es normal y cual si, lo cual tal vez era un poco más complicado en un diagrama de caja y brazos.

Comparando tiempos de expresiones

Principios de
Estadística

Clase htest

Ejemplos
prácticos

Problema de
la clase

- Bueno, ahora pasemos al ejercicio principal.
- Usando `t.test` hagan la prueba de hipótesis para ver si las medias de los tiempos² en los que corren las siguientes expresiones son iguales o diferentes. Las medias son de 100 eventos y solo necesitaremos el valor p para poder sacar nuestra conclusión.

```
> x <- rnorm(1e+06)
> sort(x)
```

²Solo fíjense en `elapsed`

- Usen un ciclo `for` o si alguien se anima, un `apply`.
- Solo necesitan los datos de 1 posición de la función `system.time`
- Al usar `t.test` usen la forma más simple. Osea, entre menos argumentos usen, mejor :)

Solución

Principios de
Estadística

Clase htest

Ejemplos
prácticos

Problema de
la clase

```
> dat1 <- NULL
> dat2 <- NULL
> for (sim in 1:100) {
+   dat1 <- c(dat1, system.time(x <- rnorm(1e+06))[3])
+   dat2 <- c(dat2, system.time(sort(x))[3])
+ }
> t.test(dat1, dat2)$p.value

[1] 9.548745e-162
```

- Dados nuestros datos, rechazamos las H_0 en favor de la H_A , osea que las medias no son iguales.

Una trampa

Principios de
Estadística

Clase htest

Ejemplos
prácticos

Problema de
la clase

■ Tengan cuidado!! Miren que pasa ahora:

```
> dat1 <- NULL
> dat2 <- NULL
> for (sim in 1:100) {
+   dat1 <- c(dat1, system.time(x <- rnorm(1000))[3])
+   dat2 <- c(dat2, system.time(sort(x))[3])
+ }
> t.test(dat1, dat2)$p.value

[1] 0.08093636
```