UAX Universidad Alfonso X el Sabio

Estadística con R

BIOINFORMATICA 25-26

Grado en Biomedicina

Estadística básica

R tiene una serie de funciones para realizar multitud de cálculos y operaciones estadísticas:

`sum()`, `mean()`, `median()`, `max()`, `min()`, `sd()`, `var()`, `length()`, `quantile()`…

También cuenta con un montón de paquetes o librerías que realizan análisis estadísticos muy concretos.

Nosotros nos vamos a centrar en los que vienen por defecto en `R::base`.

`Saphiro-Wilk, Kolgomorov-Smirnov`

`Fisher`

`T-test y ANOVA`

`Wilcox y Kruskal Wallis`

`Chi Square`

`Correlación`

`FDR`

Saphiro-Wilk test

```
> shapiro.test()
```

Se emplea para determinar la **normalidad** de una distribución (3-5000 observaciones).

***Para muestras mayores se emplea un quantile-quantile plot.

Si la distribución **es normal -> test paramétricos.**

Si la distribución **no es normal -> test no paramétricos.** will-cox = T-student

Kolgomórov-Smirnov

```
> ks.test()
```

Se emplea para evaluar si dos muestras corresponden a la misma población.

Ho: ambas muestras pertenecen a la misma población.

H1: las muestras pertenecen a distintas poblaciones.

T-Student y ANOVA

```
> t.test()
```

Método paramétrico para comparar la media de máximo 2 muestras de distribución normal.

```
> aov()
```

Método paramétrico para comparar la media de 3 o más muestras de distribución normal.

Análisis de varianzas y muestras independientes (complejo).

human

anova

Wilcoxon y Kruskal-Wallis

```
> u.test() / wilcox.test()
```

Método no paramétrico para comparar la mediana de máximo 2 muestras de distribución no normal.

```
> kruskal.test()
```

Método no paramétrico para comparar la media de 3 o más muestras de distribución no normal.

Fisher’s F-test

```
> var.test() / fisher.test()
```

Evalúa si dos muestras tienen la misma varianza.

Puede analizar tablas de contingencia si los datos son pequeños.

Parecido a fligner.test() y bartlett.test().

Chi-Square

```
> chisq.test()
```

Test de asociación. Identifica diferencias significativas entre grupos categóricos en una tabla de contingencia.

Si la tabla es demasiado pequeña (valores observados inferiores a 5), se recomienda usar el test de Fisher.

How to tell if x, y are independent?

There are two ways to tell if they are independent:

1.By looking at the p-Value: If the p-Value is less that 0.05, we fail to reject the null hypothesis that the x and y are independent. So for the example output above, (p-Value=2.954e-07), we reject the null hypothesis and conclude that x and y are not independent.

2.From Chi.sq value: For 2 x 2 contingency tables with 2 degrees of freedom (d.o.f), if the Chi-Squared calculated is greater than 3.841 (critical value), we reject the null hypothesis that the variables are independent. To find the critical value of larger d.o.f contingency tables, use qchisq(0.95, n-1), where n is the number of variables.

Correlación

```
> cor.test()
```

Test de correlación entre dos variables.

Ho: no existe correlación (independientes).

H1: existe correlación (dependientes).

False Discovery Rate (FDR)

Las correcciones de p-value cuando hacemos un gran número de test estadísticos (por ejemplo, T-Student) **son imprescindibles** para asegurar que nuestro proyecto sea publicado y tenga una calidad decente.

En algunos casos se baja el nivel de significancia (ej. 0.01) y en otros se aplica algunas correcciones como la Bonferroni o la de Benjamini-Hochberg.

https://rpubs.com/Joaquin_AR/236898

Ejercicios

Script de trabajo -> RStatistics.R
Archivo de entrada -> anova-datos.txt

1. Abrir R
2. Cambiar Dir
3. Abrir script RStatistics.R
4. Ejecutar comandos (`Ctrl+R` / `Cmd + R` / Botón Ejecutar)