Laboratorio de R

Curso: Introducción a la Estadística y Probabilidades CM-274

Lecturas Importantes

1. Un tutorial inicial de R, con aspectos fundamentales del lenguaje.
   `http://www.studytrails.com/blog/15-page-tutorial-for-r/`.
2. Para aprender a ser un mejor programador, el libro de Andrew Hunt y David Thomas *The Pragmatic Programmer* es demasiado útil.
3. Notas importantes sobre aleatoriedad y no aleatoriedad en ciencia de datos
   `http://www.kdnuggets.com/2015/10/random-pseudorandom.html`.

Preguntas

1.
   - El conjunto de datos `Orange` es almacenado como un data frame con 3 variables. Indica esas variables.
   - Calcula el promedio de años de los árboles en el conjunto de datos `Orange` usando `mean`.
   - Calcula la mayor circunferencia de los árboles en el conjunto de datos `Orange`.
2. Escribe operaciones en R, para generar cada uno de los siguientes vectores
   - El vector conteniendo los valores $1, -2, 3, -4, \ldots, 99, -100$.
   - El vector conteniendo los primeros 100 valores del factorial.
   - El vector conteniendo las primeras 100 potencias de 2.
3.
   - El conjunto de datos `exec.pay` del paquete `UsingR` es disponible desde la línea de comandos después de cargar el paquete `UsingR`. Carga el paquete y inspecciona el conjunto de datos. Encuentra el mayor valor.
   - Para este conjunto de datos, aplica las funciones `mean`, `min` y `max`. ¿Cuáles son los valores encontrados?.
   - La función `mean` tiene un argumento adicional `trim`. Cuando se da una proporción específica de los datos recorta los datos ordenados antes de que la media es tomada. Compara la diferencia entre `mean(exec.pay)` y `mean(exec.pay, trim = 0.10)`.
4. Los siguientes son una muestra de observaciones sobre la radiación solar entrante en un invernadero:
   11.1 10.6 6.3 8.8 10.7 11.2 8.9 12.2
   (a) Asigna los datos a un objeto `solar.radiacion`.
   (b) Encontrar la media, mediana y la varianza de las observaciones obtenidas sobre la radiación solar.
   (c) Agregar 10 a cada observación de `solar.radiacion` y asigna el resultado a `sr10`. Encontrar la media, la mediana y la varianza de `sr10`. Cuál de las estadística cambia y por cuanto?.
   (d) Multiplica cada observación por -2 y asigna el valor a `srm2`. Encontrar la media, la mediana y la varianza de `srm2`. Como las estadísticas cambian?.
5. Considera el conjunto de datos `islands` y prueba el siguiente código

```
> islands
> hist(log(islands,10), breaks="Scott", axes=FALSE, xlab="area",
+ main="Histograma de Areas de Islas")
> axis(1, at=1:5, labels=10^(1:5))
> axis(2)
> box()
```

(a) Explica que está ocurriendo en cada paso del código de anterior.

6. La función `dim()` devuelve las dimensiones (un vector que tiene el número de filas entonces el número de columnas) de matrices y data frames. Utilice esta función para encontrar el número de filas de los data frames de `tinting`, `possum` y `possumsites` del paquete `DAAG`.

7. La distancia al centro es calculada como $(|x_1 - \overline{x}| + \cdots + |x_n - \overline{x}|)/n$, donde $\overline{x}$ es la media del vector de datos. Calcula este valor para el conjunto de datos `rivers` usando la función `sum` para agregar los valores y `abs` para encontrar el valor absoluto.

8. El conjunto de datos *iris* contiene las medidas de la longitud y el ancho (en cm) de pétalos y sépalos de tres especies: `1: Setosa, 2: versicolor y 3: Virginica`.

   - Considera el objeto *iris*. ¿ Como está estructurado?. ¿ Cuantas observaciones(lineas) contiene?. ¿ Cuantas variables (columnas) contiene?.
   - Para tener una visión general del conjunto de valores, utiliza la función `summary()` del conjunto de dato. ¿Qué información sobre el conjunto de datos proporciona?.
   - Para la variable `Sepal.Length` verifica los resultados dados, usando las funciones `min()`, `max()`, `mean()`, `median()`, `quantile()`. Si es necesario usa la ayuda de `?quantile`.

9. - Escribe código en R que utiliza la función `seq ()` para generar un vector que contiene una secuencia numérica a partir de $0,05$ a $0,2$ en pasos de $0,05$ y asigna el resultado a un objeto llamado `pReg`.
   - Escribe código en R para la siguiente expresión matemática:

$$(1 - pReg)^{40}$$

   - Anote en palabras lo que el resultado del siguiente código en R, muestra (explica que tipo de estructura de datos es creada, que representa cada valor en la estructura)

```
> nJuegos <-seq(20, 40, 5)
> outer(pReg, nJuegos, function(p,n){
+   (1 -p)^n
+   })
```

10. El modelo de `Regresión Lineal Simple` se ajusta a una respuesta y_i mediante una función lineal de una variable predictor x_i.

$$\widehat{y}_i = a + bx_i \text{ para } (i = 1, \ldots, n).$$

Por lo general, los `mínimos cuadrados` son utilizados para estimar los parámetros desconocidos a y b, pero a veces se utiliza la `menor desviación absoluta`. Esto requiere la elección de a y b a fin de minimizar

$$Q(a, b) = \sum_{i=1}^{n} |y_i - \widehat{y}_i|.$$

- Implementa una función que calcule $Q(a, b)$. Debes definir una función de un solo argumento el cúal es un vector cuyos primer elemento es a y el segundo elemento b.

- Explica como usa R la función `optim` para obtener el mejor ajuste de valores de a y b.

11. Trabajar con nombres de archivo en R es fácil, pero requiere el uso adecuado de los separadores de archivos, que varían dependiendo del sistema operativo. Por ejemplo, suponga que tiene el directorio y el nombre de un archivo y desea obtener el archivo completo:

```
> f <- system.file("DESCRIPTION", package="UsingR")
> dname <- dirname(f)
> fname <- basename(f)
```

Para combinar `dname` y `fname` en una ruta completa, usamos `paste` con el argumento `sep` siendo `.Platform$file.sep`. Cuál es el resultado?.

12. Pon a prueba las reglas de coerción mediante la predicción de la salida de los siguientes ejemplos de la función `c()`

```
> c(1, FALSE)
> c("a", 1)
> c(list(1), "a")
> c(TRUE, 1L)
```

13.
- ¿Qué atributos posee un data frame?.
- ¿Se puede tener un data frame con 0 filas?, ¿Qué hay si se tiene 0 columnas?.
- Explica el siguiente código

```
> df <- data.frame(x = 1:3)
> df$y <- list(1:2, 1:3, 1:4)
> df
```

14.
- ¿ Qué ocurre a un factor cuando se modifica sus niveles?

```
> f1 <- factor(letters)
> levels(f1) <- rev(levels(f1))
```

- ¿Qué hace el siguiente código?. ¿ Como difiere $f2$ y $f3$ de $f1$?

```
> f2 <- rev(factor(letters))
> f3 <- factor(letters, levels = rev(letters))
```

15. ¿Como describirias los tres objetos?. ¿ Por qué son diferentes de $1 : 5$?

```
> x1 <- array(1:5, c(1, 1, 5))
> x2 <- array(1:5, c(1, 5, 1))
> x3 <- array(1:5, c(5, 1, 1))
```

16. Esta pregunta es acerca de `vectorización (vectorization)` y `reciclado (recycling)`

- Define que significa que una función R pueda ser vectorizada o que cumple la `vectorization`. Justifica con ejemplos en R.
- Define que significa que una función obedezca la regla de reciclaje. Justifica con ejemplos en R.

17. Supongamos que x es un vector numérico. **Explica en detalle**, como las siguientes expresiones son evaluadas y que valores toman

```
> sum(!is.na(x))
> c(x,x[-(1:length(x))])
> x[length(x) + 1]/length(x)
> sum(x > mean(x))
```

18. La función

```
> f <-function(x,y){
+   if(y > 0)
+     y *sin(x)
+   else
+     x*sin(y)
+ }
```

no soporta el **reciclado**. Explica como puedes modificar la función para que si pueda soportarlo.