En esta charla explicaré el uso de R con ejercicios.

Presentación personal

Satoshi Kusaka

Universidad de Educación de Naruto, Curso de educación global
Catedrático (Educación en Matemáticas)

2002–2004	Enseñanza de matemáticas en la República Dominicana como voluntario para la cooperación japonesa en el extranjero
2004–2005	Maestro de primaria en la Primaria Municipal de Ogino, Atsugi, prefectura de Kanagawa
2008–2011	Maestro de la Corporación Educacional New International School
2011–2021	Participo en proyectos educativos de JICA como experto en educación en matemáticas
2021–actualidad	Catedrático de la Universidad de Educación de Naruto

Pasatiempos:
- ✓ Viajar (ya sea dentro o fuera del país)
- ✓ Estudiar programación (estudiando desde cero)
- ✓ Resolver problemas de matemáticas de secundaria (no me gustaba en esa época, pero ahora sí)

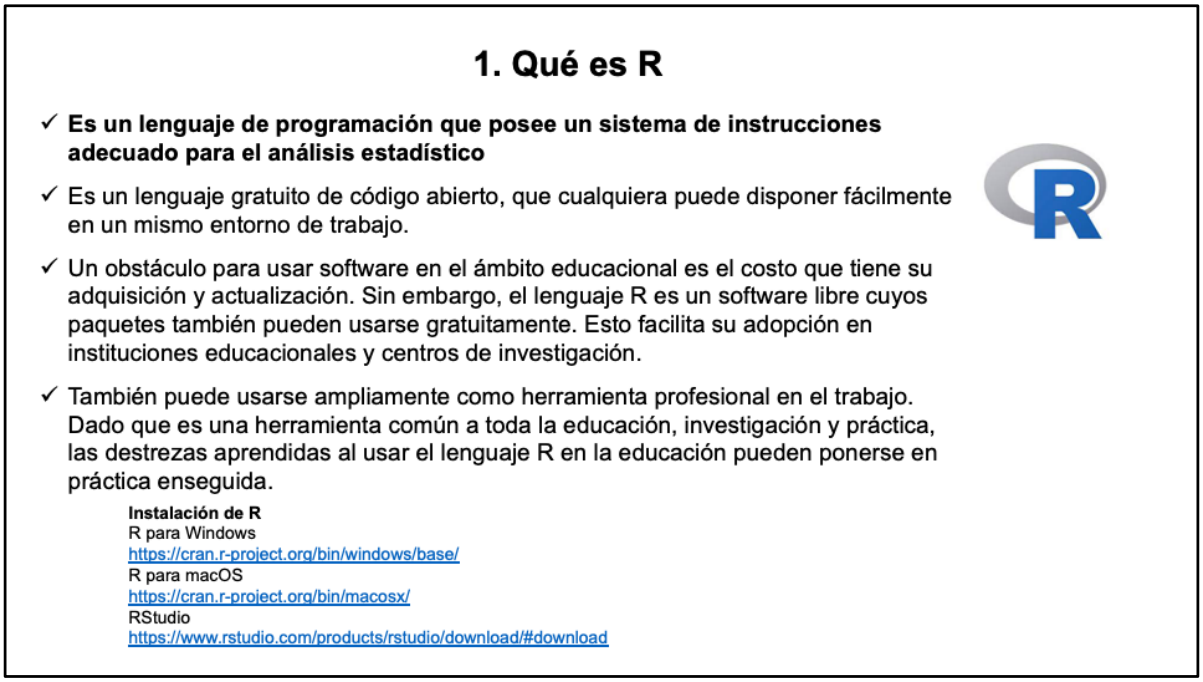
2

Me llamo Satoshi Kusaka, de la Universidad de Educación de Naruto. Mi especialidad es la pedagogía en matemáticas. A la fecha, he sido profesor y he participado como experto en educación en matemáticas en los proyectos educativos de JICA en diversos países. Desde abril de 2021, soy catedrático de la Universidad de Educación de Naruto. Mucho gusto.

Contenido

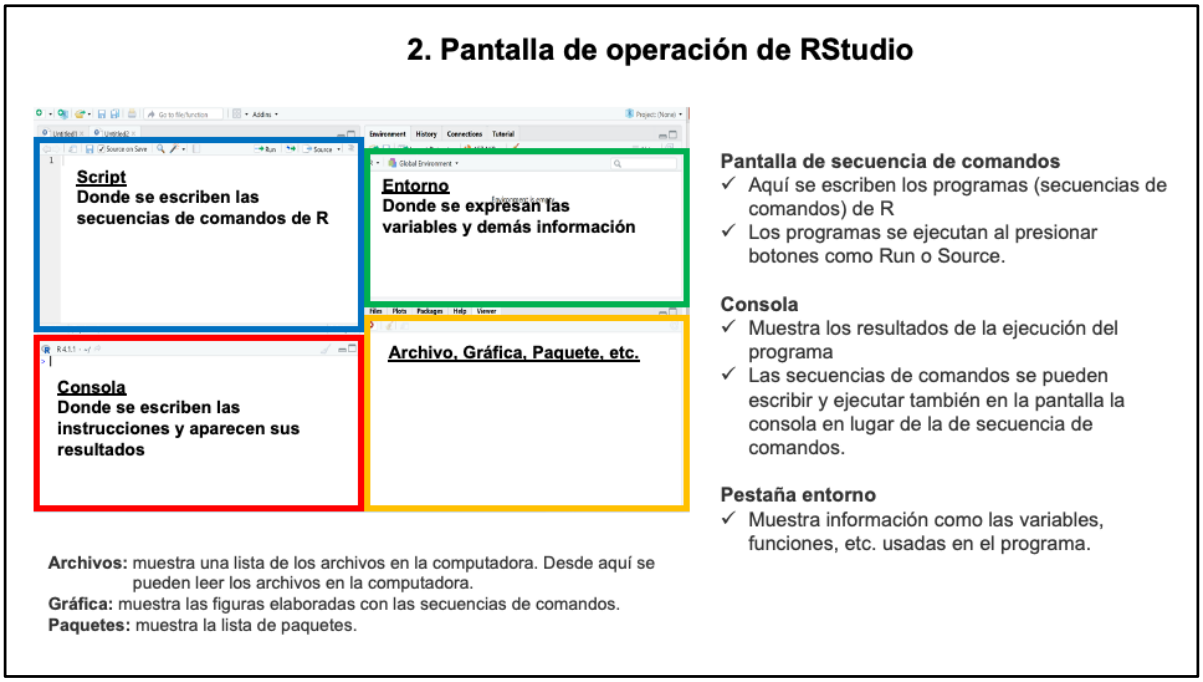
1. Qué es R
2. Pantalla de operación de RStudio
3. Práctica del análisis de datos con R

Primero explicaré brevemente acerca de R. A continuación, vamos a hacer ejercicios de análisis de datos con datos reales.



R es un lenguaje de programación que posee un sistema de instrucciones adecuado para el análisis estadístico. Entre los diversos lenguajes de programación como Python, Java o C++ también está R, que posee un sistema de instrucciones adecuado para el análisis estadístico. Además, es un lenguaje gratuito de código abierto, que cualquiera puede disponer fácilmente en un mismo entorno de trabajo. Un obstáculo para usar software en el ámbito educacional es el costo que tiene su adquisición y actualización. Sin embargo, el lenguaje R es un software libre cuyos paquetes también

pueden usarse gratuitamente. Por eso, las instituciones educativas y centros de investigación pueden adoptarlo fácilmente. R también puede usarse ampliamente como herramienta profesional en el trabajo. Dado que es una herramienta común a toda la educación, investigación y práctica, por usar el lenguaje R, pueden ponerse en práctica enseguida las destrezas aprendidas en la educación.
Aunque creo que todos lo tienen ya instalado, estas son las URL para instalar R.



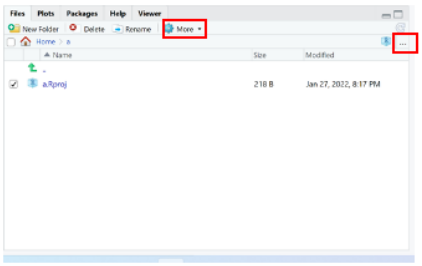
Esta es la pantalla de operación de R. Se divide en 4 grandes secciones. La superior izquierda es la pantalla de secuencia de comandos. Aquí se escriben las secuencias de comandos y al presionar Run, se ejecutan los comandos (instrucciones) señalados. También permite guardar el contenido de las secuencias de comandos aquí escritas. La inferior izquierda es la pantalla de la consola. Muestra los resultados de la ejecución del programa. Las secuencias de comandos se pueden escribir y ejecutar directamente aquí también. La superior derecha muestra lo esencial del entorno del

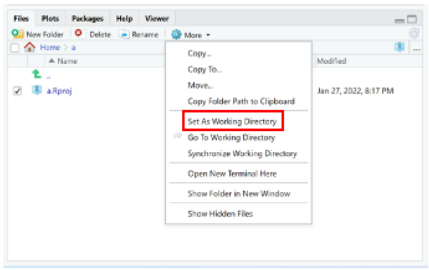
programa en ejecución. Muestra información como las variables, funciones, etc. usadas en el programa. La inferior derecha muestra los archivos en la computadora, la lista de paquetes que se usan en R, etc. Aquí también aparecen las figuras elaboradas con las secuencias de comandos.

3. Ejercicio del análisis de datos con R

1. Configuración del directorio de trabajo

- ✓ Hacer clic en los tres puntos (...) del extremo derecho de la pestaña Files (Archivos) y elegir la carpeta en que se almacenan los datos.
- ✓ Hacer clic en el botón que dice More (Más) y elegir Set As Working Directory (Elegir como directorio de trabajo).





Aquí comenzamos la práctica con datos reales.
En primer lugar, comprobamos de forma sencilla los datos a usar en la práctica. Abramos el archivo R basic. La fila al extremo izquierdo corresponde al ID del alumno. Son los datos de 500 alumnos en total. La fila B es el sexo. Puede tomar dos valores, M o F. C corresponde a Scold (regaño) y expresa en valores el haber sido regañado por el docente. D corresponde a Encouragement (incentivo) y expresa en valores el haber sido incentivado por el docente. E corresponde a Motivation (motivación) y expresa en valores la motivación del alumno. G corresponde al puntaje de la prueba.

Entonces, comencemos con la operación en R. Configuramos el directorio de trabajo. Es necesario hacerlo al inicio cada vez que se trabaja con R. Sirve para indicar los archivos en que se almacenan los datos. Hacer clic en los tres puntos (...) del extremo derecho de la pestaña Files (Archivos) y elegir la carpeta en que se almacenan los datos. Luego, hacer clic en el botón que dice More (Más), elegir Set As Working Directory (Elegir como directorio de trabajo) y la carpeta en que se almacena R basic.

2. Leer los datos

ej<-read.csv("R basic.csv")

3. Confirmación de variables

>colnames(ej)

[1] "ID" "Gender" "Scold" "Encouragement"

[5] "Anxiety" "Motivation" "Test.score"

4. Confirmación de las primeras 4 filas

>head(ej, 4)

ID	Género	Regaño	Incentivo	Ansiedad	Motivación	Test.score
1	M	9	12	16	17	66
2	F	6	6	30	16	68
3	M	7	9	22	19	44
4	M	4	12	23	16	56

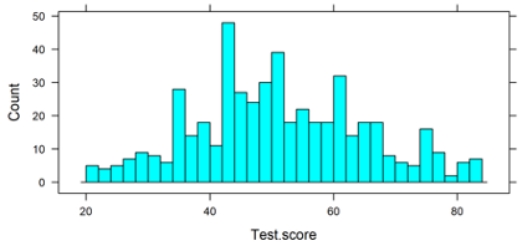
5. Elaboración del histograma

>library(lattice) *# Reading a statistical package*

> histogram(~Test.score, data=ej, breaks=30, type="count")

 # type should be "percent", "count", or "density"

Times New Roman: comandos a escribir
Arial: Resultado
 # Cursiva: explicación adicional



Luego, leemos los datos. Ingresemos el comando ej<-read.csv("R basic.csv"). ej es el nombre de los datos de R basic importados. De aquí en adelante, estos datos son ej y no R basic.

A continuación, comprobemos las variables de los datos importados. Usamos el comando Colnames. El argumento es ej, que es el nombre de los datos. Ingresamos colnames(ej). Debieran aparecer 7 variables.

A continuación, comprobemos las primeras 4 filas de los datos importados. Usamos el comando head. El argumento es (ej,4). Ingresemos >head(ej, 4).

Luego, creamos el histograma. Es necesario haber leído primero el paquete de estadística. Ingresemos >library(lattice). Para trazar el histograma se usa el comando histogram. El argumento es (~Test.score, data=ej, breaks=30, type="count"). Como Breaks y type se pueden cambiar, cambiémoslos para crear varios histogramas.

6. Encontrar tendencias centrales

```
> mean(ej$Test.score)
[1] 51.834
> median(ej$Test.score)
[1] 51
> sort(table(ej$Gender))
  F   M
202 298
```

7. Encontrar la dispersión

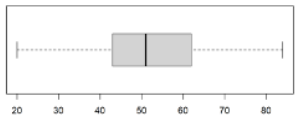
```
> sd(ej$Encouragement)
[1] 2.527548
> var(ej$Encouragement)
[1] 6.388501
```

8. Comparar la dispersión por grupos

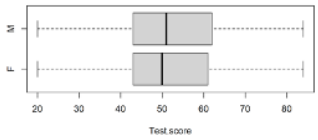
```
> tapply(ej$Test.score, ej$Gender, mean)
# tapply: (quantitative variable, variable of a group,
name of function)
       F        M
51.71287 51.91611
```

9. Dibujar diagramas de cajas

```
>boxplot(ej$Test.score, horizontal=TRUE)
  # drawing vertically: horizontal=FALSE
```




```
>boxplot(Test.score~Gender,data=ej,horizontal=TRUE)
```



10. Resumen de las estadísticas

```
> summary(ej$Motivation)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.00   12.00   14.00   13.98   17.00   20.00
```

A continuación, calculamos los valores del promedio, la mediana, etc. Ingresemos los comandos indicados para calcular el promedio y la mediana. Si queremos clasificar los datos de dos valores, usamos sort. Como el sexo toma dos valores, clasifiquémoslo en femenino y masculino.

A continuación, calculamos la dispersión. Calculamos la desviación estándar y la varianza, que corresponde al cuadrado de la desviación estándar. Se pueden calcular con los comandos sd y var respectivamente. Aquí tomamos como ejemplo la columna de valores de Encouragement, pero pueden

probar a calcularlas también para las otras variables.
A continuación, comparamos los valores representativos para cada grupo. Usamos el comando tapply. Como argumentos, ingresamos primero «nombre de variable a usar», luego «grupo» y finalmente el «valor representativo que deseamos comparar». Aquí usamos el promedio (mean). Pruebe con los valores representativos de las otras variables.
A continuación, trazamos un diagrama de cajas. Se usa el comando boxplot. Indicamos la variable de boxplot que queremos trazar y si es en vertical u horizontal. Ingresamos horizontal=FALSE si lo queremos trazar en vertical. Si lo queremos separar por sexo, lo separamos con una tilde (~) e ingresamos Gender.
Después, calculamos de una vez los valores representativos de las variables. Se usa el comando summary.

11. Prueba F de homogeneidad de la varianza

```
> var.test(Scold~Gender, data=ej)

        F test to compare two variances

data: Scold by Gender
F = 0.82005, num df = 201, denom df = 297, p-value = 0.1297
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6382503 1.0602887
sample estimates:
ratio of variances
         0.8200534
```

12. Prueba T de muestras independientes (homogeneidad de la varianza)

```
> t.test(Scold~Gender, data=ej, var.equal=TRUE)

        Two Sample t-test

data: Scold by Gender
t = -1.052, df = 498, p-value = 0.2933
alternative hypothesis: true difference in means between group F
and group M is not equal to 0
95 percent confidence interval:
 -0.6357685  0.1923503
sample estimates:
mean in group F mean in group M
       5.113861        5.335570
```

13. Prueba T de Welch (No de homogeneidad de la varianza)

```
> t.test(Scold~Gender, data=ej,var.equal=FALSE)

        Welch Two Sample t-test

data: Scold by Gender
t = -1.0722, df = 458.58, p-value = 0.2842
alternative hypothesis: true difference in means between group F and
group M is not equal to 0
95 percent confidence interval:
   -0.6280515  0.1846334
   sample estimates:
   mean in group F mean in group M
          5.113861        5.335570
```

14. Prueba T pareada

```
> score<-c(ej$Encouragement, ej$Motivation)
> result<-c(rep("Encouragement",500), rep("Motivation", 500))
> t.test(score~result, paired=TRUE)

        Paired t-test

data:  score by year
    t = -26.264, df = 499, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
 95 percent confidence interval:
  -5.165517 -4.446483
   sample estimates:
    mean of the differences
        -4.806
```

Aquí entramos a evaluar las diferencias estadísticamente significativas. Consiste en una evaluación para ver si las diferencias en los resultados son estadísticamente significativas o no; es decir, que demuestran que no son resultados obtenidos por casualidad. Esto es sumamente importante en el análisis estadístico.

Primero, realizamos una prueba f. Es una prueba para averiguar si la dispersión es equivalente. La hipótesis a probar es la hipótesis nula: «no hay diferencia de la dispersión del regaño por sexo». El comando es var.test y como comparamos Scold entre

sexos, los argumentos son (Scold~Gender, data=ej). El valor p es 0,1297. Como es mayor que 0,05 no se puede descartar la hipótesis nula. Es decir, no podemos decir que haya diferencias estadísticamente significativas entre la varianza de ambos grupos y podemos considerar que la varianza es homogénea.

Como podemos considerar que hay homogeneidad en la varianza de hombres y mujeres en Scold, realizamos la prueba t para ello. Aquí la hipótesis nula es: «no hay diferencia en la dispersión del regaño por sexo». Se usa el comando t.test. El argumento es (Scold~Gender, data=ej, var.equal=TRUE). Como vemos que el valor P es mayor que 0,05, no podemos descartar la hipótesis nula. Es decir, determinamos que las diferencias en los promedios entre hombres y mujeres no son estadísticamente significativas.

Luego, realizamos la prueba t para cuando las dispersiones son equivalentes. Con el mismo comando y argumentos, ponemos FALSE en var.equal.

Después, realizamos la prueba t pareada usando otras variables. La prueba t pareada es una prueba t en la que los datos a comparar están emparejados. En la prueba t anterior los datos no estaban pareados, ya que se comparaban hombres y mujeres. Aquí comparamos Incentivo y Motivación. Podemos decir que los datos están emparejados, porque se

trata del incentivo y la emoción para todas las mismas 500 personas. Juntamos Incentivo y Motivación como un puntaje. Luego, emparejando con el puntaje, usamos el comando rep y extraemos los 500 datos de Incentivo y los 500 datos de Motivación.
Como el valor p es 2,2e-16, podemos decir que la diferencia entre Motivación e Incentivo es estadísticamente significativa. Como los datos de Incentivo y Motivación en que se basa el puntaje son completamente distintos, es obvio que así sea.

Bibliografía

川端一光、岩間徳兼、鈴木雅之(2018).「Rによる多変量解析入門 データ分析の実践と理論」オーム社