Manejo práctico del software de análisis estadístico R

Manejo práctico del software de análisis estadístico R

Armando Cervantes Sandoval

Xavier Chiappa Carrara

Maite Mascaró Miquelajauregui

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Facultad de Estudios Superiores Zaragoza

UMDI-Sisal, Facultad de Ciencias

México 2009

Primera edición: 2009

Impreso y hecho en México

Formación, Diseño editorial y portada: Armando Cervantes Sandoval

Desarrollado con apoyo de los proyectos PAPIME PE201106 y PE101606

Contenido

Capítulo 6

Presentación

Este material es una guía práctica para realizar análisis estadístico de datos, utilizando el software de análisis estadístico R. Con ese fin se hace una revisión del entorno y el lenguaje en un formato de curso, donde se describe a detalle cómo utilizar y aplicar algunas de las muchas opciones disponibles, quedando una gran cantidad de ellas todavía por analizar.

Utilizando el "paquete" RCmdr, una herramienta que genera una interfaz de usuario tipo Windows, nos enfocamos al uso del software y en cada sección se indica cómo realizar el análisis correspondiente. A través de ejemplos se dan algunos criterios que apoyan la interpretación de resultados.

R es un conjunto integrado de programas, enfocado principalmente al análisis estadístico de datos, con grandes facilidades para la manipulación de datos, cálculo y gráficos.

Un aspecto fundamental es que R es un software de uso libre, cuya consolidación y desarrollo se está dando por la activa participación de la comunidad estadística mundial. Pero sobre todo, es una oportunidad para dejar de ser usuarios "piratas" del cada vez más costoso software comercial de análisis estadístico.

Capítulo 1

Dónde adquirir y cómo se instala R

1.1. Introducción

R es una versión de uso libre del lenguaje S y consiste en un ambiente para el cómputo de análisis estadísticos. Con un lenguaje de programación propio, proporciona una amplia variedad de técnicas gráficas y estadísticas. Ofreciendo a los usuarios avanzados un entorno de programación muy completo que les permite agregar nuevas herramientas mediante la creación de nuevas funciones. S y R han cambiado la forma de analizar, visualizar y manejar los datos. Lo que probablemente los hacen los dos lenguajes más utilizados en la investigación estadística.

- R es un GNU, que se puede distribuir con licencia GLP o General Public (la filosofía y objetivos del proyecto GNU se pueden ver en www.gnu.org.
- R y los diferentes paquetes se obtienen por 0 euros en http://cran.r-project.org

1.2. Obtener e instalar R

Según el sistema operativo (Linux, Windows o Macintosh), todo el software que conforma a R, se puede obtener en:

http://cran.r-project.org/bin

Para Windows, se puede "bajar" el ejecutable (download), desde:

http://cran.r-project.org/bin/windows/base

Para este curso se obtuvo el archivo:

http://cran.r-project.org/bin/windows/base/R.2.9.2-win32.exe

el cual es un programa ejecutable que instalará el sistema y los paquetes base, para que funcione el R.

Instalar el sistema básico de R, en ambiente Windows, requiere los siguientes pasos:

1. Bajar de la red el archivo R.2.9.2.-win32.exe.
2. Ejecutar el archivo.

Además del sistema base hay un conjunto de paquetes que extienden la funcionalidad de R, los cuales se describen en la sección 1.4.

1.3. Documentación de R

1.3.1. En la instalación de R se incluyen los siguientes manuales

1. An introduction to R, cuya lectura es altamente recomendable
2. Writing R extensions
3. R data import/export

4. The R language definition
5. Installation and administration

1.3.2. Se puede encontrar ayuda en una gran cantidad de sitios como

http://cran.r-project.org/doc/contrib/Burns-unwilling_S.pdf o

http://www.burns-stat.com/pages/tutorials.html. ¡Con sólo 8 páginas!

NOTA: Hay que recordar que R tiene como origen S, así que todo lo que funcione en S, funciona también para R.

1.3.3. R para principiantes, de E. Paraadis, en:

http://cran.r-project.org/other-docs.html o

http://cran.r-project.org/doc/contrib/rdebuts_es.pdf.

1.3.4. El sitio de Paul Johnson

http://lark.cc.ukans.edu/~pauljohn/R/statsRus.html

1.3.5. Algunos otros como el sitio de J. Faraway, de Practical Regresión and ANOVA using R, en:

http://cran.r-project.org/other-docs.html o

http://www.stat.lsa.umich.edu/~faraway/book/.

Una búsqueda en Google puede proporcionar una amplia gama de posibilidades de manuales, desde los más sencillos hasta aquellos que describen técnicas muy complejas o especializadas.

1.4. Instalación de paquetes adicionales

La instalación depende del sistema operativo en el que se esté trabajando, para el caso de Windows, ésta se puede realizar desde el CRAN (seleccionando un sitio espejo), aunque es más recomendable "bajar" los archivos comprimidos desde un sitio espejo y trabajar con la opción **Install package(s) from local zip files** (figura 1.1.).

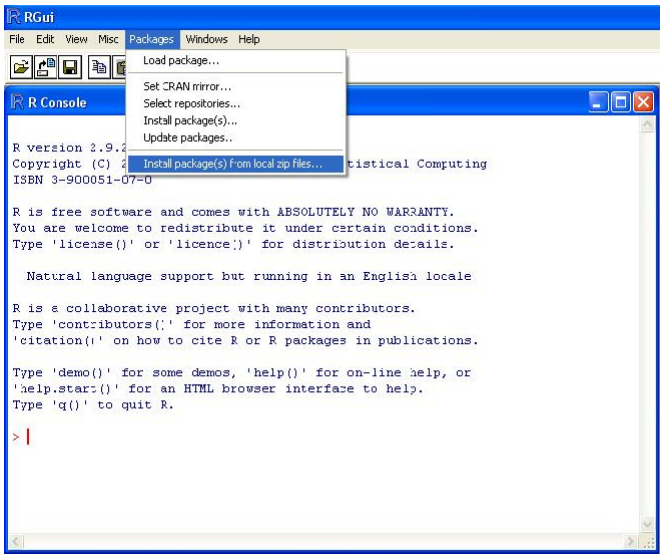


Figura 1.1. Ventana de instalación de paquetes.

El primer paquete que se debe "bajar" e instalar es el Rcmdr, una vez instalado se "carga" (Load) de la lista de paquetes disponibles, como se muestra en la figura 1.2.

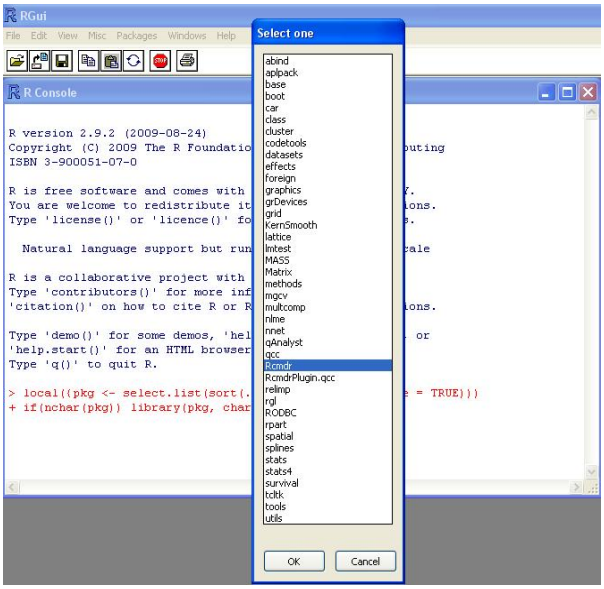


Figura 1.2. Cargar Rcmdr.

Al momento de cargar Rcmdr se despliega una lista de todos los paquetes R que se necesitan para su adecuado funcionamiento, los cuales se deben bajar e instalar siguiendo el procedimiento ya descrito.

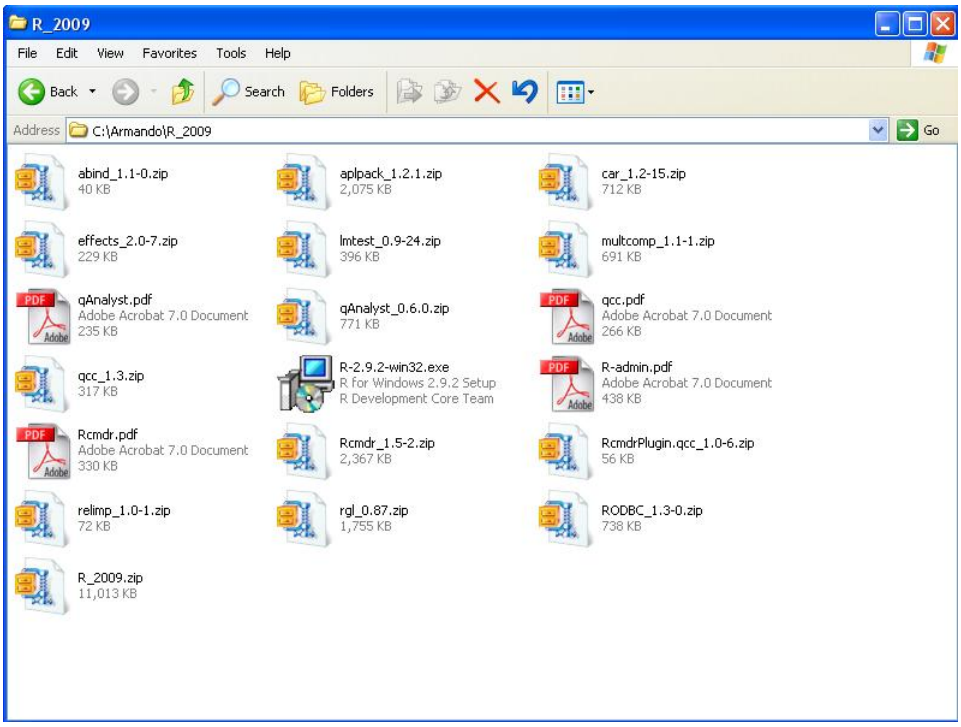


Figura 1.3. Lista de posibles "paquetes" a instalar.

Para facilidad del lector y de los potenciales usuarios de R, en el sitio enlinea.zaragoza.unam.nx/biomat, hay un vínculo a un archivo zip (de 11 MBytes) que contiene todos los paquetes que requiere Rcmdr para funcionar adecuadamente. No se "puso" el ejecutable de Windows, ya que su tamaño es de 35 Mbytes.

Capítulo 2

Entorno Rcmdr y manejo de datos

El primer paso es entrar a R, como se hace con cualquier programa en Windows. Un doble clic y listo; o ir a Inicio-Programas-R.

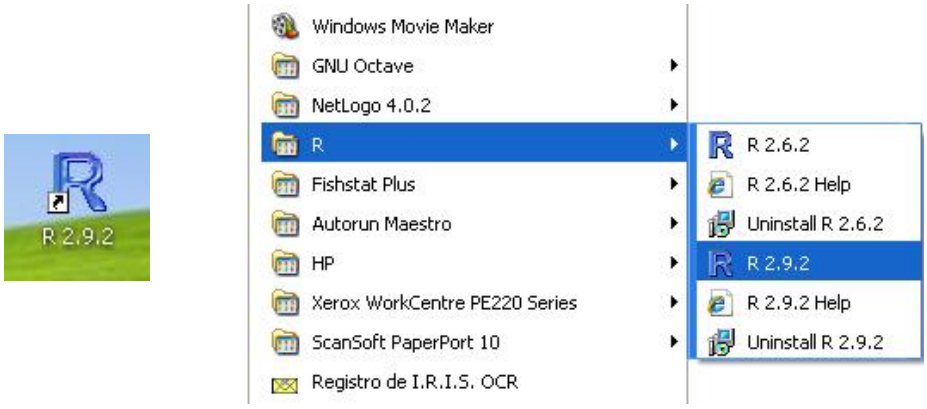


Figura 2.1. Opciones de entrada a R.

En ambos casos se despliega la misma ventana de entrada. En la parte inferior se aprecia el entorno de programación instrucción por instrucción. Pero también se puede seleccionar del menú la opción: **File -> new script** y aparece un editor donde se puede "teclear" todo un programa y probar su correcta ejecución, figura 2.2,.

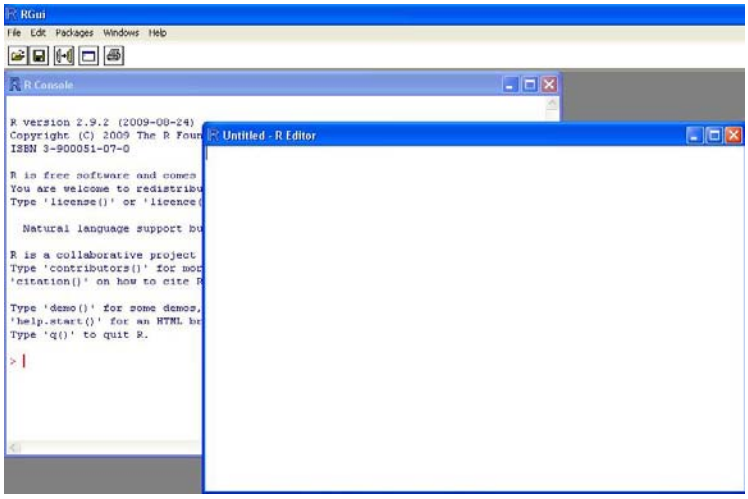


Figura 2.2. Ambiente de trabajo R.

Esa opción se revisará en el último capítulo, por el momento de la barra de menú se selecciona la opción:

Packages -> Load packages y seleccionar de la lista que aparece el Rcmdr.

Al cargar Rcmdr aparece el correspondiente entorno de trabajo, un conjunto de paquetes que permite el análisis de datos en un entorno tipo Windows, generalmente la versión más actualizada está en idioma inglés, pero es posible conseguir versiones anteriores en español.

Rcmdr presenta tres áreas de trabajo, figura 2.3.

- La superior para instrucciones de programación. Que conviene revisar constantemente para familiarizarse con la programación en R.
- La intermedia de resultados
- La inferior de mensajes

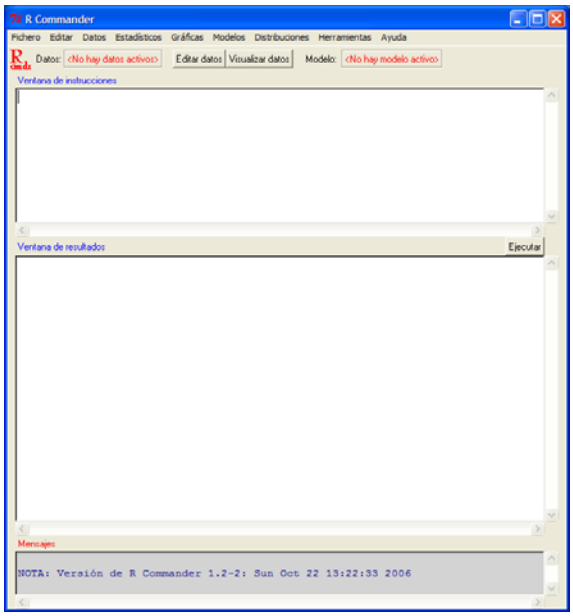


Figura 2.3. Entorno Rcmdr

Aquí se van a revisar sólo algunas de las opciones que aparecen en el menú de Rcmdr. Para lo cual se trabajan algunos ejemplos, que muestran como ingresar los datos y trabajar las diferentes opciones de análisis.

2.1. Ejemplos para manejo de datos

Se tiene la siguiente distribución de alumnos en 7 carreras:

Carrera	Número de alumnos
Biología	700
Enfermería	850
Ingeniería Química	400
Medicina	1600
Odontología	1700
Psicología	1800
Q.F.B.	1130

Para generar el archivo se deben considerar dos variables: una llamada **Carrera** de tipo **Character** y otra de nombre **Alumnos** de tipo **Numeric**, entonces se tienen 2 columnas con 7 datos cada una.

Lo cual se logra con la opción del menú: **Datos -> Nuevos datos**

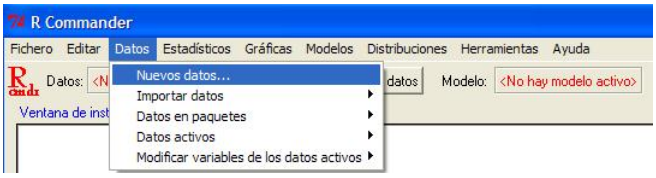


Figura 2.4. Creando archivo de datos.

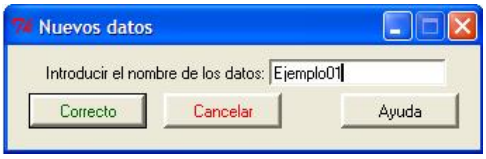


Figura 2.5. Nombrar el archivo de datos, se utiliza para distinguir y definir el conjunto de datos activos.

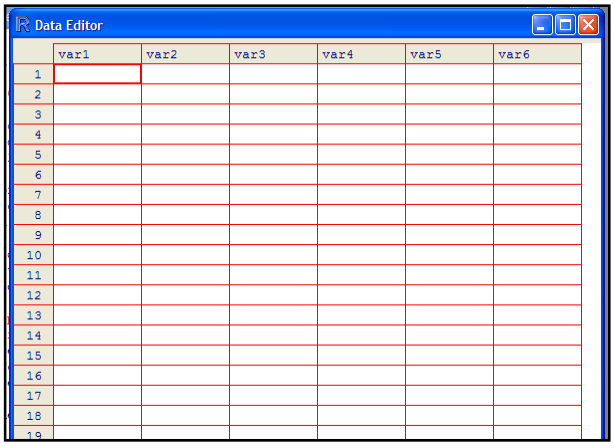


Figura 2.6. Hoja de datos, con formato de una hoja de cálculo.

Los datos se "teclean" directamente en la hoja de datos y el nombre de las variables y tipo de datos se modifican siguiendo la secuencia.

1. Dar un clic sobre el identificador de la columna (var1, var2, . . .), y en la caja de diálogo que aparece "teclear" el nombre de la variable y el tipo de datos que almacena.

Después de eso hay que ingresar los datos y almacenarlos en disco, de otra manera habría que "teclearlos" cada vez que se vaya a utilizar.

La secuencia para hacerlo es:

Datos -> Datos activos -> Exportar los datos activos

Y después seguir los diálogos que se despliegan hasta tener los datos en un archivo tipo texto, en la carpeta seleccionada.

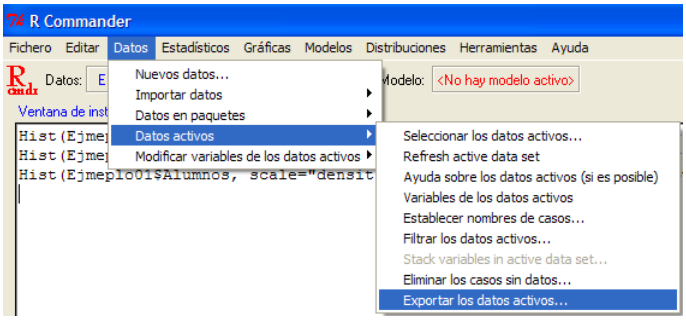


Figura 2.7. Guardar datos en disco, como archivo texto (tipo .txt).

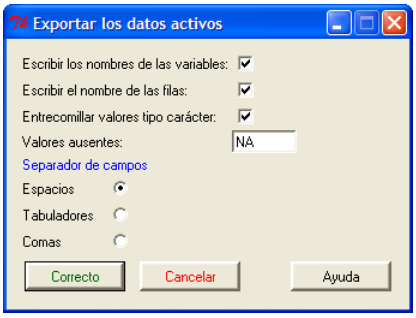


Figura 2.8. Diálogo para establecer la forma en que se exportan los datos.

Si los datos ya están guardados, el siguiente paso es recuperarlos al entorno de Rcmdr, esto se logra con la secuencia:

Datos -> Importar datos -> from text file or clipboard

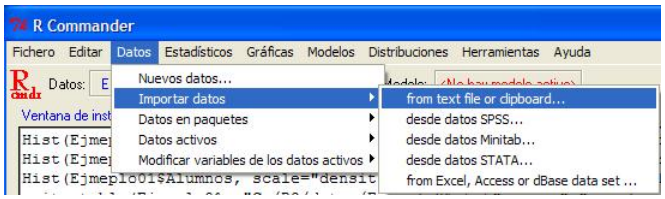


Figura 2.9. Importar datos a Rcmdr.

El siguiente paso es seguir los diálogos para ubicar el archivo texto que previamente se exportó y que ahora se va a importar. Al tener los datos en el entorno de Rcmdr, ya se puede empezar a trabajar con ellos.

Otra opción más práctica es tener los **datos en Excel e importarlos a Rcmdr**, con la secuencia que se presenta de la figura 2.10. a la 2.1.3.

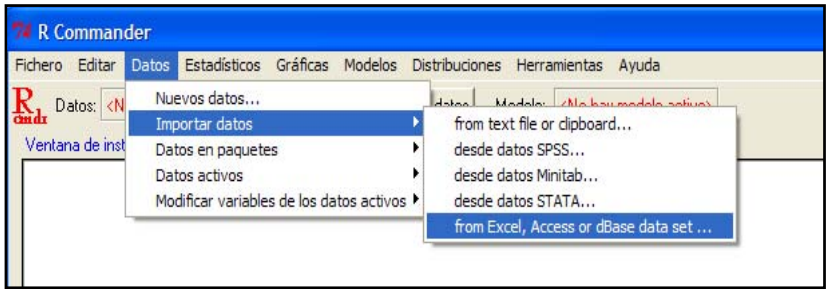


Figura 2.10. Secuencia para importar datos desde Excel.

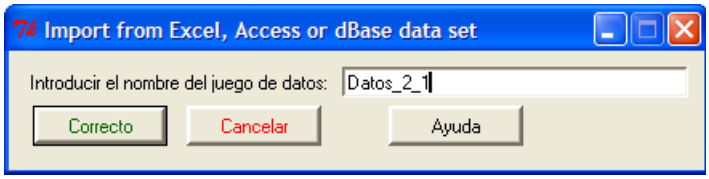


Figura 2.11. Nombre que identifica el archivo de datos, sólo dentro de R.

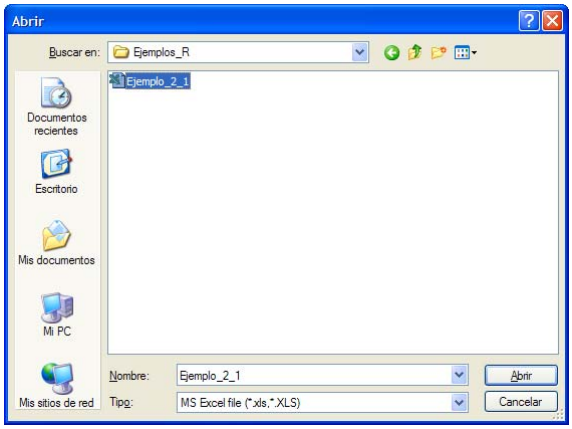


Figura 2.12. Seleccionar el archivo a importar.

Figura 2.13. Se debe seleccionar la hoja en que vienen los datos, que generalmente es la hoja1.

Los datos se pueden revisar con el botón <editar datos> o <visualizar datos>, en cuyo caso es importante "cerrar" la ventana de datos antes de realizar cualquier otra actividad, de lo contrario da la sensación de que R se "pasma".

Para empezar a trabajar con R, se van a realizar unos gráficos de histograma. Por lo que se define la secuencia: **Gráficas -> Histogramas**, y aparece una caja de diálogo para definir la o las variables a trabajar.

Figura 2.14. Elegir las variables para realizar el histograma.

El resultado se presenta en la figura 2.15.

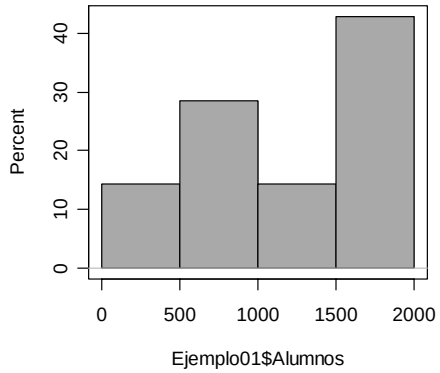


Figura 2.15. Resultado del histograma.

Para el reporte de resultados se recomienda abrir un archivo Word y mediante corte (copia) y pega elaborar un archivo con los resultados. Además de que los gráficos se pueden guardar como archivos de imagen o copiar directamente a Word o en cualquier procesador de texto, para ver las opciones dar un clic derecho en el gráfico.

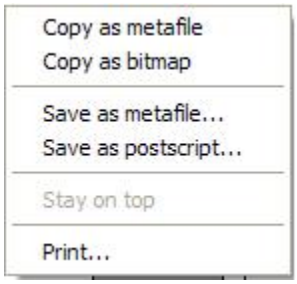


Figura 2.16. Opciones para guardar imágenes.

Capítulo 3

Exploración de datos

En el análisis estadístico la primera actividad a realizar es ver los datos, observarlos o explorarlos. Apreciar su distribución, agrupamiento, dispersión o la presencia de valores extremos. Como se dice: dejar que los datos hablen.

Este primer paso se basa en la exploración gráfica de los datos, para contestar preguntas como:

1. ¿Cuál es el valor del dato central (promedio)?
2. ¿Cómo es su variabilidad?
3. ¿Son simétricos, sesgados, bimodales?
4. ¿Tienen outliers (observaciones extremas)?
5. En caso de dos variables, ¿hay alguna relación entre ellas?, ¿Esta relación es lineal?
6. ¿Es necesario transformar los datos?

Las herramientas básicas son algunos gráficos de uso común. Como los de caja y bigote, los de tallo y hojas, así como los clásicos histogramas.

3.1. Diagrama de caja y bigote o boxplot

Estas gráficas se han vuelto muy populares, ya que ofrecen mucha información de manera compacta. Muestran el rango de los datos, la dispersión a través del rango intercuartílico y la mediana como medida de tendencia central.

Pasos para construir un BoxPlot

1. Calcular los cuartiles Q_1, Q_2 y Q_3
2. Sobre una línea, horizontal o vertical, "pintar": el valor mínimo; Q_1; Q_2; Q_3 y el valor máximo
3. Hacer un rectángulo de Q_1 a Q_3
4. Trazar una línea en Q_2=mediana
5. Revisar que los valores extremos no estén a una distancia mayor a 1.5 el valor del rango intercuartílico, si hay algún valor marcarlo.

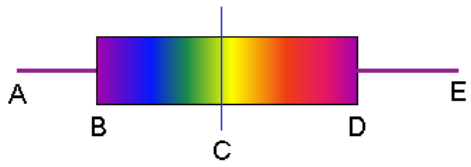


Figura 3.1. Representación de un Boxplot

3.2. Diagramas de tallo y hojas

Estos diagramas fueron desarrollados por John Tukey en 1977. Permiten observar la distribución de los datos originales y son muy útiles para resumir y describir, sobre todo cuando no se rebasan los cien datos.

Para construir un diagrama de tallo y hoja:

1. Colocar a la izquierda los dígitos más significativos del dato (Tallo)
2. Colocar a la derecha los dígitos menos significativos, en orden de menor a mayor, unidades o decimales, (Hojas). En algunos casos conviene poner en las hojas dos dígitos significativos.
3. Hacer un conteo de la frecuencia de valores asociados al valor del tallo.

3.3. Histogramas

Un histograma es un gráfico de barras que muestra la frecuencia de cada uno de los valores encontrados en la variable medida (número de veces que se repite un valor). En términos simples, consta de un eje horizontal cuya escala va desde el valor más pequeño hasta el valor máximo en los datos, valores que de preferencia deben ser cuantitativos y continuos (resultado de mediciones de peso, longitud, volumen, etc.); y de un eje vertical cuya escala puede ir desde cero hasta la máxima frecuencia encontrada.

Para elaborar un histograma se recomienda:

1) Obtener el valor máximo y el mínimo, de todo el conjunto de valores.

2) Escribir cada uno de los valores, en columna y en orden ascendente.

3) Revisar todos los valores del conjunto total de datos y colocar una marca, al frente de cada valor, por cada vez que se repita.

4) Contar el número de marcas en cada uno de los valores, anotándolo en la fila correspondiente.

Las diferentes herramientas exploratorias están encaminadas a responder estos diferentes cuestionamientos y pueden ser usadas para validar análisis estadísticos subsecuentes.

Gráficos como los histogramas de frecuencia muestran, tanto el valor central como la distribución de los datos y ~~nos~~ proporcionan una idea de la normalidad de la variable.

Un aspecto importante en la exploración de datos es la identificación de los outliers (casos extraordinarios), que son datos anómalos en comparación con el resto del conjunto de datos y pueden influir en el análisis. Así el primer cuestionamiento sería como se pueden detectar los outliers?. Una de las herramientas gráficas que apoyan esta labor es el diagrama de cajas, aspecto en el que se centra el siguiente ejemplo y los ejercicios de este capítulo.

3.4. Ejemplo en R

Peso en Kg de una muestra al azar de 50 estudiantes de una universidad, tomados de sus registros médicos.

Género	M	M	M	F	F	M	F	M	M	F
Peso	89	93	96	64	68	98	69	102	95	60

Género	F	M	F	M	M	F	F	M	M	F
Peso	49	75	54	79	81	56	59	84	85	60

Género	M	M	F	M	F	F	M	M	M	M
Peso	94	88	59	84	58	58	81	79	77	74

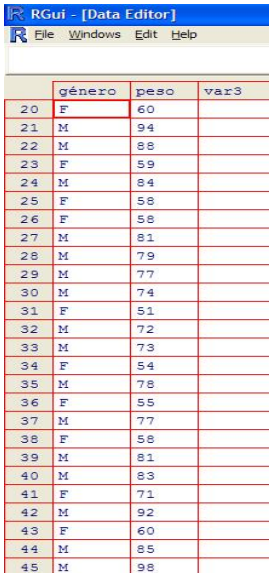
Género	F	M	M	F	M	F	M	F	M	M
Peso	51	72	73	54	78	55	77	58	81	83

Género	F	M	F	M	M	F	F	M	M	M
Peso	71	92	60	85	98	66	65	108	88	92

Actividades

1. Generar el nuevo conjunto de datos.
2. Definir el nombre y tipo de variables.
3. "teclear" los datos.

Para generar el archivo se deben considerar dos variables: una llamada **Género** de tipo **Character** y otra de nombre **Peso** de tipo **Numeric**, entonces se tienen 2 columnas con 50 datos cada una.



	género	peso	var3
20	F	60	
21	M	94	
22	M	88	
23	F	59	
24	M	84	
25	F	58	
26	F	58	
27	M	81	
28	M	79	
29	M	77	
30	M	74	
31	F	51	
32	M	72	
33	M	73	
34	F	54	
35	M	78	
36	F	55	
37	M	77	
38	F	58	
39	M	81	
40	M	83	
41	F	71	
42	M	92	
43	F	60	
44	M	85	
45	M	98	

Figura 3.2. Tabla de datos.

Otra opción para ingresar los datos es:

1) Teclearlos en Excel.
2) Seleccionar los datos y hacer una copia en Excel.
3) En Rcmdr, importar datos y definir que vienen del clipboard.

Una vez que ya se tienen los datos, no hay que olvidar guardar el archivo. Y ahora si ya se tienen los datos listos para análisis

Vamos a realizar las siguientes secuencias:

Gráficas -> Diagrama de cajas -> Gráficas por grupos

Cuyo resultado se muestra en el siguiente gráfico.

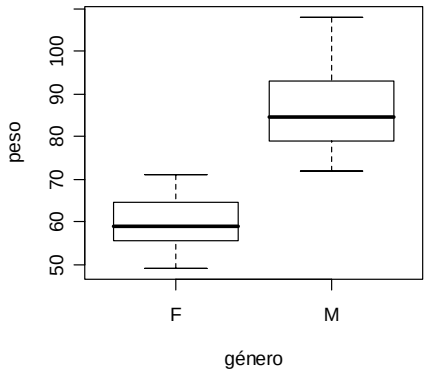


Figura 2.17. Boxplot del ejemplo.

Qué se puede decir de este gráfico, mejor vamos a formalizarlo mediante preguntas.

1) Quiénes pesan más en promedio, los hombres o las mujeres
   R = sexo masculino
2) En que datos hay más variación
   R = en los masculinos
3) Que tan diferente es la variación en los datos femeninos por arriba y por debajo de la caja
   R = Los femeninos son simétricos con respecto a la caja
4) Que tan diferente es la variación en los datos masculinos por arriba y por debajo de la caja.
   R = Los masculinos son más variables por "arriba" de la caja (hay un cierto sesgo).

Ahora, pasemos a los números, con la secuencia:

Estadísticas -> Resúmenes -> Resúmenes numéricos

Cuyos resultados son:

	mean	sd	0%	25%	50%	75%	100%	n
F	59.70000	5.939165	49	55.75	59.0	64.25	71	20
M	86.03333	9.208255	72	79.00	84.5	92.75	108	30

Aprovechando estos datos aplicar las secuencias

Estadísticos -> Medias -> Prueba t para muestras independientes

Obteniendo los resultados:

```
Welch Two Sample t-test
data:  peso by género
t = -12.2912,  df = 47.973,  p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -30.64108  -22.02559
sample estimates:
mean in group F  mean in group M
       59.70000         86.03333
```

Estadisticos -> Varianzas -> Prueba F para dos varianzas

Cuyos resultados son:

```
> tapply(Datos$peso, Datos$género,  var, na.rm=TRUE)
      F                 M
35.27368         84.79195
          F test to compare two variances
data:  peso by género
F = 0.416,  num df = 19,  denom df = 29,  p-value = 0.0498
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1864418       0.9992147
sample estimates:
ratio of variances
        0.4160027
```

3.5. Ejercicios del capítulo 3

- Utilizando los datos, de los siguientes ejercicios, explorar las opciones gráficas y descriptivas de Rcmdr. Haciendo gráficos Boxplot; de Tallo y hojas e Histogramas.

- Seguir toda la secuencia de análisis, más algunas que Ud. explore y considere que mejoran y facilitan el análisis estadístico de los siguientes datos.

3.5.1. En el primer día de clases se les preguntó a 50 estudiantes acerca del tiempo requerido para desplazarse de su casa a la universidad (redondeado a 5 minutos). Los datos resultantes son:

30	40	35	50	50	55	50	35	35	60
45	45	20	50	55	70	45	60	30	45
35	50	20	45	75	55	55	45	35	30
25	40	25	45	50	55	40	40	40	40
15	40	25	35	35	50	50	30	50	35

¿Qué se puede decir de los tiempos de desplazamiento? Sugerencia analizar con base en dispersiones y medidas de tendencia central.

3.5.2. Se toma una muestra de 50 calificaciones de una población de resultados de un examen final de estadística. Estos datos se muestran en la siguiente tabla.

75	97	71	65	84	27	99	91	99	82
96	58	94	43	10	10	91	10	94	43
74	73	68	54	50	49	81	10	97	76
10	94	79	80	82	71	88	88	47	73
71	99	86	10	84	93	77	98	44	10

¿Qué tan buenas son las calificaciones y qué se puede decir de este grupo de alumnos?

3.5.3. Se tienen los siguientes datos de emisiones de óxido de azufre, en toneladas.

(modificado de Estadística Elemental, John E. Freund y Gary A. Simon, 1992, Prentice Hall, pp. 21-22).

Planta industrial A

15.8	22.7	26.8	19.1	18.5	14.4	8.3	25.9	26.4	9.8
22.7	15.2	23.0	29.6	21.9	10.5	17.3	6.2	18.0	22.9
24.6	19.4	12.3	15.9	11.2	14.7	20.5	26.6	20.1	17.0
22.3	27.5	23.9	17.5	11.0	20.4	16.2	20.8	13.3	18.1

Planta industrial B

27.8	29.1	23.9	24.4	21.0	27.3	14.8	20.9	21.7	15.8
18.5	22.2	10.7	25.5	22.3	12.4	16.9	31.6	22.4	24.6
16.5	27.6	23.0	27.1	12.0	20.6	19.7	19.9	26.5	21.4
28.7	23.1	16.2	26.7	13.7	22.0	17.5	21.1	34.8	31.5

a. "Teclear" y guardar los datos en un archivo R.

b. Realizar un análisis descriptivo de los datos y comparar las dos plantas.

c. Guardar los resultados del análisis en un archivo Word

Capítulo 4

Inferencia estadística

La inferencia estadística se caracteriza porque a través de una muestra se pueden realizar inferencias de toda una población en estudio. De manera que utilizando modelos estadísticos se puede asignar un nivel de confiabilidad a las conclusiones que se obtengan, proporcionando soporte para la toma de decisiones.

4.1. Población y muestra

En cualquier proceso de investigación o producción es demasiado costoso, en recursos o en tiempo, revisar uno a uno todos los elementos que conforman una población, de ahí la necesidad de revisar unos cuantos, que sean representativos, y a partir de ellos predecir el comportamiento de toda la población.

El primer "viaje" a la estadística implica seleccionar una muestra de manera aleatoria, es decir, sin privilegiar o descartar de antemano elemento alguno; garantizando que todos tengan la misma posibilidad de ser elegidos. La mejor forma de hacer esto es utilizando

herramientas como tablas de números aleatorios, una urna, o algún proceso de números pseudoaleatorios como los que vienen integrados las calculadoras y en la mayoría de los paquetes estadísticos. Cualquiera de estas opciones es mejor que cerrar los ojos y estirar la mano o establecer criterios personales de selección de muestras.

Uno de los ejemplos más simples, pero nada estadístico, es lo que hacen quienes cocinan ya que a través de pequeñas "probadas" saben si un guiso está o no en su punto, esto previa homogenización del contenido de la cazuela y sin consumir todo su contenido.

Es conveniente aclarar que el tema de muestreo es una de las grandes ramas de la estadística, para la cual existen libros completos que analizan a detalle cada una de las opciones, dependiendo del propósito del muestreo.

El segundo "viaje" a la estadística consiste en analizar la muestra mediante alguna de las muchas técnicas de la estadística inferencial para tomar decisiones con respecto a la población, apoyándose en el conocimiento de causa evidenciado a partir de los datos y asignándole un nivel de confiabilidad o de incertidumbre a las conclusiones obtenidas.

4.2. Incertidumbre y distribuciones estadísticas

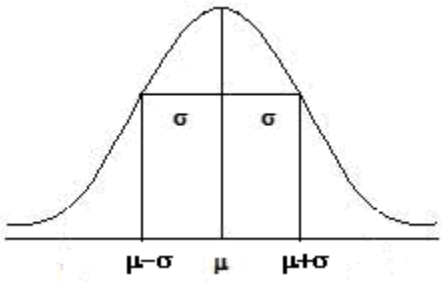
La estadística es la disciplina que estudia los procesos estocásticos, es decir aquellos que presentan variaciones, sin causa asignable (debidas al azar). Por lo que se han desarrollado técnicas que permiten detectar y diferenciar variaciones por efecto de algún factor, de las debidas al azar, con el fin de identificar su comportamiento y reducir estas últimas a un nivel aceptable para que no altere las características de calidad de los productos en manufacturación.

Con el apoyo de la teoría de la probabilidad se ha demostrado que las variables aleatorias tienen un comportamiento bien definido, que se puede representar mediante funciones de probabilidad y funciones de densidad de probabilidad, que dependiendo del tipo de unidades de medición generan las distribuciones estadísticas, base fundamental de las técnicas inferenciales. Debido a su importancia algunas de ellas se han tabulado para facilitar su uso; entre las más conocidas, sin ser las únicas, se encuentran:

- Binomial
- Poisson
- Normal (Z)
    - t-student
    - F-Fisher
    - Ji-cuadrada (χ^2)

Estas distribuciones realmente corresponden a modelos matemáticos, por ejemplo la función de densidad de la distribución normal tiene como expresión matemática la siguiente ecuación.

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{(y-\mu)}{\sigma}\right]^2}$$



Distribución Normal, mostrando los puntos de inflexión.

Donde se puede ver que la distribución queda totalmente representada por dos parámetros: μ (la media) y σ (la desviación estándar). Con las siguientes propiedades.

- Toda el área bajo la curva suma a 1.
- Los puntos de inflexión se localizan en $\mu-\sigma$ y $\mu+\sigma$.
- Entre $\mu-4\sigma$ y $\mu+4\sigma$ se encuentra la mayor parte del área bajo la curva (99.994%).

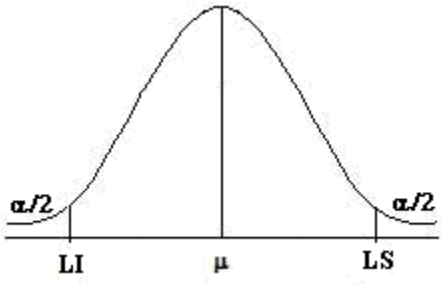
4.3. Distribución normal estándar

- A una distribución normal con μ = 0 y σ= 1 se le conoce como normal estándar y se representa por la variable z donde $z=\frac{(y-\mu)}{\sigma}$.

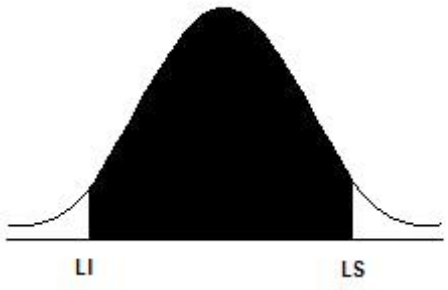
Su función densidad de probabilidad está dada por

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}[z]^2}$$

Cada conjunto de datos genera una distribución con sus propios valores de μ, σ y f(y), además es difícil que el valor estimado a partir de la media sea exactamente μ, por lo que es común establecer intervalos de confianza en los que se espera que el verdadero valor se encuentre entre un límite inferior (LI) y uno superior (LS). Valores que al representarse en la distribución, como área bajo la curva, indican una probabilidad.



Distribución normal y Límites de confianza para la media



Área bajo la curva delimitada por los límites de confianza.

Los valores de Z asociados a LI y LS acotan o delimitan cierta proporción del área, de ahí la importancia de saber, por ejemplo, que -1.96 < Z < 1.96 delimita el 95% del área bajo la curva de una distribución normal y que el área que no está sombreada corresponde al complemento a 1, en este caso al 5%, que expresado en probabilidades se le conoce como nivel de significancia, α, y a (1 - α) como nivel de confianza.

De la misma forma el valor de -2.5756 < Z < 2.5756 delimita el 99%, con un complemento de 1% que dividido entre 2 corresponde al 0.5% ($\alpha_{0.01)/2}$ = 0.005), lo interesante es que al asociar estos valores a los datos muestrales se pueden establecer intervalos de confianza para estimar los valores poblacionales.

4.4. Teorema Central del Límite

Este teorema establece que la distribución de las medias muestrales es normal aún cuando las muestras se toman de una distribución no-normal.

Si $y_1, y_2, \ldots, y_n$ son resultados de una muestra de n observaciones independientes de una variable aleatoria Y con media μ y desviación σ, la media de las $\bar{Y}$´s se distribuirá aproximadamente en forma normal con media y varianza, respectivamente:

$$\mu_{\bar{y}} = \mu \qquad \text{y} \qquad \sigma^2_{\bar{y}} = \frac{\sigma^2}{n}$$

La aproximación es mucho mejor cuando n se hace grande. En general, la población de la cual se toman las muestras no necesita ser normal, para que la distribución de las medias muestrales sea normal. Esto constituye lo más notorio y poderoso de este teorema.

4.5. Estimación (intervalos de confianza)

La estimación hace referencia al cálculo de intervalos de confianza para los parámetros de una distribución, a partir de datos muestrales.

Por ejemplo, para la estimación de la media se tiene:

$$P(LI < \mu < LS) = 1 - \alpha$$

que puede leerse como: la probabilidad de que el verdadero valor de μ esté en el intervalo acotado por LI y LS es 1-α, cuyo resultado numérico es LI $\leq \mu \leq$ LS.

4.6. Ejemplos

4.6.1. Contrastes de un Parámetro vs un valor predeterminado

Se realizaron seis determinaciones del contenido de hidrógeno de un compuesto cuya composición teórica es del 9.55% en promedio, ¿Difiere el valor promedio del teórico?

%H 9.17, 9.09, 9.14, 9.10, 9.13, 9.27

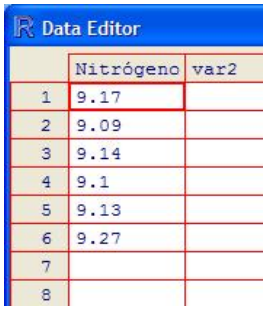
Solución

El primer paso consiste en establecer el par de hipótesis, en otras palabras: ¿quién es **Ho** y quién es **Ha?**

No realizar cálculo alguno si no sabe contra que hipótesis se está trabajando

En este caso se tiene

Ho: $\mu = 9.55$ Ha: $\mu \neq 9.55$

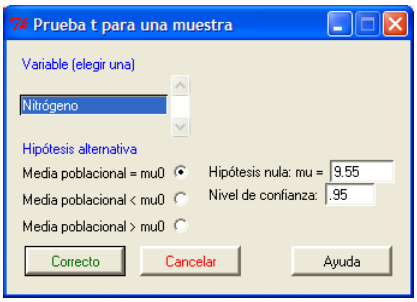


1. Ingresar los datos en R y almacenarlos en disco.

2. Seguir la secuencia

Estadísticos -> Medias -> Prueba t para una media

3. En el diálogo que aparece definir los valores adecuados a la hipótesis estadística de trabajo.



En este diálogo se puede definir el tipo de hipótesis, unilateral o bilateral y valor de referencia. En este caso, el valor es 9.55.

Resultados

One Sample t-test

data: Datos$Nitrógeno

t = -14.9766, df = 5, **p-value = 2.403e-05**

alternative hypothesis: true mean is not equal to 9.55

95 percent confidence interval:

9.081344 9.218656

sample estimates:

mean of x

9.15

Pregunta

¿Se tiene o no evidencia de que el valor promedio es de 9.55?

NOTA: Considerar principalmente el intervalo de confianza y el valor de p-value. **La regla práctica es rechazar Ho si p-value menor a 0.05** (y de manera complementaria, NO rechazar Ho si p-value es mayor de 0.05). En este caso se rechaza Ho.

Si la hipótesis es bilateral, y sólo en ese caso, se puede apoyar la conclusión en el intervalo de confianza, para este ejemplo: se tiene evidencia estadística (al 95% de confianza) de que el valor es menor a 9.55.

También se puede hacer el análisis con relación a la cantidad de error contenido en los datos, 2.0403e-05 = 0.000020403, que es el tamaño del error que podría cometer si rechazo Ho, en este caso 0.002403%, mucho menor que el 5%, el tamaño de error aceptado por convención.

Ejemplo 4.6.2. Se analizó el contenido de silicio de una muestra de agua por dos métodos independientes, en un intento por mejorar la precisión de la determinación. De acuerdo a los siguientes datos.

Método original	Método modificado
149 ppm	150 ppm
139	147
135	152
140	151
155	145

¿Qué se puede decir de las respuestas promedio y de su precisión?

Solución

El primer paso consiste en establecer el par de hipótesis, en otras palabras: ¿quién es Ho y quién es Ha?

Para las respuestas promedio se tiene:

Ho: $\mu_1 = \mu_2$ o $\mu_1 - \mu_2 = 0$

Ha: $\mu_1 \neq \mu_2$ o $\mu_1 - \mu_2 \neq 0$

Para la precisión se tienen las hipótesis

Ho: $\sigma_1^2 = \sigma_2^2$ o $\frac{\sigma_1^2}{\sigma_2^2} = 1$ y Ha: $\sigma_1^2 \neq \sigma_2^2$ o $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$

Este par de hipótesis permite probar que las precisiones son diferentes.

Ho: $\sigma_1^2 \leq \sigma_2^2$ o $\frac{\sigma_1^2}{\sigma_2^2} \leq 1$ y Ha: $\sigma_1^2 > \sigma_2^2$ o $\frac{\sigma_1^2}{\sigma_2^2} > 1$

Este último par de hipótesis permite probar que el método modificado (2) tiene menos variabilidad (mayor precisión) que el método original (1).

Los pasos a seguir en R, son:

Data Editor

	método	ppm
1	original	149
2	original	139
3	original	135
4	original	140
5	original	155
6	modificado	150
7	modificado	147
8	modificado	152
9	modificado	151
10	modificado	145
11		

1. Generar un nuevo conjunto de datos en R
2. Se tiene una columna, con una variable llamada puntajes
3. Seguir la secuencia:

   Estadísticas -> Resúmenes -> Resúmenes numéricos

4. Seguir la secuencia:

Estadísticas -> Medias -> Prueba t muestras independientes

5. Seguir la secuencia:

Estadísticas -> Varianzas -> Prueba F para dos varianzas

Resultados

```
> numSummary(Datos[,"ppm"], groups=Datos$método, statistics=c("mean", "sd",
"quantiles"))
              mean    sd  0%  25%  50%  75% 100%  n
modificado   149.0 2.915 145  147  150  151  152  5
original     143.6 8.173 135  139  140  149  155  5
```

```
> t.test(ppm~método, alternative='two.sided', conf.level=.95, var.equal=FALSE,
data=Datos)
Welch Two Sample t-test
data:  ppm by método
t = 1.3915, df = 5.002, p-value = 0.2228
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.574666 15.374666
```

El p-value es menor a 0.05, por lo tanto se rechaza Ho y por el intervalo de confianza, que va de un valor negativo a un valor positivo (lo que quiere decir que incluye al cero), se tiene evidencia estadística de que las medias son iguales.

```
sample estimates:
mean in group modificado   mean in group original
                   149.0                      143.6
> tapply(Datos$ppm, Datos$método,  var, na.rm=TRUE)
modificado   original
       8.5       66.8
```

```
> var.test(ppm ~ método, alternative='two.sided', conf.level=.95, data=Datos)
F test to compare two variances
data:  ppm by método
F = 0.1272, num df = 4, denom df = 4, p-value = 0.0707
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.01324849 1.22213329
sample estimates:
ratio of variances
0.1272455
```

El p-value está apenas por arriba de 0.05 y el intervalo de confianza incluye el valor de 1 (va de 0.01324 a 1.2221), entonces se tiene evidencia de que las varianzas son iguales. Por los valores valdría la pena probar si la varianza 2 es mayor que la varianza 1.

Ejemplo 4.6.3. Se realizó un estudio para probar que un programa de ejercicios regulares moderadamente activos beneficia a pacientes que previamente han sufrido un infarto al miocardio. Once individuos participan en el estudio, de manera que antes de iniciar el programa se les determina la capacidad de trabajo midiendo el tiempo que tardan en alcanzar una tasa de 160 latidos por minuto mientras caminaban sobre una banda sin fin. Después de 25 semanas de ejercicio controlado, se repitieron las medidas, encontrando los siguientes resultados.

Sujeto	1	2	3	4	5	6	7	8	9	10	11
Antes	7.6	9.9	8.6	9.5	8.4	9.2	6.4	9.9	8.7	10.3	8.3
Después	14.7	14.1	11.8	16.1	14.7	14.1	13.2	14.9	12.2	13.4	14.0

¿Realmente funciona el programa de ejercicios?

Solución

1. Establecer el par de hipótesis, en otras palabras: ¿quién es Ho y quién es Ha?

Para las respuestas promedio se tiene:

Ho: $\mu_d \geq 0$ **o** $\mu_A - \mu_B \geq 0$

Ha: $\mu_d < 0$ **o** $\mu_A - \mu_B < 0$

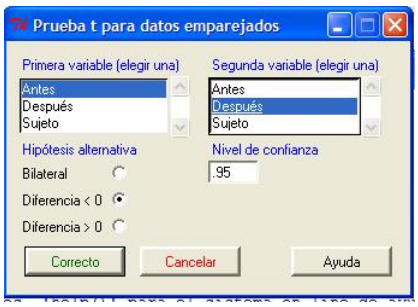
2. Ingresar los datos en R

Data Editor

	Sujeto	Antes	Después
1	1	7.6	14.7
2	2	9.9	14.1
3	3	8.6	11.8
4	4	9.5	16.1
5	5	8.4	14.7
6	6	9.2	14.1
7	7	6.4	13.2
8	8	9.9	14.9
9	9	8.7	12.2
10	10	10.3	13.4
11	11	8.3	14
12			

3. Seguir la secuencia

Estadísticos -> Medias -> Prueba t para datos emparejados



4. En el diálogo que aparece seleccionar la hipótesis unilateral, de acuerdo a la hipótesis alternativa.

5. Dar un clic al botón correcto y obtener los siguientes resultados.

```
> t.test(Datos$Antes, Datos$Después, alternative='less', conf.level=.95, paired=TRUE)
Paired t-test
data:  Datos$Antes and Datos$Después
t = -11.4749, df = 10, p-value = 2.222e-07
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf -4.317419
sample estimates:
```

```
mean of the differences
            -5.127273
```

6. Concluir, si funciona o no el programa de ejercicios.

P-value es muchísimo menor de 0.05 (realmente es 2.222×10^{-7} = 0.0000002222), entonces se tiene evidencia estadística que el programa de ejercicio SI beneficia a los pacientes.

4.7. Ejercicios

1. Las siguientes son medidas de profundidad (m), en una estación de investigación oceanográfica: 46.8, 43.8, 44.6, 38.9, 45.6, 52.1, 40.1, 53.4, 49.4, 53.2, 46.3, 47.8, 42.2 y 44.9.

   ¿Contradicen estos datos la aseveración de que la profundidad promedio en esta zona es de 42.5 m?

2. Las siguientes son calificaciones de 15 estudiantes, en un mismo examen aplicado a la mitad y final de un curso de estadística.

Mitad	66	88	75	90	63	58	75	82	73	84	85	93	70	82	90
Fin	73	91	78	86	69	67	75	80	76	89	81	96	76	90	97

Los alumnos afirman que mejoraron sus calificaciones, ¿es cierto o no?

3. Para determinar la efectividad de un nuevo sistema de control de tránsito, se observó el número de accidentes en diez cruceros peligrosos cuatro semanas antes (medición 1) y cuatro semanas después (medición 2) de la instauración del nuevo sistema, obteniendo los siguientes resultados.
4.

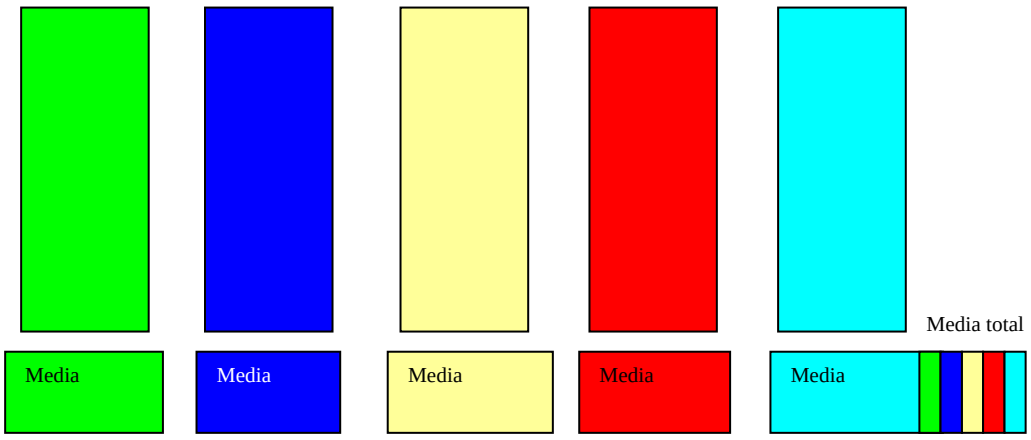
M1	3	4	2	5	3	2	3	6	1	1
M2	1	2	3	2	3	0	2	3	2	0

¿Qué se puede decir del nuevo sistema?

Capítulo 5

Análisis de varianza y diseño de experimentos

Suponga un experimento donde se quieren comparar 5 tratamientos, para ver si su respuesta promedio es la misma para los 5 o si hay algunas diferentes.



De antemano el investigador asume que hay diferencia, si no que sentido tiene el experimento. También se sabe que en cada tratamiento debe haber un efecto de

variaciones debida a la causa que se está controlando (temperatura, presión, etcétera) y una variación debida al azar, la cual es inevitable.

La variación entre tratamientos se mide como una varianza de la media de cada tratamiento con respecto a la gran media.

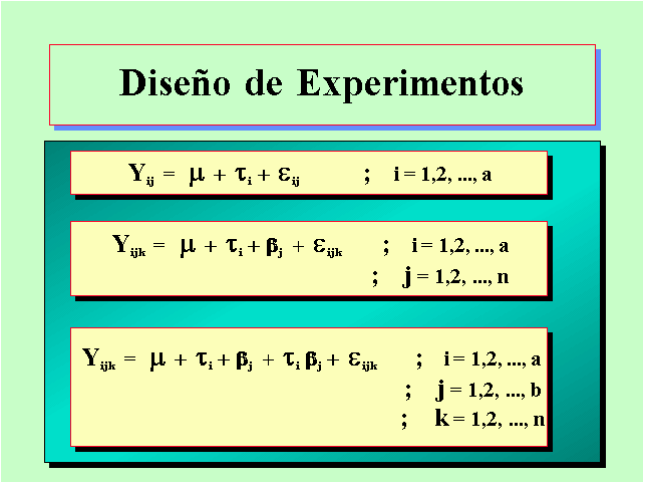
La variación dentro de tratamientos se mide comparando cada observación o medición con respecto a la media del respectivo tratamiento y en términos del análisis de varianza se le conoce como cuadrado medio del error.

Ahora, si se tienen dos varianzas (entre tratamientos y dentro de tratamientos) lo que se puede hacer es compararlas mediante una prueba de F.

$$F = \frac{\text{Varianza entre tratamientos}}{\text{Varianza dentro tratamientos}}$$

La variación dentro de tratamientos se debe al azar y si no se puede establecer diferencia estadística entre estas varianzas, entonces no hay efecto de tratamiento y la variación se debe al azar.

5.1. Modelos más comunes en el diseño de experimentos



Ec. (1)

Ec. (2)

Ec. (3)

5.1.1. Diseño Completamente al Azar (DCA), de un factor o One-Way

La característica esencial es que todas las posibles fuentes de variación o de influencia están controladas y sólo hay efecto del factor en estudio. Este es el experimento ideal, todo controlado y lo único que influye es el factor de estudio.

5.1.2. Diseño de Bloques al Azar Completo (DBAC)

Sigue siendo un diseño de una vía pero hay alguna fuente con un gradiente de variación, que influye o afecta en el experimento, por lo tanto hay que cuantificar su efecto y eliminarlo de la varianza dentro de tratamientos, para evitar que nos conduzca a valores bajos de F y se llegue a conclusiones erróneas.

5.1.3. Diseños Factoriales

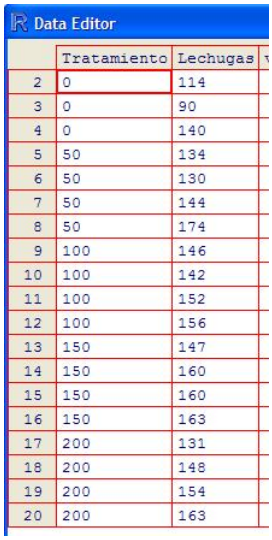
La tercera ecuación, de la figura anterior, muestra un diseño con dos factores de estudio, donde el mayor interés está en el efecto de la interacción, $\tau_i\beta_\varphi$. Nótese la semejanza entre el modelo de la ecuación 2 y la 3, en la figura anterior.

5.2. Ejemplo 1

Se realizó un experimento para probar los efectos de un fertilizante nitrogenado en la producción de lechuga. Se aplicaron cinco dosis diferentes de nitrato de amonio a cuatro parcelas (réplicas). Los datos son el número de lechugas cosechadas de la parcela.

Tratamiento (Kg N/Ha)				
0	104	114	90	140
50	134	130	144	174
100	146	142	152	156
150	147	160	160	163
200	131	148	154	163

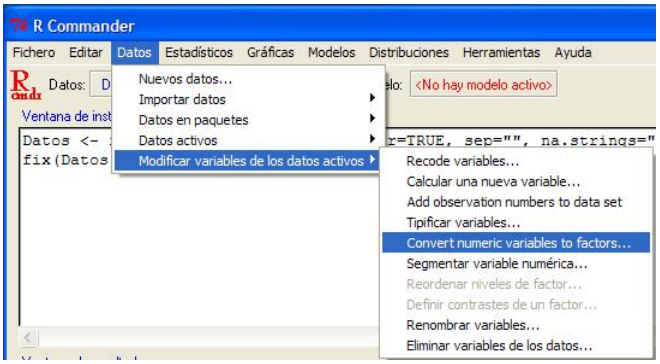
Solución



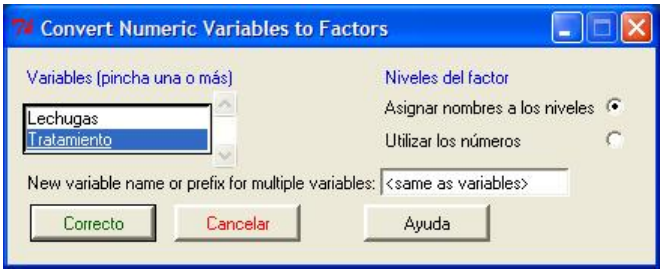
1. Ingresar los datos a R, en dos columnas. Una para el tratamiento y otra para el número de lechugas.

2. Seguir la secuencia:

Datos -> Modificar variables de los datos activos -> Convert numeric variables to factor

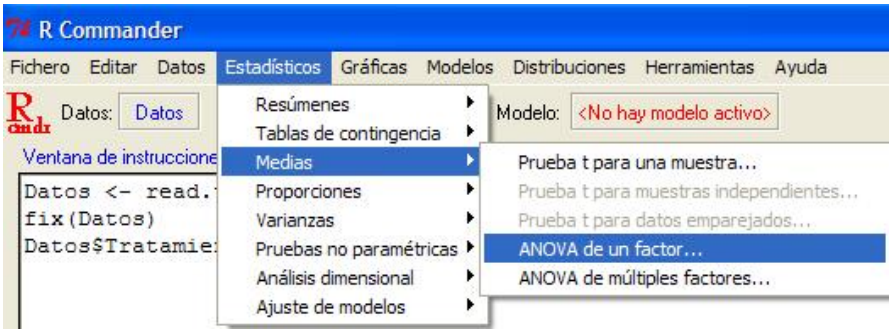


3. Seleccionar la variable Tratamiento como factor



5. Seguir la secuencia:

Estadísticos -> Medias -> ANOVA de un factor



NOTA: Al llamar al procedimiento ANOVA, R presenta el siguiente mensaje en la consola.

Versión del Rcmdr 1.2-2

Loading required package: multcomp

Error: package 'mvtnorm' required by 'multcomp' could not be found

Esto quiere decir que se debe conseguir el "paquete" mvtnorm" e instalarlo, antes de lograr que se despliegue el diálogo de ANOVA y empezar a tener resultados.

Resultados

```
> anova(lm(Lechuga ~ Tratamiento, data=Datos))
Analysis of Variance Table
Response: Lechuga
             Df  Sum Sq Mean Sq F value    Pr(>F)
Tratamiento   4  4994.8  1248.7  5.6113 0.005757 **
Residuals    15  3338.0   222.5
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Aquí Pr(>F) equivale al p-value y esto indica que se rechaza Ho, entonces al menos un par de medias es diferente, ¿PERO CUALES?

```
> tapply(Datos$Lechuga, Datos$Tratamiento, mean, na.rm=TRUE) # means
   0      50     100     150     200
112.0  145.5  149.0  157.5  149.0
```

Estas son las medias de cada tratamiento, aquí se aprecia cuál es la menor y la mayor, pero hay aportar evidencia estadística de esta diferencia.

```
> tapply(Datos$Lechuga, Datos$Tratamiento, sd, na.rm=TRUE) # std. Deviations
       0            50           100           150           200
21.102922  19.891372   6.218253   7.141428  13.490738
```

Estas son las desviaciones estándar de cada tratamiento

```
> tapply(Datos$Lechuga, Datos$Tratamiento, function(x) sum(!is.na(x))) # counts
 0  50 100 150 200
 4   4   4   4   4
```

Conteo de cuantos datos por tratamiento

```
> .Pairs <- simint(Lechuga ~ Tratamiento, type="Tukey", data=Datos)
> summary(.Pairs)
        Simultaneous 95% confidence intervals: Tukey contrasts
Call:
simint.formula(formula = Lechuga ~ Tratamiento, data = Datos,    type = "Tukey")
Tukey contrasts for factor Tratamiento
```

Contrast matrix: **Esta matriz define la estructura de los contrastes a realizar, para fines prácticos se puede obviar.**

	Tratamiento0	Tratamiento50	Tratamiento100
Tratamiento50-Tratamiento0	0	-1	1	0
Tratamiento100-Tratamiento0	0	-1	0	1
Tratamiento150-Tratamiento0	0	-1	0	0
Tratamiento200-Tratamiento0	0	-1	0	0
Tratamiento100-Tratamiento50	0	0	-1	1
Tratamiento150-Tratamiento50	0	0	-1	0
Tratamiento200-Tratamiento50	0	0	-1	0
Tratamiento150-Tratamiento100	0	0	0	-1
Tratamiento200-Tratamiento100	0	0	0	-1
Tratamiento200-Tratamiento150	0	0	0	0

	Tratamiento150	Tratamiento200
Tratamiento50-Tratamiento0	0	0
Tratamiento100-Tratamiento0	0	0
Tratamiento150-Tratamiento0	1	0
Tratamiento200-Tratamiento0	0	1
Tratamiento100-Tratamiento50	0	0
Tratamiento150-Tratamiento50	1	0
Tratamiento200-Tratamiento50	0	1
Tratamiento150-Tratamiento100	1	0
Tratamiento200-Tratamiento100	0	1
Tratamiento200-Tratamiento150	-1	1

Absolute Error Tolerance: 0.001

95 % quantile: 3.088

Para la prueba de tukey, la ho es que el par de medias es igual y la alternativa que el par de medias es diferente.

Todos los pares con p raw menor a 0.05 son diferentes

Coefficients:

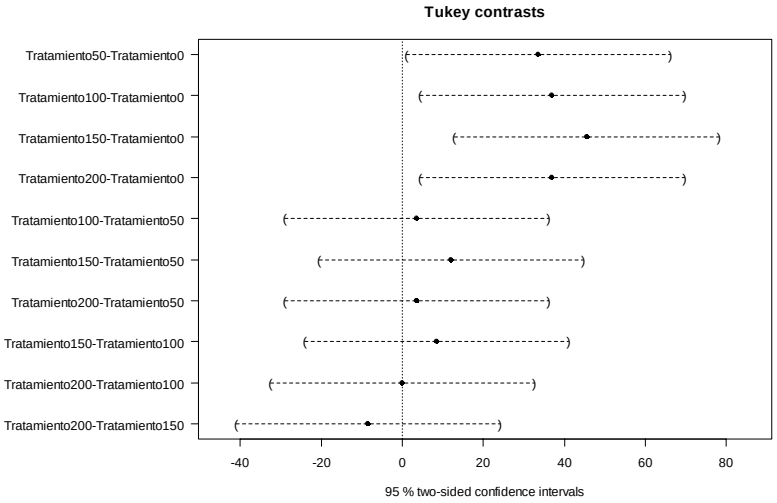
	Estimate	2.5 %	97.5 %	t value	Std.Err.	p raw
Tratamiento50-Tratamiento0	33.5	0.923	66.077	3.176	10.548	0.006
Tratamiento100-Tratamiento0	37.0	4.423	69.577	3.508	10.548	0.003
Tratamiento150-Tratamiento0	45.5	12.923	78.077	4.313	10.548	0.001
Tratamiento200-Tratamiento0	37.0	4.423	69.577	3.508	10.548	0.003
Tratamiento100-Tratamiento50	3.5	-29.077	36.077	0.332	10.548	0.745
Tratamiento150-Tratamiento50	12.0	-20.577	44.577	1.138	10.548	0.273
Tratamiento200-Tratamiento50	3.5	-29.077	36.077	0.332	10.548	0.745
Tratamiento150-Tratamiento100	8.5	-24.077	41.077	0.806	10.548	0.433
Tratamiento200-Tratamiento100	0.0	-32.577	32.577	0.000	10.548	1.000
Tratamiento200-Tratamiento150	-8.5	-41.077	24.077	-0.806	10.548	0.433

La prueba de Bonferroni es útil cuando el número de repeticiones es diferente en cada tratamiento (desbalanceado). Cuando el diseño está balanceado p raw y p Bonf conducen a la misma conclusión.

	p Bonf	p adj
Tratamiento50-Tratamiento0	0.063	0.042
Tratamiento100-Tratamiento0	0.032	0.022
Tratamiento150-Tratamiento0	0.006	0.005
Tratamiento200-Tratamiento0	0.032	0.023
Tratamiento100-Tratamiento50	1.000	0.997
Tratamiento150-Tratamiento50	1.000	0.785
Tratamiento200-Tratamiento50	1.000	0.997
Tratamiento150-Tratamiento100	1.000	0.925
Tratamiento200-Tratamiento100	1.000	1.000
Tratamiento200-Tratamiento150	1.000	0.925

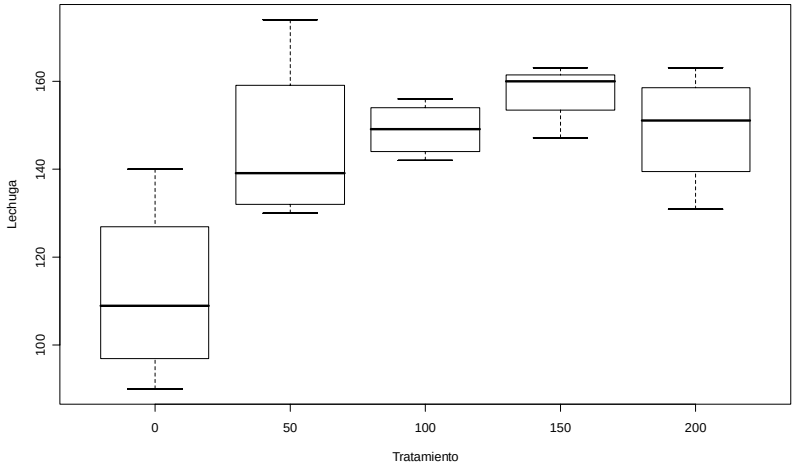
> plot(.Pairs)

> remove(.Pairs)



Este gráfico es un resumen de la prueba de Tukey, de manera que aquellas diferencias de medias que intercepten al cero son iguales y las que no lo intercepten son diferentes entre ellas. En este caso, a excepción del tratamiento 0, todos los demás tratamientos son iguales entre ellos.

Aquí valdría la pena, realizar un gráfico de cajas, por cada tratamiento, que puede apoyar la conclusión final. El cual, a estas alturas del partido, es un hecho que Ud. apreciable lector, ya sabe cómo se hace.



El criterio práctico es: los tratamientos cuyas cajas se interceptan son iguales y los que no, pues entonces son diferentes. Lo mejor es concluir con base en los resultados de la prueba de Tukey.

5.3 Ejercicio del capítulo

En cada uno de los siguientes ejercicios: seleccionar el tipo de diseño experimental, plantear el correspondiente juego de hipótesis y realizar el análisis de comparación múltiple de medias.

5.3.1. Un fabricante supone que existe diferencia en el contenido de calcio en lotes de materia prima que le son suministrados por su proveedor. Actualmente hay una gran cantidad de lotes en la bodega. Cinco de estos son elegidos aleatoriamente. Un químico realiza cinco pruebas sobre cada lote y obtiene los siguientes resultados.

Lote				
1	**2**	**3**	**4**	**5**
23.46	23.59	23.51	23.28	23.29
23.48	23.46	23.64	23.40	23.46
23.56	23.42	23.46	23.37	23.37
23.39	23.49	23.52	23.46	23.32
23.40	23.50	23.49	23.39	23.38

5.3.2.. Para probar 4 dietas diferentes sobre el incremento en peso de cerdos se tienen los siguientes datos.

Peso de los cerdos (Kgs)			
Dieta 1	**Dieta 2**	**Dieta 3**	**Dieta 4**
60.8	68.7	102.6	87.9
57.0	67.7	102.1	84.2
65.0	74.7	100.2	83.1
58.6	66.3	96.5	85.7
61.7	69.8		90.3

5.3.3. Un ingeniero industrial está realizando un experimento sobre el tiempo de enfoque del ojo. Se interesa en el efecto de la distancia del objeto al ojo sobre el tiempo de enfoque. Cuatro distancias diferentes son de interés. Se cuenta con cinco sujetos no homogéneos para el experimento.

	Sujeto				
Distancia (pies)	**1**	**2**	**3**	**4**	**5**
4	10	6	6	6	6
6	7	6	6	1	6
8	5	3	3	2	5
10	6	4	4	2	3

5.3.4. Aumento en el peso de camarón cultivado en acuarios con diferentes niveles de temperatura (***T***), Densidad de Población (***D***) y Salinidad de Agua (***S***), luego de cuatro semanas.

T (°C)	D (organismos/40 litros)	S (%)	Aumento de Peso (mg)
25	80	10	86, 52, 73
		25	544, 371, 482
		40	390, 290, 397
	160	10	53, 73, 86
		25	393, 398, 208
		40	249, 265, 243
35	80	10	439, 436, 349
		25	249, 245, 330
		40	247, 277, 205
	160	10	324, 305, 364
		25	352, 267, 316
		40	188, 223, 281

Tomado de: Robert O. Kuehl, 2001, Diseño de experimentos. Principios estadísticos de diseño y análisis de investigación, 2ª. Ed., Thomson Leraning editores, México, pág. 201.

Capítulo 6

Análisis de Regresión lineal

6.1. Problemas que se plantean:

1) ¿Cuál es el modelo matemático más apropiado para describir la relación entre una o más variables independientes ($X_1, X_2, \cdots, X_k$) y una variable dependiente (Y)?

2) Dado un modelo especifico, ¿qué significa éste y cómo se encuentran los parámetros del modelo que mejor ajustan a nuestros datos? Si el modelo es una línea recta: ¿cómo se encuentra la "mejor recta"?

La ecuación de una línea recta es:

$$Y = f(X) = \beta_0 + \beta_1 X$$

Donde:

β_0 ordenada al origen

β_1 pendiente

En un análisis de regresión lineal simple, el problema es encontrar los valores que mejor estimen a los parámetros β_0 y β_1. A partir de una muestra aleatoria.

El modelo de regresión lineal es:

$$Y_i = \mu_{Y/X} + \varepsilon_i = \beta_0 + \beta_1 X + \varepsilon_i \text{, con } i = 1,2,3,\cdots,n$$

Para cada observación el modelo es:

$$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2$$

$$\vdots$$

$$Y_n = \beta_0 + \beta_1 X_n + \varepsilon_n$$

Que se puede escribir como:

$${}_nY_1 = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad {}_nX_2 = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ & \vdots \\ 1 & X_n \end{pmatrix} \quad {}_2\beta_1 = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad {}_n\varepsilon_1 = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

donde:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ & \vdots \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = Y = X\beta + \varepsilon$$

6.2. Estimación por mínimos cuadrados

Sea $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ la respuesta estimada en X_i con base en la línea de regresión ajustada. La distancia vertical entre el punto (X_i, Y_i) y el punto $(X_i, \hat{Y}_i)$ de la recta

ajustada está dada por el valor absoluto de $|Y_i - \hat{Y}_i|$ o $|\hat{Y}_i - \hat{\beta}_0 - \hat{\beta}_1 X_i|$, cuya suma de cuadrados es:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

El problema ahora es encontrar los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ tales que $\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$ sea mínimo.

Solución:

Si Q = $\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$, entonces

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \qquad (1)$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-X_i) = 0 \qquad (2)$$

(NOTA: las derivadas parciales se igualan a cero para determinar los puntos críticos, que serán mínimos). Esto conduce a las **Ecuaciones Normales de Mínimos Cuadrados**

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2$$

En notación matricial se tiene:

$$\begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix}$$

$$\mathbf{X}^t\mathbf{X}\beta = \mathbf{X}^t\mathbf{Y}$$

de donde

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{Y})$$

Solución matricial para calcular los parámetros de la ecuación de regresión

La solución algebraica de las ecuaciones normales, para datos muestrales, genera las siguientes ecuaciones:

$$b_0 = \frac{\sum_{i=}^{n} Y_i \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

$$b_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

6.3. Algo de geometría

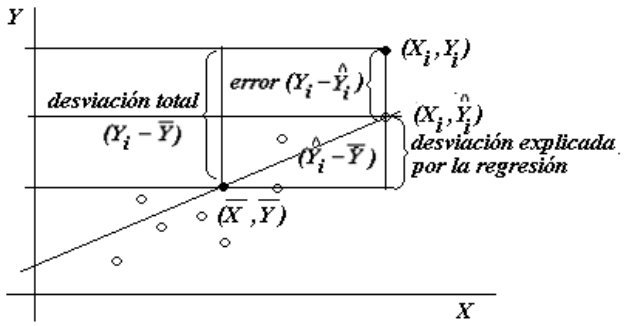


Figura 6.1. Desviaciones: total, explicada por la regresión y error.

1) $Y_i - \overline{Y}$ desviación total

2) $\hat{Y}_i - \overline{Y}$ desviación explicada por la regresión

3) $Y_i - \hat{Y}_i$ error

$$Y_i - \overline{Y} = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$$

Total = Regresion + Error

Al aplicar sumatorias y elevar al cuadrado se tiene:

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}[(\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)]^2$$

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$

$SC_{Total} = SC_{Regresion} + SC_{Error}$

Cantidades que permiten realizar un ANOVA, para contrastar las hipótesis:

$$Ho: \ \beta_i = 0$$

$$Ha: \ \beta_i \neq 0$$

Fuente de variación	**g.l.**	**SC**	**CM**	**F_C**	**F_t**
Regresión	1	SC_{Reg}	CM_{Reg}	$\frac{CM_{Reg}}{CM_{Error}}$	$F_{1-\alpha,1,n-2}$
Error Residual	$n-2$	SC_{Error}	CM_{Error}		
Total	$n-1$	SC_{Total}			

Este ANOVA considera el siguiente par de hipótesis:

Ho: $\beta_i = 0$, es decir que todos los coeficientes del modelo son iguales a cero y por lo tanto no hay un modelo lineal que describa el comportamiento de los datos.

Contra *Ha*: $\beta_i \neq 0$ de que al menos uno de los coeficientes es diferente de cero y entonces si hay un modelo lineal.

6.4. Interpretando a β_0 y β_1

Ho: $\beta_1 = 0$

Caso 1.- *Ho*: $\beta_1 = 0$ **No se rechaza. Es decir que la pendiente es cero o que no hay pendiente,** entonces se tienen dos opciones de interpretación.

a) Si la suposición de línea recta es correcta significa que *X* no proporciona ayuda para predecir *Y*, esto quiere decir que $\overline{Y}$ predice a *Y*.

Ho: $\beta_1 = 0$

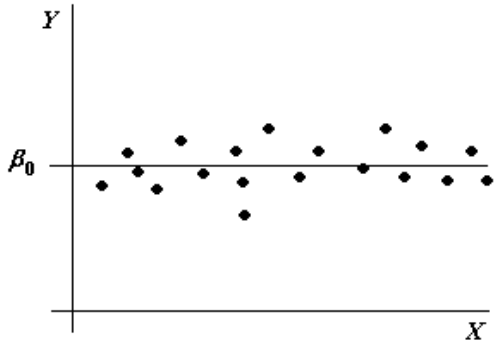


Figura 6.2. *X* no proporciona ayuda para predecir *Y.*

b) La verdadera relación entre *X* y *Y* no es lineal, esto significa que el modelo puede involucrar funciones cuadráticas, cúbicas o funciones más complejas.

Figura 6.3. La relación entre *X* y *Y* no es lineal.

NOTA: Si hay una curvatura se requiere un elemento cuadrático en el modelo, si hay dos curvaturas entonces se requiere un cúbico y así sucesivamente.

Caso 2.- *Ho*: $\beta_1 = 0$ **se rechaza (es decir, si hay pendiente o en otras palabras si hay un modelo lineal que describe el comportamiento de los datos).**

a) *X* proporciona información significativa para predecir *Y*

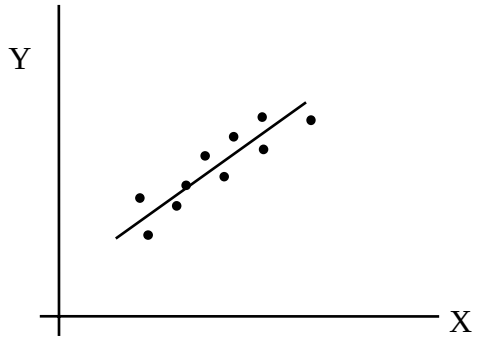


Figura 6.4. La relación entre *X* y *Y* es lineal.

b). El modelo puede tener un término lineal más, quizás un término cuadrático.

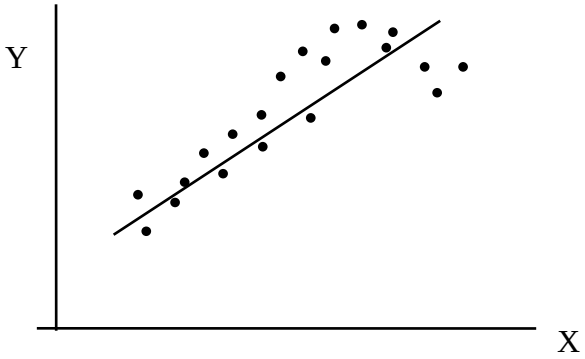


Figura 6.5. La relación entre *X* y *Y* no es lineal.

Caso 3. Prueba. *Ho*: $\beta_0 = 0$, Si NO se rechaza esta Hipótesis, puede ser apropiado ajustar un modelo sin β_0, siempre y cuando exista experiencia previa o teoría que sugiera que la recta ajustada debe pasar por el origen y que existan datos alrededor del origen para mejorar la información sobre β_0.

6.5. Correlacion

Si *X* y *Y* son dos variables aleatorias (no existe causa-efecto), entonces el coeficiente de correlación se define como:

1) r $\in$ [-1,1]

2) r es independiente de las unidades de *X* y *Y*

3) $\hat{\beta}_1 > 0 \Leftrightarrow$ r > 0

$\hat{\beta}_1 < 0 \Leftrightarrow$ r < 0

$\hat{\beta}_1 = 0 \Leftrightarrow$ r = 0

r es una medida de la fuerza de asociación lineal entre *X* y *Y*

NOTA: NO se puede ni se deben establecer relaciones causales a partir de los valores de r, ya que ambas variables son aleatorias.

6.6. Coeficiente de determinación r^2

$$r^2 = \frac{SC_{total} - SC_{error}}{SC_{total}} = \frac{SC_{Reg}}{SC_{Total}}$$

donde, $r^2 \in [0,1]$

Esta r-cuadrada es una medida de la variación de *Y* explicada por los cambios o variación en la *X*. Es común leerla como porcentaje de variación en *Y* explicada por los cambios en *X*.

6.7. Diagnóstico del modelo de regresión lineal simple

Las técnicas de diagnóstico son esenciales para detectar desacuerdos entre el modelo y los datos para los cuales se ajusta éste. Esto se hace a través del análisis de los residuos.

Los supuestos que se hacen del estudio del análisis de regresión son:

- La relación entre Y y X es lineal.
- Los errores tienen media cero
- Los errores tienen varianza constante σ^2.
- Los errores no están correlacionados (son independientes).
- Los errores se distribuyen normalmente.

Las posibles violaciones al modelo se pueden detectar a través de los residuos y son:

- Evidencias que sugieren que la forma del modelo no es la apropiada.
- Presencia de casos extraordinarios (outliers) en los datos.
- Evidencia que sugieren varianza no constante.
- Evidencia de que la distribución de los errores no proviene de una distribución normal.
- Autocorrelación, que se define como la falta de independencia de los residuos (errores).

6.8. Ejemplo

Con los siguientes datos,

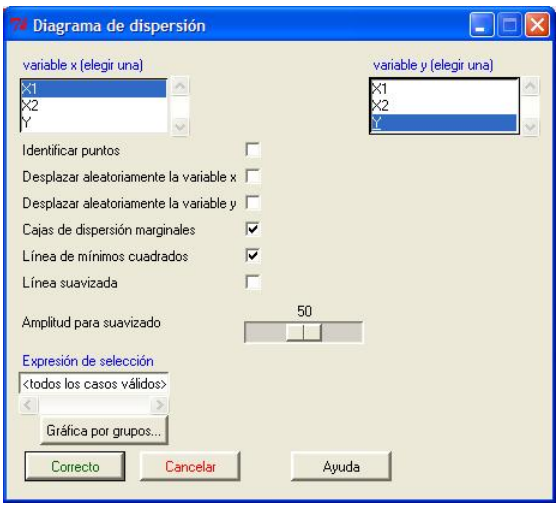
a) Identificar la ecuación de regresión.
b) ¿Hay evidencia suficiente para establecer una relación lineal positiva entre x_1 con y?
c) ¿Hay evidencia para establecer una relación lineal negativa entre x_2 con y?

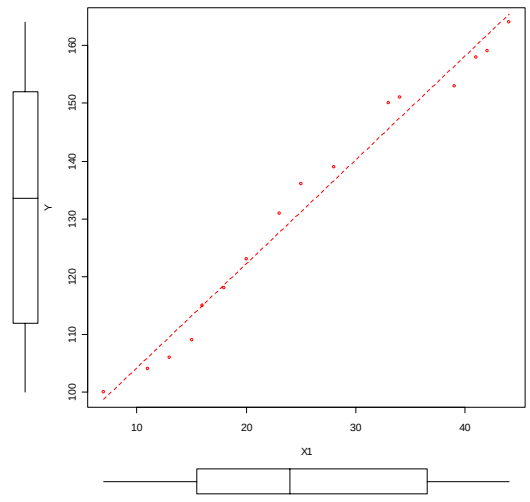
d) ¿Hay suficiente evidencia para establecer que el modelo de regresión es útil?

Y	X_1	X_2
100	7	28
104	11	27
106	13	29
109	15	31
115	16	26
118	18	24
123	20	20
131	23	18
136	25	22
139	28	20
150	33	19
151	34	17
153	39	14
158	41	12
159	42	14
164	44	13

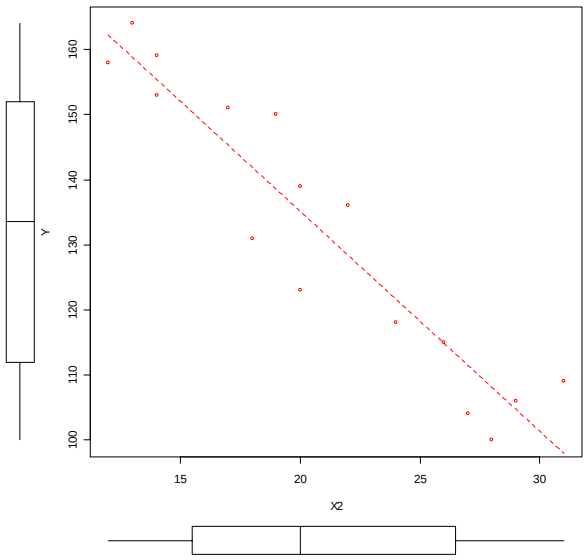
1. Seguir la secuencia:

Gráficos -> Diagrama de dispersión. Para X1 – Y.



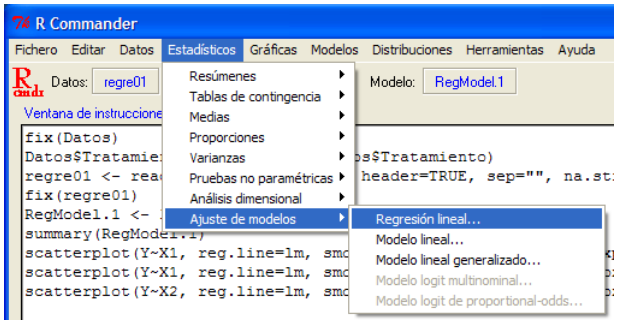


2. Repetir los pasos del punto 1, para la variable X2



3. Hacer el ajuste de manera numérica, con la secuencia:

Estadísticas -> Ajuste de modelos -> Regresión líneal



4. Seleccionar las variables en el modelo. Para Y–X1, Y–X2 y finalmente para Y y las dos X's, X1 y X2.

5. Analizar los resultados

```
> RegModel.3 <- lm(Y~X1, data=regre01)
> summary(RegModel.3)

Call:
lm(formula = Y ~ X1, data = regre01)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2100 -2.2784 -0.3152  2.6094  4.7640

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  86.17092    1.87143   46.05  < 2e-16 ***
X1            1.80260    0.06661   27.06  1.73e-13 ***
```

Este último valor indica que hay un buen ajuste lineal entre la variable Y y la variable X1.

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.106 on 14 degrees of freedom
Multiple R-Squared: 0.9812,     Adjusted R-squared: 0.9799
F-statistic: 732.3 on 1 and 14 DF,  p-value: 1.726e-13
```

Si hay un modelo estadístico.

```
> RegModel.4 <- lm(Y~X2, data=regre01)
> summary(RegModel.4)

Call:
lm(formula = Y ~ X2, data = regre01)
Residuals:
```

```
    Min      1Q  Median      3Q     Max
-12.2080  -5.0753  0.6461  5.2587  11.4115

Coefficients:
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  202.819      7.051    28.76  7.46e-14 ***
X2            -3.381      0.325   -10.40  5.73e-08 ***
```

La variable X2 modela o ajusta bien con la variable Y.

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.676 on 14 degrees of freedom
Multiple R-Squared: 0.8854,     Adjusted R-squared: 0.8772
F-statistic: 108.2 on 1 and 14 DF,  p-value: 5.732e-08
```

Si hay modelo que describe la relación entre las dos variables.

```
> RegModel.5 <- lm(Y~X1+X2, data=regre01)
> summary(RegModel.5)

Call:
lm(formula = Y ~ X1 + X2, data = regre01)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7098  -2.3772  -0.8203  1.7529  5.1548

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  101.1684    11.9726    8.450  1.22e-06 ***
X1             1.5875     0.1818    8.731  8.48e-07 ***
```

X1 si “entra” en el modelo multiple.

```
X2              -0.4550      0.3590    -1.268      0.227
```

X2 no “entra” en el modelo múltiple.

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.041 on 13 degrees of freedom
Multiple R-Squared: 0.9833,      Adjusted R-squared: 0.9807
F-statistic: 382.8 on 2 and 13 DF,  p-value: 2.798e-12
```

Si hay un modelo múltiple que describe la relación entre las variables.

6.9. Ejercicio del capítulo 6

Los siguientes datos son parte de un estudio grande conducido por el Dr. Rick Linthurst de la Universidad estatal de North Carolina. El propósito de la investigación fue identificar las principales características del suelo que influyen en la producción de biomasa aérea en el pasto de pantano *Spartina alterniflora* en el estuario del Cabo de Miedo en North Carolina.

Una fase de la investigación consistió en muestrear tres tipos de Spartina (áreas revegetativas "muertas" (DVEG), áreas de Spartina "enana" (SHRT) y áreas de Spartina "alta" (TALL)), en cada una de tres localidades (Oak Island (OI), Smith Island (SI) y Snows Marsh (SM)). Las muestras de sustrato de 5 lugares seleccionados al azar dentro de cada tipo de localidad-vegetación, dan un total de 45 muestras que fueron analizadas para 14 características del suelo para cada mes por varios meses. Además se midió la biomasa superficial para cada lugar de muestreo cada mes. Los datos usados en este estudio de caso involucraron solo los del mes de septiembre y las siguientes 5 medidas del sustrato.

X_1 = Salinidad ‰ (SAL)

X_2 = Acidez del agua (pH)

X_3 = Potasio en ppm (K)

X_4 = Sodio en ppm (Na)

X_5 = Zinc en ppm (Zn)

La variable dependiente Y es la biomasa aérea en $\frac{g}{m^2}$. Los datos para el mes de septiembre y las 6 variables se dan a continuación.

Obs.	Loc.	Tipo	BIO	SAL	pH	K	Na	Zn
1	OI	DVEG	676	33	5.00	1441.67	35184.5	16.4524
2	OI	DVEG	516	35	4.75	1299.19	28170.4	13.9852
3	OI	DVEG	1052	32	4.20	1154.27	26455.0	15.3276
4	OI	DVEG	868	30	4.40	1045.15	25072.9	17.3128
5	OI	DVEG	1008	33	5.55	521.62	31664.2	22.3312
6	OI	SHRT	436	33	5.05	1273.02	25491.7	12.2778
7	OI	SHRT	544	36	4.25	1346.35	20877.3	17.8225
8	OI	SHRT	680	30	4.45	1253.88	25621.3	14.3516
9	OI	SHRT	640	38	4.75	1242.65	27587.3	13.6826
10	OI	SHRT	492	30	4.60	1282.95	26511.7	11.7566
11	OI	TALL	984	30	4.10	553.69	7886.5	9.8820
12	OI	TALL	1400	37	3.45	494.74	14596.0	16.6752
13	OI	TALL	1276	33	3.45	526.97	9826.8	12.3730

14	OI	TALL	1736	36	4.10	571.14	11978.4	9.4058
15	OI	TALL	1004	30	3.50	408.64	10368.6	14.9302
16	SI	DVEG	396	30	3.25	646.65	17307.4	31.2865
17	SI	DVEG	352	27	3.35	514.03	12822.0	30.1652
18	SI	DVEG	328	29	3.20	350.73	8582.6	28.5901
19	SI	DVEG	392	34	3.35	496.29	12369.5	19.8795
20	SI	DVEG	236	36	3.30	580.92	14731.9	18.5056
21	SI	SHRT	392	30	3.25	535.82	15060.6	22.1344
22	SI	SHRT	268	28	3.25	490.34	11056.3	28.6101
23	SI	SHRT	252	31	3.20	552.39	8118.9	23.1908
24	SI	SHRT	236	31	3.20	661.32	13009.5	24.6917
25	SI	SHRT	340	35	3.35	372.15	15003.7	22.6758
26	SI	TALL	2436	29	7.10	525.65	10225.0	0.3729
27	SI	TALL	2216	35	7.35	563.13	8024.2	0.2703
28	SI	TALL	2096	35	7.45	497.96	10393.0	0.3705
29	SI	TALL	1660	30	7.45	458.38	8711.6	0.2648
30	SI	TALL	2272	30	7.40	498.25	10239.6	0.2105
31	SM	DVEG	824	26	4.85	936.26	20436.0	18.9875
32	SM	DVEG	1196	29	4.60	894.79	12519.9	20.9687
33	SM	DVEG	1960	25	5.20	941.36	18979.0	23.9841
34	SM	DVEG	2080	26	4.75	1038.79	22986.1	19.9727
35	SM	DVEG	1764	26	5.20	898.05	11704.5	21.3864
36	SM	SHRT	412	25	4.55	989.87	17721.0	23.7063
37	SM	SHRT	416	26	3.95	951.28	16485.2	30.5589
38	SM	SHRT	504	26	3.70	939.83	17101.3	26.8415
39	SM	SHRT	492	27	3.75	925.42	17849.0	27.7292
40	SM	SHRT	636	27	4.15	954.11	16949.6	21.5699
41	SM	TALL	1756	24	5.60	720.72	11344.6	19.6534
42	SM	TALL	1232	27	5.35	782.09	14752.4	20.3295
43	SM	TALL	1400	26	5.50	773.30	13649.8	19.5880
44	SM	TALL	1620	28	5.50	829.26	14533.0	20.1328
45	SM	TALL	1560	28	5.40	856.96	16892.2	19.2420

a) Ajustar un modelo de regresión lineal múltiple para los datos.
b) Analizar la adecuación del modelo mediante un ANOVA.
c) Realizar el diagnóstico del modelo (normalidad y homoscedasticidad mediante residuos, casos extraordinarios (outliers) y medidas de influencia, multicolinealidad y autocorrelación) e identifique las inadecuaciones que presente.
d) Analizar los datos mediante la selección de variables y proponga el modelo más adecuado.

Bibliografía

Cervantes, S. A., Rivera, G. P. y De la Paz L. J. M., 2004, *SPSS. Una Herramienta para el Análisis Estadístico de Datos*, FES Zaragoza UNAM, México, 77 pp.

Cervantes, S. A., Marques D. S. M. J., Rivera G. P., 2006, *Análisis Estadístico. Un enfoque práctico con Statgraphics*, FES Zaragoza UNAM, México, 113 pp.

Cervantes, S. A., Marques D. S. M. J., 2007, *Diseño de Experimentos. Curso Práctico*, FES Zaragoza UNAM, México, 151 pp.

Devore, J. L., 2001, *Probabilidad y Estadística para Ingeniería y Ciencias.* 5ª. edición. Ed. Thomson Learning, México.

Freund, J. E. y Simon, G. A., 1992, *Estadística Elemental*, Prentice Hall, Inc. U.S.A.

Kuehl O. R., 2001, *Diseño de experimentos. Principios estadísticos de diseño y análisis de investigación*, 2ª edición, Thomson Learning, México, 666 pp.

Marques D. S. M. J., 2004, *Probabilidad y Estadística para Ciencias Químico Biológicas*, 2ª. edición, FES Zaragoza UNAM, México.

Marques D. S. M. J., Galindo D. S. M. C, Cervantes S. A., *Análisis de Regresión. Un Enfoque Práctico*. FES Zaragoza UNAM, México, 101 pp.

Milton J. S. y J. O. Tsokos, 1987, *Estadística para biología y ciencias de la salud*, Ed. Interamericana McGraw-Hill, España, 527 pp.

Velleman, P. F. Y Hoaglin, D. C. (1981), *Applications, Basics, and Computing of Exploratory Data Analysis.* Duxbury Press, Boston, Massachusets, U.S.A.

Manejo práctico del software

de análisis estadístico R

1ª. Edición

Se imprimió en el Laboratorio de Aplicaciones

Computacionales de la FES Zaragoza

Con un tiraje de 100 ejemplares y su edición en formato

electrónico para difundirlo en los sitios:

www.sisal.unam.mx

enlinea.zaragoza.unam.mx/biomat

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES ZARAGOZA

UMDI-SISAL, FACULTAD DE CIENCIAS