Cartas sobre Estadística de la Revista Argentina de Bioingeniería

Dr. Ing. Marcelo R. Risk

Facultad Regional Buenos Aires,

Universidad Tecnológica Nacional,

Argentina

Versión 1.01

2003

ISBN 987-43-6130-1

Índice

Cartas sobre Estadística 1:
Estadística Descriptiva, Representación Gráfica y Distribución Normal.

"Es remarcable que una ciencia la cual comenzó con el estudio sobre las chances en juegos de azar, se haya convertido en el objeto más importante del conocimiento humano… las preguntas más importantes sobre la vida son, en su mayor parte, en realidad sólo problemas de probabilidad"

Pierre Simon, Marqués de Laplace (1749-1827)

Introducción

La famosa reflexión hecha por el Marqués de Laplace, muestra como a partir del estudio de algo que a simple vista puede ser considerado de una relativa importancia, se puede llegar a algo muy importante, en este caso la estadística es realmente muy importante, cualquiera sea nuestra actividad científica o tecnológica [1]. Nosotros como bioingenieros dependemos mucho de la estadística, tanto para nuestra actividad específica, como así también para interactuar con otros profesionales de la salud [2].

Los objetivos de esta serie de "Cartas sobre Estadística" son aclarar con ejemplos concretos los conceptos básicos de la estadística aplicada a la bioingeniería y la medicina; un segundo objetivo es crear un foro a través de nuestra revista, para lo cual invitamos a nuestros lectores a que tomen contacto con nosotros con preguntas, críticas, propuestas o cualquier inquietud.

Los ejemplos que se presentarán en estas cartas fueron implementados con el "Lenguaje R" [3], dicho lenguaje es un entorno con capacidad de programación y graficación, desarrollado originalmente en los laboratorios Bell por John Chambers y colegas, es fácil de usar (por lo menos para aquellos que han experimentado otros lenguajes tales como C y C++), y se ha convertido en un projecto de colaboración entre investigadores a lo largo del mundo, es gratis (algo muy importante en estos tiempos !!), se lo puedo "bajar" por Internet en el sitio oficial del projecto (R project), así como en otros sitios espejos; están disponibles versiones de R para Windows de Microsoft, Unix, Linux y MacOS [4].

Si bien el lenguaje R fue pensado para la estadística, es posible hacer procesamiento de señales, en esta área el lenguaje R compite con MatLab entre otros. El lenguaje R es interpretado, por eso el caso de necesitarse cálculos intensivos se pueden encadenar en forma dinámica (DLL) programas hechos en C, C++ y Fortran. El lenguaje R tiene un paquete básico, con las funciones más utilizadas, y paquetes adicionales, todos ellos disponibles gratis en el sitio oficial (R project). Por supuesto que recomendamos el lenguaje R, pero los ejemplos de estas cartas pueden en su mayor parte probarse con Excel [5].

No es en absoluto un objetivo de las "Cartas sobre Estadística" comparar el lenguaje R a los otros lenguajes anteriormente mencionados.

Estadística descriptiva

Una definición de "estadística descriptiva" es "describir los datos en forma concisa" [6], la forma más común de describir un conjunto de datos relacionados entre sí es reportar un valor *medio* y una dispersión alrededor de dicho valor *medio*.

Para comenzar nuestros ejemplos, necesitamos un conjunto de datos, la tabla 1 muestra la edad para cada sujeto, proveniente de un estudio de reproducibilidad de la maniobra de Valsalva [7].

En lenguaje R podemos ingresar este conjunto de datos de la siguiente forma:

```
> edad <- c(22,22,23,24,25,25,26,27,28,29,29,29,29,29,
31,31,32,33,34,35,35,35,36,38,39,39,42,42,44,44,45,45,
45,47,48,52,59,66,67,69,69)
```

Podemos verificar, al menos, que la cantidad de datos ingresados en la computadora es correcto, para eso se puede utilizar la función *length* que devuelve la longitud del vector:

```
> length(edad)
[1] 41
```

La medidad básica para describir el valor central de un conjunto de datos es el valor medio o *media* del mismo, definido por la Ec. 1:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(Ec. 1)}$$

Tabla 1: edad para cada sujeto (en años).

Sujeto	Edad	Sujeto	Edad	Sujeto	Edad	Sujeto	Edad	Sujeto	Edad
1	22	10	29	19	34	28	42	37	59
2	22	11	29	20	35	29	44	38	66
3	23	12	29	21	35	30	44	39	67
4	24	13	29	22	35	31	45	40	69
5	25	14	29	23	36	32	45	41	69
6	25	15	31	24	38	33	45		
7	26	16	31	25	39	34	47		
8	27	17	32	26	39	35	48		
9	28	18	33	27	42	36	52		

En R se calcula simplemente así:

```
> mean(edad)
[1] 38.26829
```

Aquí nos podemos preguntar: tiene sentido describir la edad en años con 5 decimales de presición ? la respuesta es *no*, a menos que tengamos una buena razón para hacerlo, entonces ese valor medio de edad se puede reportar con un redondeo a 1 decimal: 38.3 años, el criterio que se aplicó en este caso fue reportar un decimal extra al mostrado para los valores en la tabla [8].

La Segunda medida de valor "central" de un conjunto de datos es la *mediana*, definida como el valor en el medio cuando los datos son ordenados de menor a mayor [1], en lenguaje R:

```
> median(edad)
[1] 35
```

Note que la *mediana* para este conjunto de datos es un valor entero, en realidad es uno de los valores que componen el conjunto, porque no proviene de ningún cálculo, sino

de observar el valor dentro del conjunto después de haberlo ordenado. La tabla 1 muestra los valores ordenados de menor a mayor.

La media geométrica es otra forma de describir el valor central de un conjunto de datos, se define como:

$$\overline{log\, x} = \frac{1}{n}\sum_{i=1}^{n} log\, x_i \qquad \text{(Ec. 2)}$$

La media geométrica es en realidad una forma *transformada* de calcular la media, en ciertos casos como por ejemplo un conjunto de datos que proviene de mediciones múltiplo de 2, el valor medio no describe adecuadamente el valor central, vamos a volver sobre este tema más adelante.

En R se puede calcular así:

```
> mean(log(edad))
[1] 3.592762
```

La media geométrica calculada con la Ec. 2 está expresada en una escala logarítmica, generalmente es útil volver a la escala original, para lo cual el antilogaritmo del resultado de Ec. 2 nos provee dicho valor:

```
> exp(mean(log(edad)))
[1] 36.3343
```

La moda es otra forma de describir el valor central de un conjunto de datos, se calcula como el valor más frecuente, siguiendo con nuestro ejemplo:

```
> moda(edad)
[1] 29
```

En el caso que ningún dato del conjunto tenga una frecuencia mayor a 1, el resultado de la moda es nulo. El lenguaje R no proporciona una función para calcular la moda, por lo cual dicha función fue implementada, en una futura carta se describirá.

Para poder describir mejor un conjunto de datos necesitamos una medida de *dispersión* ademas de una del valor central, la más simple es el *rango*, el cual muestra los valores mínimo y máximo del conjunto de datos, en R:

```
> range(edad)
[1] 22 69
```

La varianza y el desvío estándar son las medidas de dispersión más populares, la Ec. 3 define la varinza de un conjunto de datos:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \quad \text{(Ec. 3)}$$

El desvío estándar se define como la raíz cuadrada de la varianza (Ec. 4):

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \quad \text{(Ec. 4)}$$

En el lenguaje R se pueden calcular así:

```
> var(edad)
[1] 171.5512
> sd(edad)
[1] 13.09776
```

Otras medidas de dispersión mas sofisticadas pero no menos útiles son los cuantilos o percentilos, por ejemplo la *mediana* se puede interpretar como el valor que separa los datos en dos mitades, en otras palabras el 50% de los valores es menor que la mediana, y el otro 50% es mayor que la mediana, la *mediana* puede ser considerada un caso especial de cuantilo [6]. El lenguaje R se calcula de esta forma:

```
> quantile(edad,0.5)
50%
```

```
 35
```

Como podemos apreciar en el párrafo anterior la función de R "quantile" necesita un argumento adicional al conjunto de datos, en este ejemplo es 0.5 y se denomina probabilidad (debe ser de 0 a 1). La denominación percentilo se utiliza cuando la probabilidad tiene un valor entre 0% y 100%, siendo los valores mínimo y máximo del conjunto de datos respectivamente. Por ejemplo el percentilo 10^{to} se puede calcular en R así:

```
> quantile(edad,0.1)
10%
 25
```

El percentilo 10^{to} se interpreta entonces como el valor para el cual el 10% de los datos del conjunto son menores al mismo.

El lenguaje R provee una función denominada *fivenum*, que significa *cinco números* en Castellano, propuesta por el famoso estadístico John W. Tukey [9], la cual calcula cinco valores que describen concisamente un conjunto de datos, son los valores mínimo, los percentilos 25^{to}, 50^{ta} y 75^{to}, y el valor máximo:

```
> fivenum(edad)
[1] 22 29 35 45 69
```

Otra medida de dispersión es el *coeficiente de variación*, definido con la Ec. 5:

$$CV\% = 100\% \frac{S}{\overline{x}} \qquad \text{(Ec. 5)}$$

En palabras es el cociente del desvío estándar y la media expresado en %, esta medida es fácil de interpretar, pero en la mayoría de los casos no es adecuada para comparar dos más conjuntos de datos. En R se calcula asi:

```
> 100*sd(edad)/mean(edad)
[1] 34.22613
```

Representación gráfica

La estadística descriptiva nos permite caracterizar con números un conjunto de datos, sin embargo en ciertas ocasiones un gráfico permite comunicar mejor las características de los datos. El gráfico de caja (box-plot en Inglés) es la forma gráfica de los *cinco números*, como podemos ver en la figura 1 la caja muestra los percentilos 25to y 75to , la línea en el medio de la caja es la mediana (percentilo 50ta), los extremos muestran los valores mínimo y máximo.

```
> boxplot(edad,main="Estudio reproducibilidad PV",ylab="Edad
(años)")
```

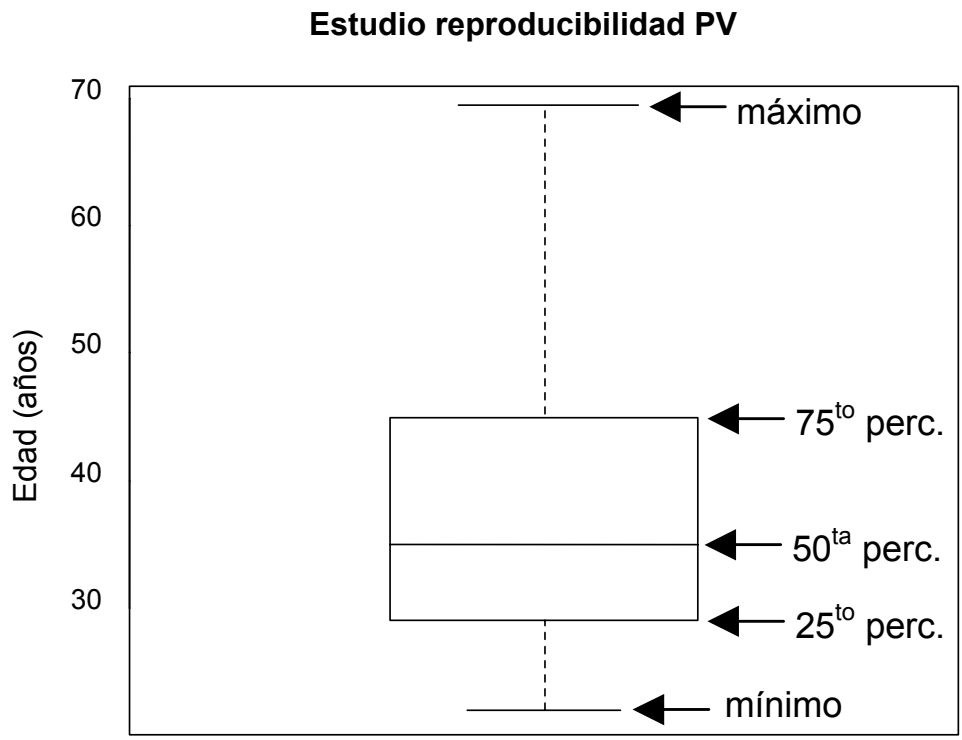

Figura 1: gráfico de caja.

Otra opción, muy utilizada por estadísticos, es el *gráfico de rama y hoja* (stem-and-leaf plot en Inglés), en lenguaje R se puede calcular de esta forma:

```
> stem(edad)
  The decimal point is 1 digit(s) to the right of the |
```

```
2 | 2234
2 | 5567899999
3 | 11234
3 | 5556899
4 | 2244
4 | 55578
5 | 2
5 | 9
6 |
6 | 6799
```

Los números que se muestran a la izquierda del carácter | son los dígitos mas significativos, y como advierte la leyenda anterior al gráfico el punto decimal está ubicado un dígito a la derecha del carácter | , en otras palabras la primera línea **2 | 2234** se lee como el primer valor 22 (por el **2 | 2**), luego hay otro 22, un 23 y un 24, correspondiente a los tres siguientes números, todos ellos corresponden a los sujetos 1 a 4 de la tabla 1.

La representación gráfica más popular de un conjunto de datos es el histograma, el cual representa la frecuencia de aparición de valores dentro del rango del conjunto de datos. La figura 2 muestra el histograma para los datos de la tabla 1, las frecuencias de aparición se calcularon para cada intervalo de 5 años dentro del rango de 20 a 70 años, totalizando 10 intervalos, por ejemplo para el primer intervalo de 20 a 25 años de encontraron 6 valores de edad (correspondientes a los sujetos 1 a 6 de la tabla 1).

El histograma en lenguaje R se calcula como se muestra a continuación, note los parámetros adicionales de la función hist para determinar el título principal (main) y los rótulos de cada eje (xlab e ylab):

```
> hist(edad,main="Estudio reproducibilidad PV",
xlab="Edad (años)",ylab="Frecuencia")
```

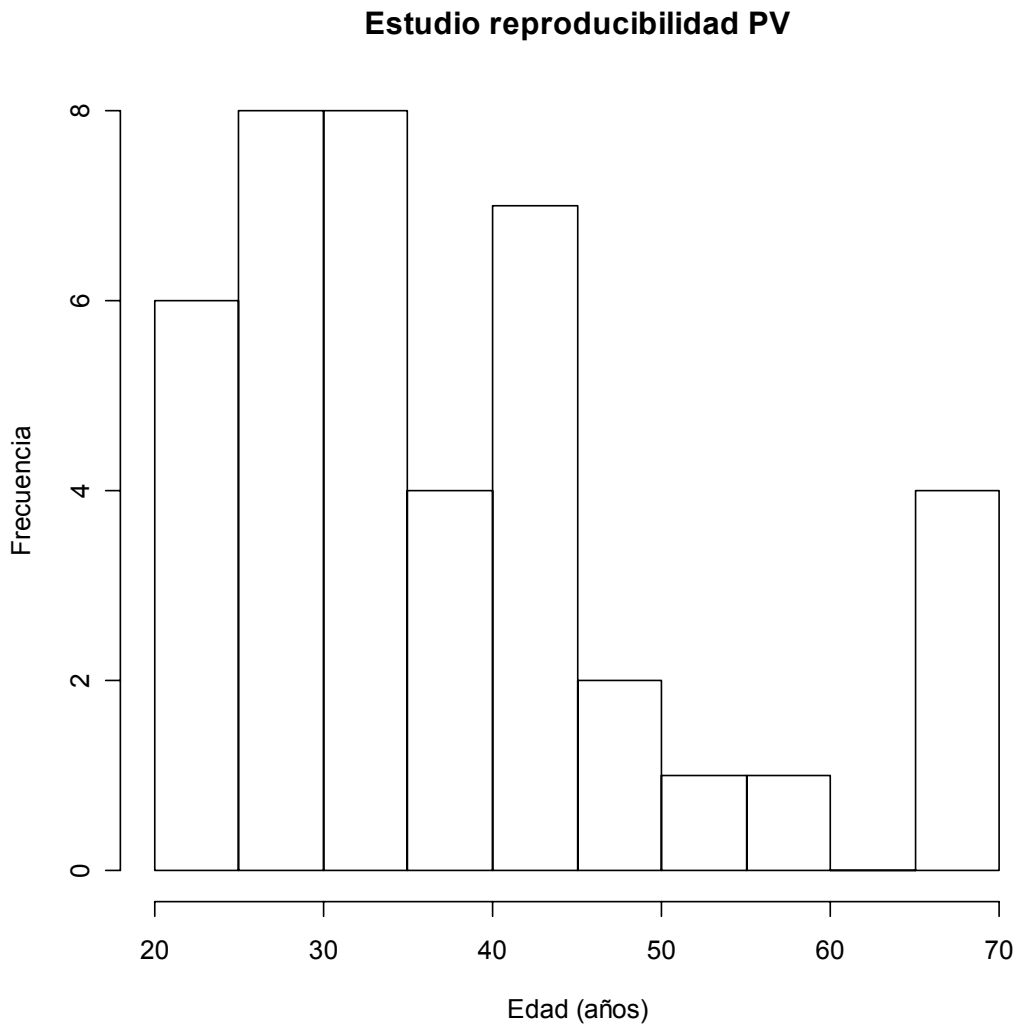

Figura 2: histograma con los datos de la tabla 1.

Distribución normal

La distribución normal, la cual lo único que tiene de normal es el nombre, fue descripta orginalmente por el matemático francés Abraham de Moivre (1667-1754), más tarde fue utilizada por Pierre Simon Laplace en una variedad de fenómenos de la ciencias naturales y sociales, pero fue el "Príncipe de los Matemáticos", el alemán Karl Gauss (1777-1855) quien aplicó la distribución normal al estudio de la forma de la tierra y los movimientos de los planetas, dicho trabajo fue tan influyente que la distribución normal se denomina con mucha frecuencia "Gaussiana".

La distribución normal se define con su función de densidad de probabilidad, como lo muestra la Ec. 6:

$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], \text{ para } -\infty < x < \infty \quad \text{(Ec. 6)}$$

donde los parámetros μ y σ son la media y el desvío estándar respectivamente. La Ec. 6 en lenguaje R se puede implementar así:

```
> mu <- 0
> sigma <- 1
> x <- c(-400:400)/100
> fx <- (1/sqrt(2*pi*sigma))*exp((x-mu)*(x-mu)/(-2*sigma*sigma))
> plot(x,fx,main="Distribución normal",type="l")
```

La figura 3 fue generada con la implementación anterior en R.

Por definición para $-\sigma<x<\sigma$ el área bajo la curva es el 68% del área total, para $-1.96\sigma<x<1.96\sigma$ es el 95% del área, y para $-2.576\sigma<x<2.576\sigma$ el 99% del área total, esto se puede apreciar en la figura 3.

La distribución normal es muy importante porque muchas pruebas estadísticas asumen que los datos tienen una *distribución normal*, por lo cual dicho conjunto de datos puede caracterizarse con dos *parámetros*, la media y el desvío estándar. Una población determinada puede tener una distribución normal, por lo cual dicha población podría eventualmente ser descripta con sus dos parámetros, una gráfica de la población en cuestión se parecería a la de la figura 3, pero todo esto no significa que una muestra de observaciones a dicha población tenga una distribución normal, esto sucede generalmente cuando la cantidad de observaciones es insuficiente.

El ejemplo que sigue muestra cómo muestras de diferentes tamaños hechas a una población normal, la función de R *rnorm* devuelve una vector de n cantidad de muestras provenientes de una población de números aleatorios con distribución normal (para más detalles sobre esta u otra función de R puede utilizar la función *help*, para este caso por ejemplo *help(rnorm)*).

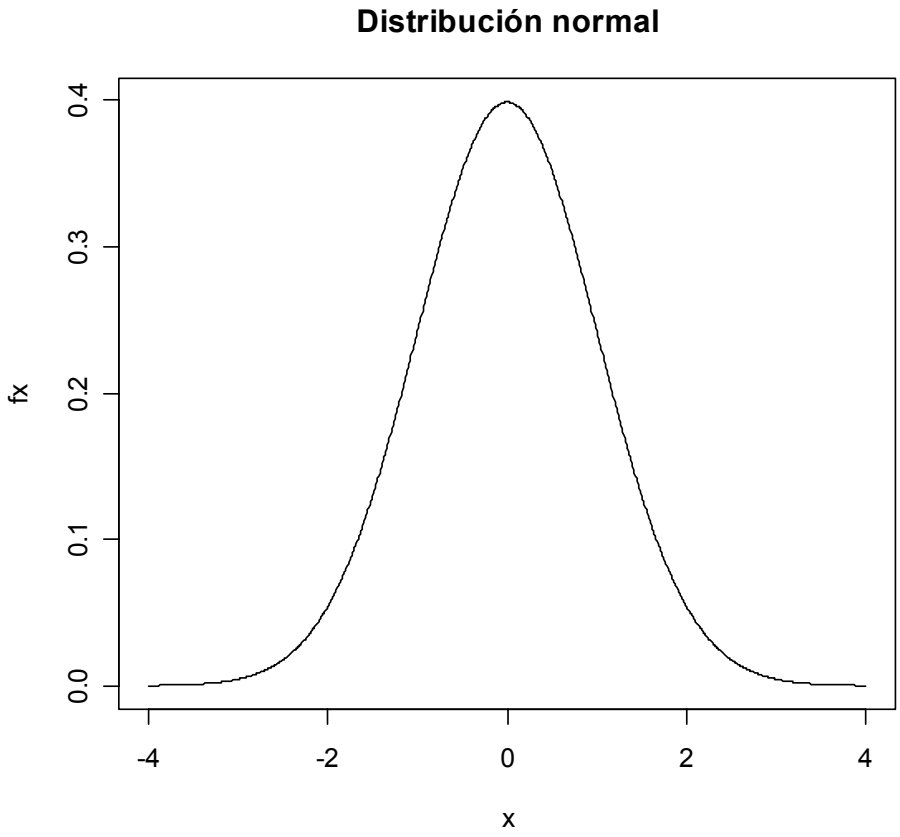

Figure 3: distribución normal calculada para $\mu=0$ y $\sigma=1$.

El ejemplo citado en el párrafo anterior se puede implementar de la siguiente forma en R:

```
> op <- par(mfrow=c(2,2))
> ruido10 <- rnorm(10)
> hist(ruido10,main="A. Histograma ruido10",ylab="Frecuencia")
> ruido50 <- rnorm(50)
> hist(ruido50,main="B. Histograma ruido50",ylab="Frecuencia")
> ruido500 <- rnorm(500)
> hist(ruido500,main="C. Histograma ruido500",ylab="Frecuencia")
> ruido1000 <- rnorm(1000)
> hist(ruido1000,main="D. Histograma
ruido1000",ylab="Frecuencia")
```

Los gráficos generados por la implementación anterior se pueden apreciar en la figura 4, donde *ruido10* es la muestra de 10 observaciones (panel A), *ruido50* tiene 50 observaciones (panel B), *ruido500* tiene 500 observaciones (panel C) y finalmente *ruido1000* tiene 1000 observaciones (panel D).

Note que a medida que aumentamos la cantidad de observaciones el histograma presenta una curva en forma de "campana" (otro de los nombres de la distribución normal), y se parece cada vez más a la distribución normal de la figura 3.

Como pudimos ver que muestras de una misma población normal pueden ser diferentes, y por lo tanto la media de las mismas también será diferente, y dichas medias pueden también tener una distribución. Un teorema fundamental de la estadística dice que las medias de muestras aleatorias provenientes de cualquier distribución tiene una distribución normal, dicho teorema se conoce con el nombre de "teorema del límite central", una consecuencia de este teorema es que cuando trabajamos con muestras de cientos de observaciones podemos olvidarnos de la distribución de la población y asumir que es normal. Una regla práctica muy utilizada dice que muestras con 30 o más observaciones tienen una distribución aceptablemente normal, como lo podemos verificar con nuestro experimento en R.

Una de las aplicaciones más importantes del teorema del límite central es la posibilidad de calcular los denominados *intervalos de confianza* (IC), el más utilizado es el IC del 95%, por ejemplo si conocemos la media (μ) el desvío estándar (σ) de una muestra por definición el 95% de los datos se encuentran dentro del intervalo determinado por μ–1.96σ y μ+1.96σ.

Comentarios

Qué podemos decir de lo visto hasta el momento ? la estadística descriptiva nos permitió caracterizar con números los datos de la tabla 1, calculando la media, el desvío estándar, el rango, los *cinco números*, y como podemos apreciar mirando la figura 2 la distribución de la edad de los sujetos provenientes de la tabla 1 *no es normal*, en realidad está bien que así sea porque para el estudio de reproducibilidad de la maniobra de Valsalva los sujetos fueron seleccionados de forma tal de cubrir aproximadamente con la misma cantidad las edades a intervalos de 5 años, para el rango de 20 a 70 años [7].

Por este motivo la muestra de la tabla 1 *no puede* ser descripta solamente con la media y el desvío estándar, es recomendable incluir la mediana y el rango, los cinco números serían inclusive más adecuados en este caso. Como comentario final podemos decir que se deberían calcular y graficar todo lo descripto en esta carta cuando estamos en presencia de datos experimentales, una vez verificado la distribución de la muestra podemos empezar a decidir que resultados se reportarán en el trabajo, así como también qué métodos estadísticos serán utilizados a posteriori.

Hasta la próxima carta: "Prueba de hipótesis e inferencias sobre una y dos muestras".

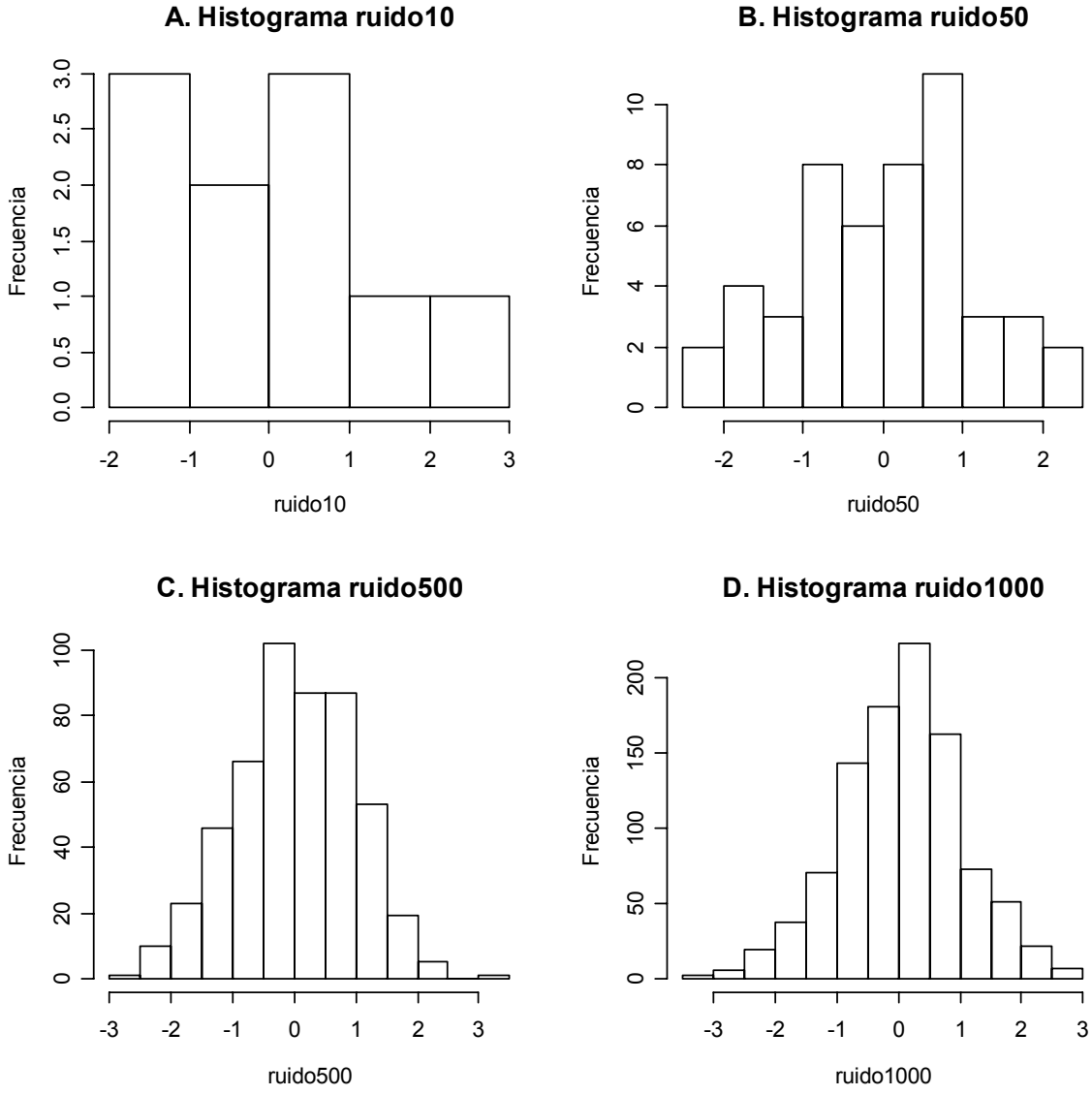

Figura 4: histogramas de 4 muestras con diferentes cantidad de observaciones.

Referencias

1. Smith G. Introduction to Statistical Reasoning. WCB/McGraw-Hill, 1998.
2. Altman DG. Practical Statistics for Medical Research. Chapman & Hall/CRC, 1999.
3. Venables WN, Smith DM. An Introduction to R, Version 1.5.1. R Development Core Team, 2002.
4. The R Project for Statistical Computing, http://www.r-project.org/.
5. Excel electronic spreadsheet. Microsoft Corporation, http://www.microsoft.com/office/.
6. Rosner B. Fundamentals of Biostatistics. Duxbury Press, 1995.
7. Risk MR. Reproducibilidad de la prueba de Valsalva. Revista Argentina de Bioingeniería. Vol a, No 1, pp 23-29, 2002.
8. Altman DG, Bland JM. Statistics notes: Presentation of numerical data. BMJ 1996, 312:572.
9. Cleveland W. The Collected Works of John W. Tukey. New York, Chapman & Hall, 1988.

Cartas sobre Estadística 2:
Prueba de hipótesis y de tendencia central sobre una y dos muestras

"Una vez uno de mis amigos me dijo que si algunos aseguraban que la tierra rotaba de este a oeste y otros que rotaba de oeste a este, siempre habría un grupo de bien intencionados ciudadanos que sugerirían que quizás haya algo de verdad de ambos lados, y que puede haber un poco de uno y otro poco del otro; o que probablemente la verdad esta entre los extremos y quizás no rota en absoluto"

Sir Maurice G. Kendall (1907-1983), estadístico británico.

Introducción

En la primera carta sobre estadística [1], revisamos los temas fundamentales sobre la estadística descriptiva, la representación gráfica y la distribución normal; la distribución normal es un tema muy importante en estadística, en esta segunda carta vamos a utilizar la misma en forma práctica, como primera prueba *antes* de cualquier otra prueba estadística. Antes de esto explicaremos la formulación y prueba de hipótesis, luego desarrollaremos como probar la normalidad de una muestra, pruebas estadísticas sobre una y dos muestras, dos muestras no normales, y finalmente una alternativa a la prueba de hipótesis utilizando el intervalo de confianza del 95% (IC 95%) de la media.

Formulación y prueba de hipótesis

El estadístico británico Sir Maurice G. Kendall expresó irónicamente una forma no científica de abordar un problema. El método científico en cambio es un proceso con el cual se investigan en forma sistemática observaciones, se resuelven problemas y se prueban hipótesis. Como parte del método científico la propuesta de una hipótesis, y luego su prueba, son temas muy bien definidos, y a pesar de la incertidumbre asociada al problema es posible cuantificar el error de la conclusion planteada por la hipótesis.

Los pasos del método científico se pueden resumir de la siguiente forma: 1) plantear el problema a resolver, 2) efectuar las observaciones, 3) formular una o mas hipótesis, 4) probar dichas hipótesis, y 5) proclamar las conclusiones; la estadística nos puede ayudar en los pasos 2) (diseño de las observaciones) y 4) (prueba de hipótesis). Una definición de hipótesis es la siguiente: "una explicación tentativa que cuenta con un conjunto de hechos y puede ser probada con una investigación posterior". La formulación de una hipótesis se logra examinando cuidadosamente las observaciones, para luego proponer un resultado posible.

Por ejemplo un problema a resolver podría ser la importancia del estado nutricional en pacientes diabéticos con complicaciones; ya tenemos el paso 1) del método científico; luego efectuamos observaciones en dos grupos de sujetos, uno control (saludables, denominados de aqui en adelante como *controles*) y otro de diabéticos con complicaciones (denominados de aqui en adelante como *pacientes*); el tamaño de dichas muestras se basa en estudios similares ya publicados y/o experiencia de los investigadores sobre y/o cálculos sobre tamaño de las muestras, este último tema sera tratado en una futura carta sobre estadística.

Uno de los indicadores más comunes del estado nutricional de una persona se puede cuantificar con el denominado índice de masa corporal (IMC), el cual se define con la siguiente ecuación [2,3]:

$$IMC = \frac{Peso[kg]}{(Altura[m])^2} \quad \text{(Ec. 1)}$$

Los valores normales (y por lo tanto saludables) del IMC van de 20 a 25 kg/m^2, valores superiores a 25 kg/m^2 y menores de 30 kg/m^2 se consideran como sobrepeso, finalmente IMC iguales o superiores a 30 kg/m^2 se consideran como indicativos de obesidad [4]. Valores altos de IMC son predictores de muerte en algunas patologías como enfermedades cardiovasculares, diabetes, cancer, hipertensión arterial y osteoartritis. La obesidad por sí sola es un factor de riesgo de muerte prematura [5].

Una vez efectuadas las observaciones en los sujetos de los dos grupos en estudio construimos las tablas con los valores de IMC para cada sujeto en cada grupo (Tablas 1 y 2); dichas tablas contienen valores simulados para los fines didácticos de la presente carta; debido al espacio limitado se presenta la siguiente tabla en formato horizontal, sin

embargo la forma más difundida y cómoda para cargar los datos las filas son sujetos y las columnas variables para facilitar tests posteriores. Por ejemplo Col1: sujeto(1...32), col2: grupo (1: control, 2: diabético), col3: IMC.

Tabla 1: IMC para cada sujeto, grupo control (en kg/m^2).

Sujeto	1	2	3	4	5	6	7	8	9
IMC	23.6	22.7	21.2	21.7	20.7	22	21.8	24.2	20.1
Sujeto	10	11	12	13	14	15	16	17	18
IMC	21.3	20.5	21.1	21.4	22.2	22.6	20.4	23.3	24.8

Tabla 2: IMC para cada sujeto (en kg/m^2), grupo de pacientes.

Sujeto	1	2	3	4	5	6	7
IMC	25.6	22.7	25.9	24.3	25.2	29.6	21.3
Sujeto	8	9	10	11	12	13	14
IMC	25.5	27.4	22.3	24.4	23.7	20.6	22.8

Por la obervación cuidadosa de las tablas 1 y 2, y utilizando métodos de estadística descriptiva y representación gráfica [1], podríamos aventurarnos a decir que el IMC en el grupo de pacientes es más alto con respecto al grupo de controles, en este punto es más seguro plantear que el IMC es distinto en lugar de mayor, aqui es donde formulamos la hipótesis:

> Hipótesis: *el IMC en pacientes con diabetes y complicaciones es distinto con respecto a sujetos saludables*.

Cabe destacar que teóricamente los sujetos de ambos debieran ser pacientes, o ambos sujetos –sanos y enfermos- para que sean comparables. La formulación *formal* de una hipótesis en el método científico se realiza definiendo la hipótesis nula (H_0) y la hipótesis alternativa (H_1); generalmente la H_0 establece que no hay diferencias entre el grupo control y el de pacientes, siguiendo con nuestro ejemplo. La hipótesis alternativa (H_1) por otra parte, suele indicarse como el complemento de la H_0, por lo tanto la H_1 establecerá que si hay diferencias entre los grupos en estudio. Por lo tanto la prueba de nuestra hipótesis consistiría en arbitrar los procedimientos necesarios para intentar rebatir

H_0, esto es, rechazarla. Esto último es válido para nuestro ejemplo en particular, en otras situaciones puede ser diferente como veremos más adelante.

H_0: IMC controles = IMC pacientes; H_1: IMC controles ≠ IMC pacientes

A la hora de tomar una decisión respecto de la hipótesis nula, surgen situaciones que nos pueden llevar a cometer diferentes errores. Así, una vez arbitradas las técnicas (o realizadas las pruebas) para probar esta hipótesis, puede que lleguemos a la conclusión de que el enunciado de nuestra H_0 sea verdadero en tal caso no rechazamos nuestra H_0 o bien que sea falso, en cuyo caso rechazaremos la H_0. En esta situación puede que hayamos rechazado la H_0 cuando en realidad era cierta, o que la evidencia no haya sido suficiente para rechazarla siendo falsa. Estas diferentes situaciones plantean la existencia de diferentes tipos de errores que se resúmen en la Tabla 3.

Una hipótesis no se acepta, simplemente la evidencia no alcanza para rechazarla, y se mantiene como cierta mientras no se rechace, este escepticismo es la base del avance del conocimiento científico.

Tabla 3: Situaciones y conclusiones posibles en la prueba de una hipótesis.

		Situación	
		H_0 verdadera	H_0 falsa
Conclusión	H_0 no rechazada	Decisión correcta	Error Tipo II (β)
	H_0 rechazada	Error Tipo I (α)	Decisión correcta (1-β)

En los casos que la H_0 se acepte y sea verdadera, así como también en el caso que H_0 se rechace y sea falsa, la desición habrá sido la correcta. Pero en los otros dos casos se producen los denominados errores tipo I y tipo II.

El error tipo I, también denominado error α, se produce cuando se rechazó la H_0 y es verdadera. Èste, representa la probabilidad de haber cometido este tipo de error. Se establece a priori α como el nivel de significancia o error máximo aceptable para la conclusión. El uso ha impuesto que en estudios de tipo clínico este error asuma valores no mayores a 0.01 o 0.05 en estudios experimentales en general. En el caso que la H_0 sea aceptada siendo falsa, se cometerá el error denominado de tipo II, o β.

El error de tipo II esta asociado con la *potencia* del método estadístico utilizado para poder detectar diferencias. La potencia de un método estadístico en una determinada situación se calcula como (1-β) , lo que se corresponde con la situación de haber rechazado correctamente la H_0, pues era falsa. Al igual que el valor de significancia α, la potencia del método estadístico se establece por el tamaño de la muestra y la prueba estadística utilizada.

En cualquier caso rechazar la H_0 es lo mismo que aceptar la H_1 y viceversa [6,7]. El resultado final de un método estadístico para la prueba de una hipótesis es el valor *P,* que indica *la* probabilidad de obtener un valor más extremo que el observado si la H_0 es verdadera. Cuando *P* es menor que α se procede a rechazar la H_0.

Pruebas de normalidad de una muestra

Antes de proceder a la prueba de una hipótesis debemos determinar la distribución las variables consideradas en nuestra muestra. En los métodos convencionales se trabaja con la distribución normal de dichas variables. El paso inicial entonces, es determinar si las variables en estudio pueden ser representadas por una distribución *normal*. En otras palabras necesitamos verificar esta primer hipótesis. O sea, si las variables medidas en la muestra pueden ser descriptas con parámetros de tendencia central y dispersión simétrica alrededor de dichos parámetros y relación media-dispersión conocidas.[1,6,7].

La importancia de verificar la normalidad de las muestras en estudio es fundamental en estadística porque si las muestras son normales se pueden aplicar métodos estadísticos paramétricos convencionales, en caso contrario se deben o bien transfromar los datos (como veremos más adelante), o bien utilizar métodos como los no paramétricos u otros métodos estadísticos más sofisticados.

Los métodos de la estadística descriptiva nos pueden ayudar a verificar la normalidad de las variables, un histograma (Figura 1A) y un gráfico de cajas (Figura 1B) nos muestra en dos formas distintas la distribución de los datos, para el ejemplo de la tabla 1 podemos decir por la forma del histograma y por lo espacios intercuartilos similares del gráfico de cajas que la muestra parece tener una distribución normal. El cálculo de los cinco números de Tukey nos muestran numéricamente que la no hay evidencia suficiente como para rechazar la distribución normal de la variable IMC [1].

Pruebas de normalidad más formales, no paramétricas, muy recomendables para verificar la normalidad de una variable son las pruebas de Shapiro-Wilk [8,9], y de Kolmogorov-Smirnov [8,10].

Contrariamente a lo que se desea en la mayoría de los casos, en las pruebas de normalidad se busca aceptar la H_0, dado que en la mayoría de los métodos estadísticos convencionales es necesaria la distribución normal de la variable de interés, pues siendo así es posible conocer los parámetros que la describen por completo, su media (μ), su desvío (σ) y la relación entre ambos y en este sentido estos métodos son más potentes. Un valor $P \geq 0.05$ en los tests de normalidad indicaría que no hay prueba suficiente para rechazar la normalidad de la variable.

Para la muestra de la tabla 1 se obtuvieron los siguientes resultados: a) media y (DE): 21.98 (1.34) kg/m^2, b) Cinco números de Tukey: 20.10, 21.10, 21.75, 22.70 y 24.80 kg/m^2, c) valores bajo las áreas de 0.5%, 2.5%, 50%, 97.5% y 99.5% en la distribución normal equivalente: 18.54, 19.36, 21.98, 24.6 y 25.42 kg/m^2 respectivamente, d) prueba Shapiro-Wilk: $P = 0.48$, y e) prueba de Kolmogorov-Smirnov: $P = 0.98$.

Como podemos apreciar en los resultados obtenidos se pudo verificar que la distribución de la variable de la tabla 1 es normal.

El siguiente código en lenguaje R fué utilizado para calcular los resultados descriptos arriba:

```
IMC <- c(23.6,22.7,21.2,21.7,20.7,22,21.8 ,24.2,20.1,
21.3,20.5,21.1,21.4,22.2,22.6,20.4,23.3,24.8)
par(mfrow=c(1,2))
hist(IMC,main="A",xlab="IMC (kg/m2)",ylab="Frecuencia")
boxplot(IMC,main="B",ymos lab="IMC (kg/m2)",ylim=c(20,25))
m.IMC <- mean(IMC)
ds.IMC <- sd(IMC)
fn.IMC <- fivenum(IMC)
n <- length(IMC)
print(paste("IMC media y ds =",round(m.IMC,2),round(ds.IMC,2)))
print("Percentilos 0, 25, 50, 75 y 100")
print(fn.IMC)
print("Valores bajo las áreas 0.5%, 2.5%, 50%, 97.5% y 99.5%")
```

```
print(paste(round(-2.576*ds.IMC+m.IMC,2),round(-
1.96*ds.IMC+m.IMC,2),
round(m.IMC,2),round(1.96*ds.IMC+m.IMC,2),round(2.576*ds.IMC+m.IM
C,2)))
sw <- shapiro.test(IMC)
print(sw)
ks <- ks.test(IMC,"pnorm",mean=mean(IMC),sd=sd(IMC))
print(ks)
```

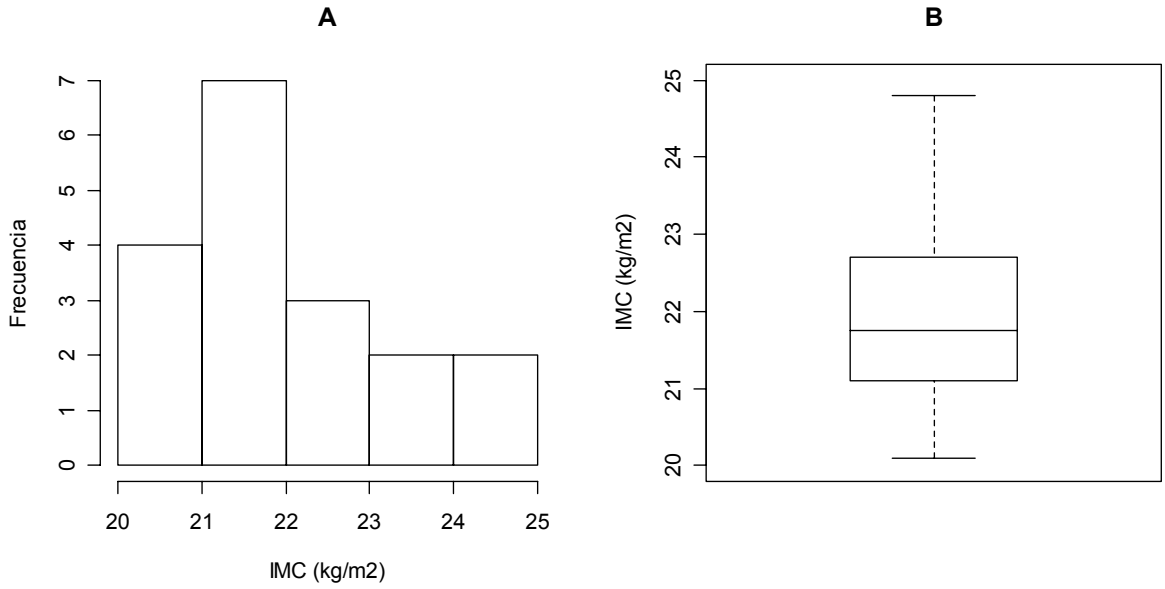

Figura 1: histograma (A) y gráfico de caja (B) para la muestra de IMC.

Pruebas sobre una muestra

Una vez verificada la normalidad, podemos realizar la prueba para verificar nuestra H_0, esto es, que la media del IMC de ambas muestras son iguales. Suponga para ello en primer lugar, que consideramos al grupo 1, los valores para la población normal de una ciudad y los del grupo 2 una muestra. Los parámetros del grupo 1 son considerados los poblacionales y los del grupo 2 los de la muestra. En este caso existe un estadístico que permite comparar la media muestral con la poblacional.

En ciertas ocasiones se cuenta con una muestra y con la media y el DE de una población, en dicho caso se utiliza la prueba sobre una muestra, una de las más poderosas

es la basada en la distribución normal [7], para la misma se debe calcular el *estadístico z* con la siguiente ecuación:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \quad \text{(Ec. 2)}$$

donde $\bar{x}$ es la media de la muestra (grupo en estudio), μ_0 es la media la población, σ es el desvío estándar de la población, y n es el tamaño de la muestra. Para los datos de la tabla 1, y comparándolos con una población descripta con $\mu_0 = 24$ y $\sigma = 4$, $z = -2.145$, el valor negativo de z simplemente significa que μ_0 es mayor que $\bar{x}$. La tabla 3 muestra los valores críticos de la distribución normal para distintos niveles de significancia α.

Como nuestra H_0 es que la muestra es igual a la población, utilizamos la columna de "dos colas"; note que para $z = 2.145$ el valor de P está entre 0.02 y 0.0456, por lo cual al ser $P < 0.05$ rechazamos la H_0 y concluimos diciendo que la muestra de la tabla 1 es distinta a la población.

Pruebas sobre dos muestras

Suponga ahora que los sujetos del grupo 1 y 2 corresponden ambos a muestras de una supuesta población subyacente. El test implicado intentará probar si ambas medias no difieren, lo que implica que ambas muestras provienen de la misma población y contrariamente si difieren.

En el caso de contar con dos muestras, para nuestro ejemplo los grupos control y de pacientes, la prueba más difundida es la "t del estudiante", publicado por el estadístico británico William Gosset (1876-1937) en 1908 bajo el seudónimo de "estudiante", según piensan algunos no lo publicó bajo su nombre porque la prueba *t* fue desarrollada como parte de su trabajo en control de calidad para la cervecería Guinness en el Reino Unido; los resultados de sus estudios probaban que algunos lotes de cerveza no eran de la calidad esperada por Guinness.

Tabla 3: Valores críticos para una distribución normal estándar.

Nivel de significancia α		
Dos colas	Una cola	z
0.001	0.0005	3.29
0.002	0.001	3.09
0.0026	0.0013	3.00
0.01	0.05	2.58
0.02	0.01	2.33
0.0456	0.0228	2.00
0.05	0.025	1.96
0.1	0.05	1.64
0.2	0.1	1.28
0.318	0.159	1.00

La prueba *t* es la prueba paramétrica más utilizada; la misma está basada en el cálculo del estadístico t y de los grados de libertad, con estos dos resultados y utlizando o bien una tabla o bien un cálculo de la distribución t se puede calcular el valor de *P*.

La prueba t del estudiante se basa en tres supuestos: a) uno es el de la distribución normal de los errores, b) es la independencia de los mismo, y c) es el de la homogeneidad de varianzas, considerado este último como el supuesto más importante.

La ecuación 3 muestra como calcular el estadístico t:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{(Ec. 3)}$$

donde $\bar{x}_1$ y $\bar{x}_2$ son las medias de cada muestra (grupos); s_1^2 y s_2^2 son las varianzas de las muestras; n_1 y n_2 son los tamaños de la muestras.

Los grados de libertad se pueden calcular con la ecuación 4:

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}} \quad \text{(Ec. 4)}$$

Como dijimos anteriormente se debe verificar la normalidad de las variables, ya lo habíamos hecho para la de la tabla 1 en el caso de los datos de la tabla 2 los resultados son los siguientes: a) Grupo controles media y (DE): 21.98 (1.34) kg/m^2, b) Grupo pacientes media y (DE): 24.38 (2.42) kg/m^2, c) prueba de Shapiro-Wilk: $P = 0.93$, y d) prueba de Kolmogorov-Smirnov: $P = 0.99$; la normalidad de las muestras fueron verificadas con las dos pruebas. La figura 2 muestra gráficamente las datos de las tablas 1 y 2 (controles y pacientes).

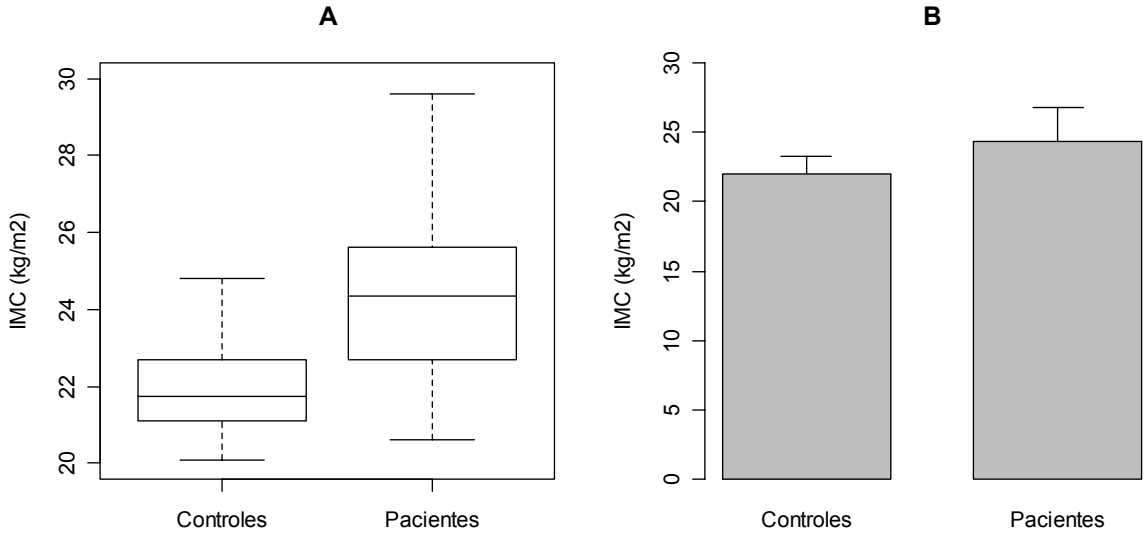

Figura 2: gráfico de cajas (A) y gráfico de barras con desvío estándar como error (B), para el IMC en los grupos control y de pacientes, $P = 0.0034$.

En lenguaje R está implementada la prueba t, el siguiente código ejemplo la calcula para las dos muestras:

```
IMC <- c(23.6,22.7,21.2,21.7,20.7,22,21.8,24.2,
20.1,21.3,20.5,21.1,21.4,22.2,22.6,20.4,23.3,24.8)
IMCp <- c(25.6,22.7,25.9,24.3,25.2,29.6,21.3,
25.5,27.4,22.3,24.4,23.7,20.6,22.8)
test.IMC <- t.test(IMC,IMCp)
print(test.IMC)
```

Los valores obtenidos son: a) t = -3.34, b) gl = 19 (redondeado), y c) P = 0.0034; como $P < 0.05$ se rechaza la H_0 y se concluye que las dos muestras son diferentes.

Pruebas sobre dos muestras apareadas

El ejemplo de la sección anterior fué sobre dos muestras provenientes de dos grupos de distintos sujetos, en ciertas ocasiones necesitamos trabajar sobre un mismo grupo de sujetos al cual se los observa en forma repetida, por ejemplo antes y después de un tratamiento, en este caso los sujetos son controles de ellos mismos. La prueba t es distinta para poder tener en cuenta que las observaciones son repetidas sobre el mismo grupo de sujetos. El primer paso es calcular el desvío estándar de las diferencia con la siguiente ecuación:

$$s = \sqrt{\sum_{i=1}^{n} \frac{\left(d_i - \bar{d}\right)^2}{n-1}} \quad \text{(Ec. 5)}$$

donde d_i es la diferencia entre dos mediciones consecutivas para cada sujeto; $\bar{d}$ es la media de las diferencias; n es la cantidad de pares de observaciones.

La ecuación 6 muestra como calcular el estadístico t para el caso de muestras apareadas:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s/\sqrt{n}} \quad \text{(Ec. 6)}$$

donde $\bar{x}_1$ y $\bar{x}_2$ son las medias de cada par de observaciones. Los grados de libertad se calcula como $gl = n\text{-}1$ (Ec.7).

La tabla 4 muestra los datos simulados (con fines didácticos), de las observaciones de la presión arterial sistólica (PAS) en un grupo de 10 pacientes antes y despúes de un tratamiento consistente en una dieta especial de bajo sodio y medicamentos.

Tabla 4: PAS antes y después del tratamiento (en mmHg).

Sujeto	PAS antes	después
1	160	139
2	155	135
3	180	175
4	140	120
5	150	145
6	130	140
7	190	170
8	192	180
9	170	149
10	165	146

La H_0 es este ejemplo es que la PAS es igual antes y después del tratamiento, por lo cual la H_1 es que la PAS es distinta para los mismos cambios.

Como siempre primero verificamos la normalidad de las variable de interés, los resultados de las pruebas Shapiro-Wilk y Kolmogorov-Smirnov fueron: a) antes del tratamiento: $P = 0.89$ y $P = 1$, y b) después del tratamiento: $P = 0.40$ y $P = 0.73$; la normalidad de las muestras es verificada.

El código en lenguaje R para calcular la prueba t para dos muestras apareadas es el siguiente:

```
PAS.antes   <- c(160,155,180,140,150,130,190,192,170,165)
PAS.despues <- c(139,135,175,120,145,140,170,180,149,146)
test.PAS <- t.test(PAS.antes,PAS.despues,paired=TRUE)
print(test.PAS)
```

El valor del estadístico t es 4.0552, con $gl = 9$, $P = 0.0029$. Con estos resultados se rechaza la H_0 y por lo tanto se concluye que la PAS antes y después del tratamiento son distintas, es decir, el tratamiento ha sido efectivo?.

La figura 3 muestra la forma correcta de representar gráficamente los estudios de dos muestras apareadas o repetidas, los gráficos de cajas y de barras (como en la figura 2) no son apropiados para este tipo de estudio porque se pierde la referencia de los cambios en cada sujeto. Como se puede apreciar en la fig. 3 en la mayoría de los sujetos el efecto del tratamiento fué la disminución de la PAS.

Pruebas sobre dos muestras no normales

Hasta el momento todos los ejemplos tuvieron distribuciones normales, por lo cual la aplicación de pruebas paramétricas normales es totalmente válido; que pasa si estamos ante muestras no normales ? la respuesta obvia es que nos olvidamos de las pruebas paramétricas y buscamos la equivalente no paramétrica, pero siempre que se pueda es aconsejable *transformar* la muestra para que sea de distribución normal y asi poder aplicar los métodos clásicos.

La transformación de la cual estamos hablando es numérica, puede ser simplemente calcular el logaritmo natural de cada observación, y luego verificar la normalidad de la muestra transformada [11]. Por lo tanto el test me dirá si los *logaritmos* de las variables difieren o no, en este caso se debería considerar si esto tiene interpretación biológica. Las tablas 5a y 5b muestran las observaciones de densidad de potencia espectral (DPE) calculados sobre los intervalos RR (RRi) provenientes de 30 minutos de ECG en reposo, en dos grupos: control y de pacientes con neuropatía autonómica cardíaca (datos simulados con fines didácticos).

En lenguaje R se puede calcular la transformación con el logaritmo natural de la siguiente forma:

```
lnDPEc <- log(DPEc)
lnDPEp <- log(DPEp)
```

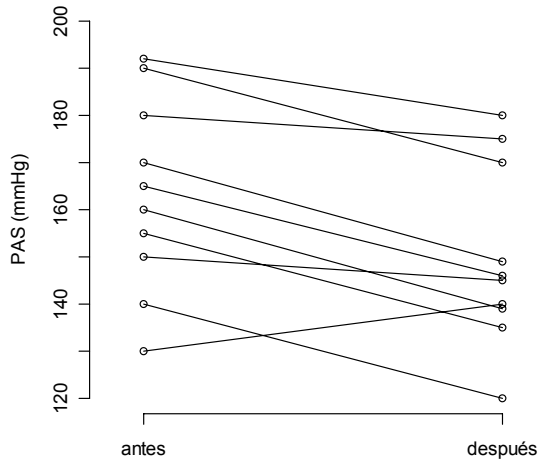

Figura 3: PAS antes y despúes del tratamiento, $P = 0.0029$.

Tabla 5a: DPE RRi grupo control (en ms^2).

Sujeto	DPE RRi	Sujeto	DPE RRi	Sujeto	DPE RRi	Sujeto	DPE RRi	Sujeto	DPE RRi
1	2098	9	2766	17	3174	25	4230	33	4739
2	2082	10	3112	18	3220	26	3707	34	4912
3	2246	11	3030	19	3464	27	4158	35	4494
4	2340	12	3375	20	3870	28	4315	36	5698
5	2714	13	3038	21	3689	29	4790	37	6349
6	2777	14	3017	22	3783	30	4464	38	6630
7	2625	15	3136	23	3457	31	4499	39	7585
8	2388	16	3204	24	4151	32	4819	40	8183

La figura 4 muestra los gráficos de cajas para los grupos de controles y pacientes en su escala original (Fig. 4A) y en la escala transformada (Fig. 4B), note que en la escala original los intervalos intercuartilos son diferentes sugiriendo distribuciones no normales.

El cálculo de los cinco números de Tukey arrojó los siguientes resultados: a) controles escala original: 2082.0, 3023.5, 3576.5, 4496.5 y 8183.0 ms^2, b) pacientes escala original: 1115.0, 1402.5, 1799.0, 2261.5 y 3627.0 ms^2, c) controles escala transformada: 7.64, 8.01, 8.18, 8.41 y 9.01 $\ln(ms^2)$, y d) pacientes escala transformada: 7.02, 7.25, 7.49, 7.72 y 8.19 $\ln(ms^2)$.

Tabla 5b: DPE RRi grupo pacientes (en ms^2).

Sujeto	DPE RRi	Sujeto	DPE RRi	Sujeto	DPE RRi	Sujeto	DPE RRi	Sujeto	DPE RRi
1	1209	9	1359	17	1661	25	2097	33	2187
2	1115	10	1337	18	1562	26	2110	34	2399
3	1151	11	1415	19	1764	27	2214	35	2630
4	1208	12	1530	20	1796	28	2069	36	2722
5	1170	13	1453	21	1976	29	2324	37	2998
6	1198	14	1324	22	1802	30	2309	38	3392
7	1390	15	1477	23	2000	31	2353	39	3379
8	1480	16	1501	24	1923	32	2091	40	3627

El resultado de las pruebas de normalidad confirma nuestras sospechas: pruebas Shapiro-Wilk: a) escala original controles y pacientes: $P = 0.0013$ y $P = 0.0034$, y b) escala transformada controles y pacientes: $P = 0.47$ y $P = 0.15$; la normalidad de las variables en la escala tranformada es verificada.

La H_0 se formula así:

la DPE del RRi en sujetos controles es igual al DPE del RRi en pacientes con neuropatía autonómica cardíaca (NAC).

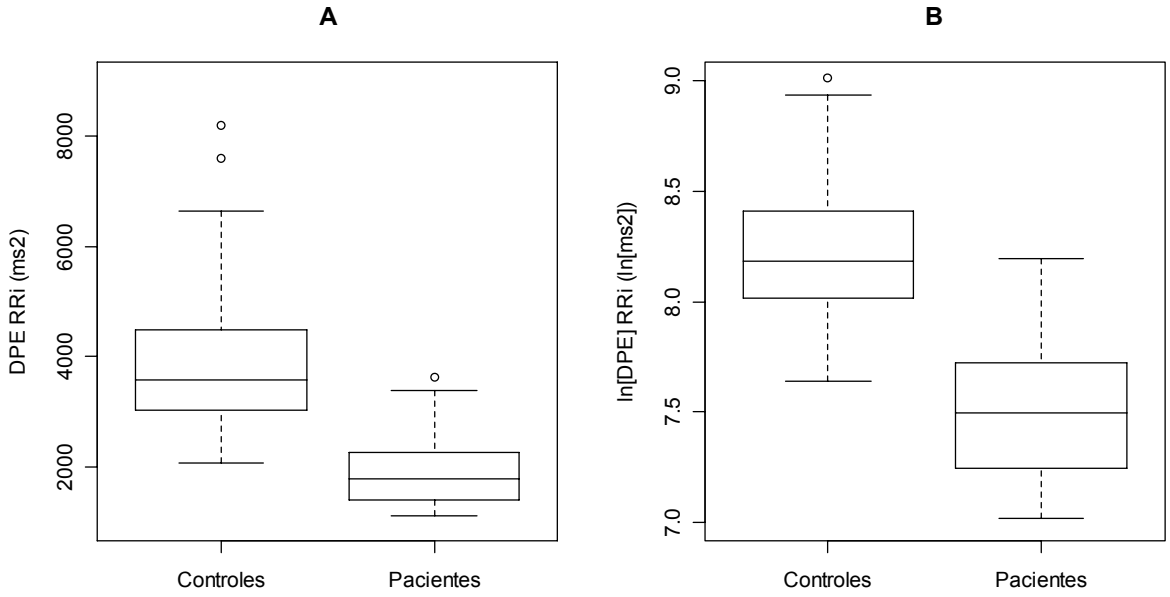

Figura 4: DPE (A) y ln[DPE] (B) en controles y pacientes.

La prueba *t* sobre las muestras transformadas brinda los siguientes resultados: t = 9.6373, gl = 78 (redondeado), P = 6.401e-15; el valor de $P < 0.05$ nos lleva a rechazar la H_0 y concluimos diciendo que el logaritmo de la DPE del RRi en pacientes con NAC es menor con respecto a la DPE del grupo control.

La figura 5 muestra la representación gráfica del estudio de DPE del RRi en controles y pacientes, note que en este ejemplo el valor de P se muestra como < 0.001 por ser muy pequeño.

Alternativa a la prueba de hipótesis

Una alternativa a las pruebas de hipotesis es el cálculo del intervalo de confianza del 95% de la media (IC 95% de la media), también denominado de la *verdadera* media, siempre que sea posible. La siguiente ecuación muestra como calcularlo:

$$IC95\%media = \bar{x} \pm t_{gl,0.05} \frac{s}{\sqrt{n}} \quad \text{(Ec. 8)}$$

donde $\bar{x}$, s y n son la media, el desvío estándar y el tamaño de la muestra respectivamente; $t_{gl,0.05}$ es el valor crítico de la distribución t para un determinado grado de libertad $gl=n-1$ y α=0.05.

Por supuesto siempre es aconsejable verificar la normalidad de las variables. Como podemos apreciar este método asume que las muestras son normales.

El siguiente código en lenguaje R muestra como calcular los IC 95% de las medias para los ejemplos del IMC de las tablas 1 y 2:

```
m.IMC <- mean(IMC)
ds.IMC <- sd(IMC)
m.IMCp <- mean(IMCp)
ds.IMCp <- sd(IMCp)

print(paste("Control   =",round(m.IMC,3),round(ds.IMC,3)))
print(paste("Pacientes =",round(m.IMCp,3),round(ds.IMCp,3)))

glc <- length(IMC)-1
tc <- qt(1-0.05/2,glc)
print(paste("tc =",tc))

glp <- length(IMCp)-1
tp <- qt(1-0.05/2,glp)
print(paste("tp =",tp))

upper95IMCc <- m.IMC+tc*ds.IMC/sqrt(length(IMC))
lower95IMCc <- m.IMC-tc*ds.IMC/sqrt(length(IMC))
print(paste("IMC IC 95% media =",round(lower95IMCc,3),
round(upper95IMCc,3)))

upper95IMCp <- m.IMCp+tp*ds.IMCp/sqrt(length(IMCp))
lower95IMCp <- m.IMCp-tp*ds.IMCp/sqrt(length(IMCp))
print(paste("IMCp IC 95% media =",
round(lower95IMCp,3),round(upper95IMCp,3)))
```

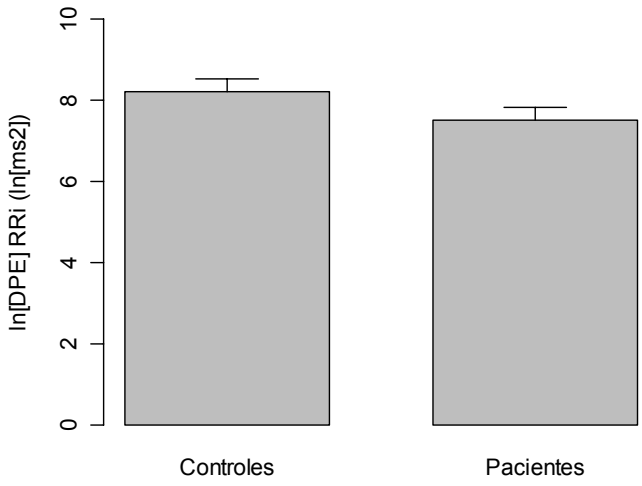

Figura 5: ln[DPE] en controles y pacientes, $P < 0.001$.

Los resultados obtenidos fueron los siguientes: a) IMC grupo de controles media y (DE): 21.99 (1.34) kg/m^2, b) IMC grupo de controles media y (DE): 24.38 (2.42) kg/m^2, c) IC 95% de la media grupo de controles: 21.31 a 22.64 kg/m^2, y d) IC 95% de la media grupo de pacientes: 22.98 a 25.77 kg/m^2.

Como podemos notar no se calculó ningún valor de *P*, tampoco se utilizaron los criterios de la tabla 3 de la prueba de hipótesis, cómo se interpretan estos resultados ? La respuesta es la siguiente: la estimación del IC 95% de la media es un método alternativo a la prueba de hipótesis, la utilización del IC 95% de la media significa que la verdadera media puede tener por chance un valor dentro de dicho intervalo de confianza, como lo muestran Beth Dawson y Robert G. Trapp [6].

En el ejemplo descripto los valores medios de cada grupo dentro de los respectivos intervalos de confianza no se solapan, con lo cual podemos concluir diciendo que son distintos, esta conclusión concuerda con la hecha anteriormente por la prueba de hipótesis. La formulación de la hipótesis fué la misma, pero la prueba de la misma fué diferente. La figura 6 muestra gráficamente los IC 95% de la media para cada grupo, de esta forma podemos ver gráficamente como llegar a la conclusión descripta arriba.

También podemos utilizar esta alternativa a la prueba de hipótesis para el ejemplo de la DPE en controles y pacientes con NAC, en este caso los resultados obtenidos fueron

los siguientes: a) IC 95% de la media grupo de controles: 8.11 a 8.32 ln(ms), y b) IC 95% de la media grupo de pacientes: 7.40 a 7.61 ln(ms).

Al igual que con la prueba de hipótesis, con el uso del IC 95% de la media, debido a que los mismos no se solapan, podemos rechazar la H_0 y concluir diciendo que la DPE del RRi en pacientes con NAC es menor con respecto a la DPE del grupo control. La figura 7 muestra gráficamente los IC 95% de la media de la DPE para los dos grupos, donde se puede apreciar una considerable separación entre los mismos.

Comentarios finales

En la presente carta se definieron conceptos fundamentales sobre la formulación y la prueba de hipótesis, así como también ejemplos prácticos para pruebas de normalidad, pruebas con una muestra, pruebas de dos muestras apareadas y no apareadas, y transformación de muestras no normales. Finalmente se describió una alternativa a la prueba de hipótesis utilizando el intervalo de confianza del 95% de la media como forma de estimar la verdadera media.

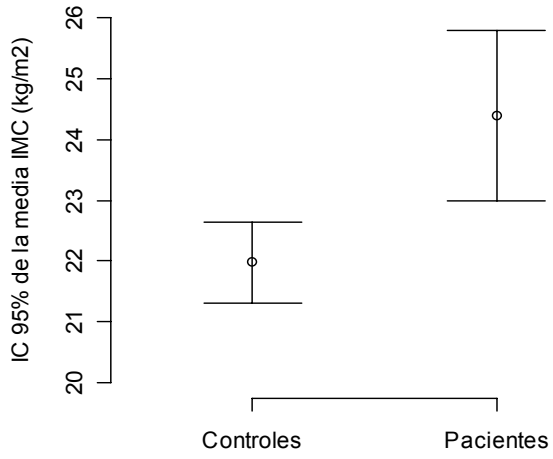

Figura 6: IC 95% de la media del IMC en controles y pacientes.

La formulación de la hipótesis se hizo en cada ejemplo con el criterio mas conservador, planteando la hipótesis nula como la igualdad entre las muestras en estudio,

por lo cual la hipótesis alternativa siempre fué la desigualdad entre las mismas. Esta forma se apoya en el concepto de las "dos colas" y no asume a priori que la diferencia es por menor o mayor [6], es aconsejable plantear las hipótesis siempre de esta forma, a menos que el conocimiento sobre el problema en estudio amerite lo contrario [11].

Un comentario especial merecen las pruebas de normalidad, a veces omitidas por algunos investigadores, pero que se consideran como fundamentales para poder verficar la normalidad de las muestras, y de esta forma poder aplicar apropiadamente las pruebas estadísticas paramétricas [6,11,12]. La prueba de normalidad de Shapiro-Wilk está considerada como más las poderosa para verificar la normalidad de una muestra [9], por lo cual algunos estadísticos consideran que por sí sola es suficiente, ese criterio fué seguido en nuestro ejemplo de dos muestras no normales.

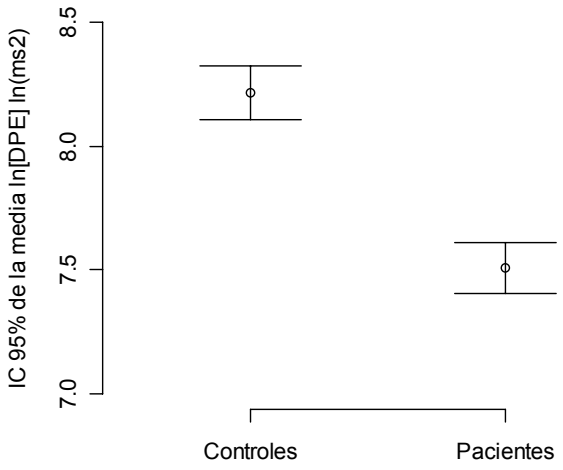

Figura 7: IC 95% del ln[DPE] del RRi en controles y pacientes.

El ejemplo de dos muestras no normales presentó una situación frecuente en el análisis estadístico de datos, el concepto de transformar los valores de las observaciones a fin de evitar el uso de pruebas no paramétricas es aconsejado por estadísticos muy prestigiosos [11], esto se debe a que las pruebas paramétricas son consideradas más poderosas que las no paramétricas [11].

La alternativa a la prueba de hipótesis utilizando el IC 95 % de la media está cada día más difundidom, en parte porque es una forma muy clara de mostrar una diferencia,

pero más importante es la posibilidad de comparar resultados provenientes de estudios diferentes, por ejemplo si un nuevo estudio sobre el IMC de un nuevo grupo de pacientes brinda una estimación de la verdadera media con el IC 95% de la media, podemos compararla con las correspondientes a nuestro estudio y sacar conclusiones; esto no hubiera sido posible a menos que contemos con tdos las observaciones del nuevo estudio y calculando nuevamente las pruebas estadísticas. Algunas revistas científicas como el British Medical Journal aconsejan siempre que sea posible reportar el IC 95% de la media [6,11]. Hasta la próxima carta: "correlación y regresión".

Agradecimientos

El autor agradece a los árbitros y al editor por sus comentarios y sugerencias, los cuales sin duda enriquecieron el presente trabajo.

Referencias

1. Risk MR. Cartas sobre estadística 1: Estadística Descriptiva, Representación Gráfica y Distribución Normal. Revista Argentina de Bioingeniería. En prensa.
2. Keen H, Thomas BJ, Jarrett RJ, Fuller JH. Nutrient intake, adiposity, and diabetes. Br Med J 1979 Mar 10;1(6164):655-8.
3. Garrow JS, Webster J. Quetelet's index (W/H^2) as a measure of fatness. International Journal of Obesity 1985;9:147-153.World Health Organization.
4. Physical status: The use and interpretation of anthropometry. Geneva, Switzerland: World Health Organization 1995. WHO Technical Report Series.
5. Calle EE, et al. BMI and mortality in a prospective cohort of U.S. adults. New England Journal of Medicine 1999;341:1097-1105.
6. Dawson B, Trapp RG. Basic and Clinical Biostatistics 3rd edition. Lange Medical Books, MacGraw-Hill. 2001.
7. Kanji GK. 100 Statistical tests. Sage Publications. 1999.
8. Venables WN, Smith DM. An Introduction to R, Version 1.5.1. R Development Core Team, 2002.
9. Royston P. A Remark on Algorithm AS 181: The W Test for Normality. Applied Statistics 1995;44:547-551.
10. Conover WJ. Practical nonparametric statistics. New York: John Wiley & Sons. 1971.

11. Altman DG. Practical Statistics for Medical Research. Chapman & Hall/CRC, 1999.
12. Smith G. Introduction to Statistical Reasoning. WCB/McGraw-Hill, 1998.

Cartas sobre Estadística 3:
Regresión y Correlación.

"Una respuesta apropiada para un problema bien formulado es mucho mejor que una respuesta exacta para un problema aproximado"

John Wilder Tukey (1915-2000), estadístico estadounidense.

Introducción

El estadístico estadounidense John Wilder Tukey realizó muy importantes contribuciones al campo de la estadística, pero para los ingenieros electrónicos y los bioingenieros quizás sea más familiar su nombre por sus contribuciones al procesamiento de señales, en dicho campo fue un pionero de la estimación espectral y del tratamiento de series temporales. La frase formulada por John W. Tukey transcripta arriba nos deja como enseñanza la importancia de formular correctamente un problema, nuestra primera preocupación debe ser esa, una vez formulado el problema en forma correcta podemos elegir el método más apropiado para resolverlo, una respuesta apropiada puede no ser exacta, como es el caso del resultado de pruebas estadísticas.

En esta tercera carta de nuestra serie de cartas sobre estadística trataremos el tema de la regresión y la correlación. En la tabla 1 podemos apreciar los métodos estadísticos más utilizados de acuerdo a la escala de las variables [1,2]; las variables se pueden dividir en dos grupos: a) variable dependiente, y b) variables independientes.

Tabla 1: métodos estadísticas más utilizados de acuerdo a la escala de las variables.

Escala de las variable dependiente	Escala de las variables independientes	Método estadístico
Intervalar	Intervalar	Regresión, múltiple en el caso de más de una variable independiente
Intervalar	Nominal u ordinal	Análisis de la varianza (ANOVA)
Intervalar	Nominal e intervalar	Análisis de la covarianza (ANCOVA)
Nominal (dicotómica)	Nominal e intervalar	Regresión logística

En la presente carta vamos a revisar la regresión, la correlación y el análisis de la covarianza; en futuras cartas revisaremos el análisis de la varianza y la regresión logística. El análisis de la varianza puede ser considerado como un caso especial de la regresión [3].

Antes de comenzar la revisión vamos a introducir la forma un conjunto de datos en lenguaje R, a través de un ejemplo (tabla 2). La tabla 2 muestra para 41 sujetos tres variables en escala intervalar (Edad, PAS: presión arterial sistólica, y SBR: sensibilidad del barorreflejo) y una variable en escala nominal (Grupo, control o paciente).

Tabla 2: Edad (años), PAS (mmHg) y SBR (ms/mmHg) y grupo (C: control, P: paciente) para cada sujeto del estudio.

Sujeto	Edad	PAS	SBR	Grupo	Sujeto	Edad	PAS	SBR	Grupo	Sujeto	Edad	PAS	SBR	Grupo
1	22	112	6.3	C	15	31	123	7.4	C	29	44	127	4	P
2	22	109	6.8	P	16	31	112	5.1	P	30	44	143	4.4	C
3	23	110	7	C	17	32	124	6	C	31	45	145	4.9	P
4	24	121	7.1	C	18	33	120	7.5	C	32	45	136	6.1	C
5	25	123	5.6	P	19	34	127	6.1	C	33	45	139	5.8	C
6	25	109	7.4	C	20	35	133	6.8	C	34	47	147	6.2	C
7	26	123	8.6	C	21	35	125	7.9	C	35	48	136	2.9	P
8	27	124	5.6	C	22	35	124	6.8	C	36	52	140	5.6	C
9	28	129	6.9	C	23	36	128	4.9	P	37	59	148	3.9	P
10	29	112	5.2	P	24	38	118	6	C	38	66	154	4.4	P
11	29	118	6.1	C	25	39	131	4.9	P	39	67	165	5.7	C
12	29	117	5.6	C	26	39	134	6.8	C	40	69	168	4.9	P
13	29	110	4.8	P	27	42	126	5.3	P	41	69	155	4.6	C
14	29	116	6.2	C	28	42	136	5.8	C					

Un conjunto de datos se define como la colección de variables (en columnas) correspondientes a sujetos (en filas); el primer paso para crear un conjunto de datos en R es definir cada una de las variables por separado, como podemos apreciar en el siguiente ejemplo de código de R:

```
Edad <- c(22,22,23,24,25,25,26,27,28,29,29,29,29,29,
31,31,32,33,34,35,35,35,36,38,39,39,42,42,44,44,45,45,
45,47,48,52,59,66,67,69,69)

PAS <- c(112,109,110,121,123,109,123,124,129,112,118,
117,110,116,123,112,124,120,127,133,125,124,128,118,
131,134,126,136,127,143,145,136,139,147,136,140,148,
154,165,168,155)

SBR <- c(6.3,6.8,7,7.1,5.6,7.4,8.6,5.6,6.9,5.2,6.1,5.6,4.8,
```

```
6.2,7.4,5.1,6,7.5,6.1,6.8,7.9,6.8,4.9,6,4.9,6.8,5.3,5.8,4,
4.4,4.9,6.1,5.8,6.2,2.9,5.6,3.9,4.4,5.7,4.9,4.6)

Grupo <- c("C","P","C","C","P","C","C","C","C","P","C","C",
"P","C","C","P","C","C","C","C","C","C","P","C","P","C","P",
"C","P","C","P","C","C","C","P","C","P","P","C","P","C")

Sujeto <- c(1:41)
```

La función de R "data.frame" concatena todas las variables en un solo conjunto de datos, también podemos guardar en un archivo dicho conjunto de datos utilizando la función "write.table". En el siguiente ejemplo podemos ver los dos pasos antes descriptos:

```
PAS.SBR <- data.frame(Sujeto,Edad,PAS,SBR,Grupo)
write.table(PAS.SBR,"c:/CartasEstadistica/RegresionData.csv",sep=
",")
```

Note que el formato del archivo es CSV (*comma separated variables*, variables separadas por comas), el cual es un formato de texto muy fácil de leer con cualquier editor de texto o con Excel.

El archivo se puede leer en R a través de la función "read.csv", por ejemplo en otra sesión de R:

```
PAS.SBR <- read.csv("c:/CartasEstadistica/RegresionData.csv")
```

Podemos verificar los datos de este conjunto utilizando la función "summary", la cual nos brinda para cada variable dentro del conjunto los valores mínimos, mediana, media, máximo y los quantilos 1ro y 3ro, todo esto para el caso de las variables en escala intervalar; en el caso de la variable en escala nominal (Grupo) nos muestra la cantidad de filas (sujetos) correspondientes a cada grupo.

```
> summary(PAS.SBR)
     Sujeto         Edad            PAS             SBR          Grupo
Min.   : 1   Min.   :22.00   Min.   :109.0   Min.   :2.900   C:27
1st Qu.:11   1st Qu.:29.00   1st Qu.:118.0   1st Qu.:4.900   P:14
Median :21   Median :35.00   Median :126.0   Median :5.800
Mean   :21   Mean   :38.27   Mean   :129.2   Mean   :5.851
3rd Qu.:31   3rd Qu.:45.00   3rd Qu.:136.0   3rd Qu.:6.800
Max.   :41   Max.   :69.00   Max.   :168.0   Max.   :8.600
```

La forma gráfica de verificar los datos se puede realizar con el código ejemplo, que utiliza la función "plot" para describir las tres variables en escala intervalar; la figura 1 muestra el resultado de la graficación. Para acceder a las variables dentro del conjunto de datos usamos el símbolo $ entre el nombre del conjunto y el nombre de la variable.

```
op <- par(mfrow = c(2, 2), pty = "m")
plot(PAS.SBR$Edad,PAS.SBR$SBR,xlim=c(20,70),pch=19,
      main="A",xlab="Edad (años)",ylab="SBR (ms/mmHg)")
plot(PAS.SBR$Edad,PAS.SBR$PAS,xlim=c(20,70),pch=19,
      main="B",xlab="Edad (años)",ylab="PAS (mmHg)")
plot(PAS.SBR$PAS,PAS.SBR$SBR,pch=19,
      main="C",xlab="PAS (mmHg)",ylab="SBR (ms/mmHg)")
```

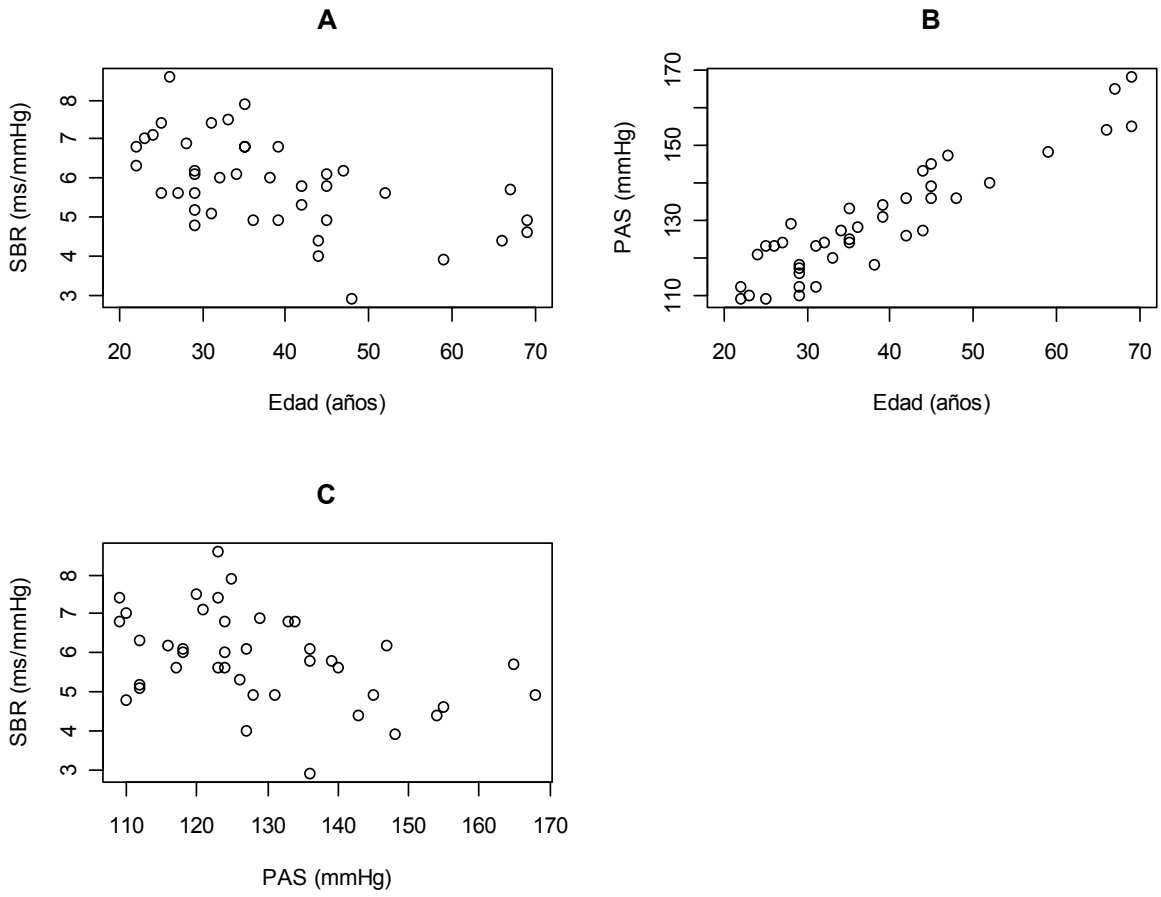

Figura 1: SBR versus edad (A), PAS versus edad (B), y SBR versus PAS (C).

Análisis de Regresión

El análisis de regresión fue introducido por Sir Francis Galton (1822-1911), científico británico de la época victoriana, el cual contribuyo al conocimento de la antropometría, la psicología diferencial, la geografía y la estadística en tre otras; además fue primo de Sir Charles Darwin. Uno de sus trabajos más influentes es "Regression Towards Mediocrity in Hereditary Stature", publicado en 1886 en el Journal of the Anthropological Institute; en dicho trabajo Francis Galton analizó graficamente la relación entre la altura de padres e hijos, concluyendo que la altura media de hijos nacidos de padres de una dada altura tienden a valores de la media de la población, luego Galton explicó esto diciendo que fue una *regresión* a los valores medios de la población.

El análisis de regresión se puede utilizar para describir la *relación*, su extensión, dirección e intensidad, entre una o varias variables independientes con escala intervalar y una variable dependiente tambien intervalar. Cabe destacar que *no* debe utilizarse el análisis de regresión como prueba de *causalidad*, en realidad no hay métodos estadísticos para probar causalidad.

La correcta aplicación del análisis de regresión asume que se cumplen las siguientes condiciones: a) existencia: para cada valor fijo de *X*, existe un valor *Y* aleatorio con una distribución de probabilidad con valores finitos de media y varianza, esta condición debe cumplirse siempre; b) independencia: los valores Y son estadisticamente independientes uno de otro, esta condición puede no cumplirse en el caso de varias observaciones (valores de *Y*) sobre un mismo valor de *X*, en ese caso debe tenerse en cuenta dicha condición; c) linearidad: el valor medio de Y es una función lineal de X; d) distribución normal: por cada valor fijo de *X*, *Y* tiene una distribuci;on normal.

En el análisis de regresión moderno, en su forma más simple es decir lineal y con una sola variable independiente (ver tabla 1), la variable independiente *X* re relaciona con la variable dependiente *Y* de acuerdo con la siguiente ecuación:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \text{Ec. 1}$$

donde ε son los residuos, es decir la diferencia entre la estimación y los valores reales para cada par de puntos *XY*. Los coeficientes de la ecuación 1 son: β_0 denominado intersección (cuando *X*=0) y β_1 pendiente.

En el siguiente ejemplo podemos apreciar el uso de la función "lm" para calcular la regresión lineal entre la Edad como variable independiente y la SBR como variable dependiente:

```
lm.SBR.Edad <- lm(SBR~Edad,data=PAS.SBR)
print(summary(lm.SBR.Edad))
plot(PAS.SBR$Edad,PAS.SBR$SBR,xlim=c(20,70),
     main="SBR versus Edad",xlab="Edad (años)",ylab="SBR
(ms/mmHg)")
abline(lm.SBR.Edad)
```

La figura 2 muestra gráficamente la relación entre Edad y SBR, junto con la recta de regresión, note en el código en R la función "abline" para graficar dicha recta.

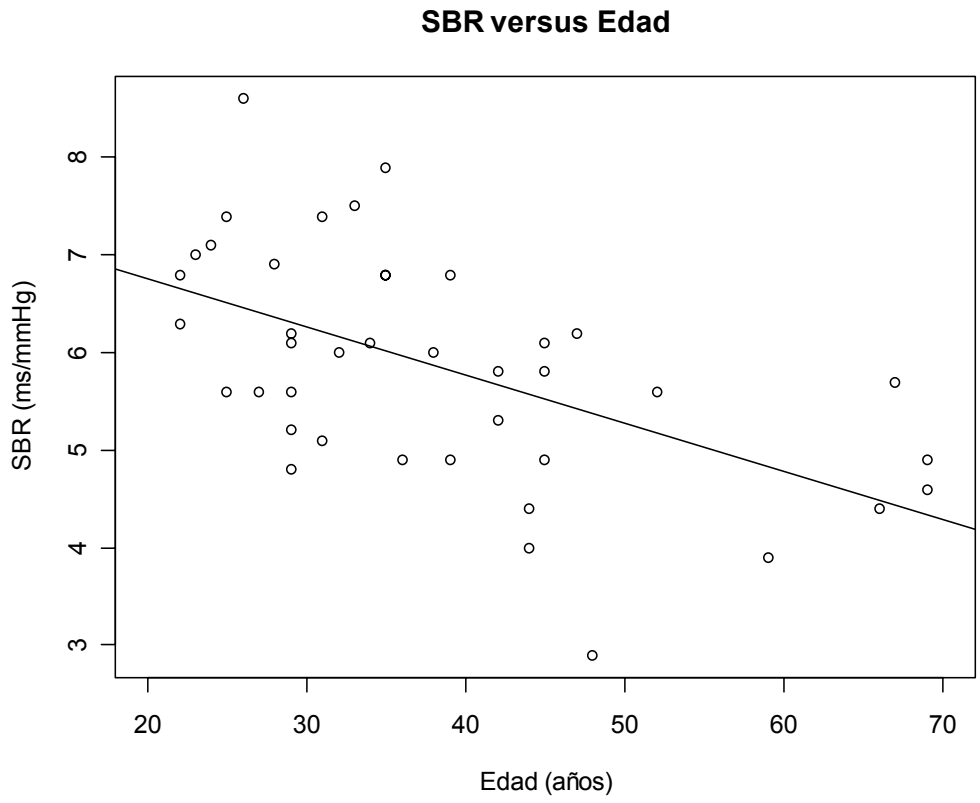

Figura 2: SBR versus edad, con la correspondiente recta de regresión.

La función "lm" utiliza la técnica denominada "método de los cuadrados mínimos" para calcular los parámetros de la regresión [3].

El resultado del análisis de regresión con la función "lm" muestra primero las variables utilizadas en la fórmula de regresión y el conjunto de datos (data) de donde provienen dichas variables. En una segunda parte muestra los valores mínimos, mediana, máximo y cuartilos 1ro y 3ro de los residuos, esto es muy importante para poder verificar la distribución de los residuos. Finalmente muestra para cada coeficiente el valor estimado, el error estándar de la estimación, el estadístico *t* y la probabilidad. El estadístico *t* se utiliza para rechazar la hipótesis nula que dice que el coeficiente es igual a cero, en el caso que no pueda rechazarce la hipótesis nula se dice que no hay asociación entre la variable independiente y la dependiente:

```
Call:
lm(formula = SBR ~ Edad, data = PAS.SBR)
Residuals:
    Min      1Q  Median      3Q     Max
-2.4694 -0.8092  0.1336  0.5821  2.1413
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.74601    0.48590  15.942  < 2e-16 ***
Edad        -0.04951    0.01203  -4.116 0.000193 ***
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

El estadístico *t* se calcula simplemente como el cociente entre la estimación y el error estándar, luego se calcula el valor de *P* de acuerdo a *n*-2 grados de libertad, siendo *n* la cantidad de pares de puntos.

Qué conclusiones podemos sacar de nuestro ejemplo ? la primera es que la estimación de la pendiente es estadísticamente significativa (*P*=0.0002) con lo cual podemos conlcluir que existe una asociación entre la Edad y la SBR, el signo negativo de la estimación nos indica que a medida que se incrementa la Edad la SBR se decrementa, a un ritmo de -0.04951 ms.mmHg^{-1}.Año^{-1}, el coeficiente de intersección es estadísticamente significativo (distinto de cero), con una estimación de 7.74 ms.mmHg^{-1}, dicho valor toma SBR cuando la Edad = 0.

Es conveniente en muchos casos verificar gráficamente los residuos para cada punto, el siguiento ejemplo muestra una forma de hacerlo, el resultado se aprecia en la figura 3.

```
plot(PAS.SBR$Edad,PAS.SBR$SBR,xlim=c(20,70),
      main="SBR versus Edad y residuos",xlab="Edad
      (años)",ylab="SBR (ms/mmHg)")
abline(lm.SBR.Edad)
interseccion <- lm.SBR.Edad$coefficients[1]
pendiente <- lm.SBR.Edad$coefficients[2]
n <- length(PAS.SBR$Edad)
for(i in 1:n){
      xx <- PAS.SBR$Edad[i]
      y1 <- interseccion+pendiente*xx
      y2 <- y1+lm.SBR.Edad$residuals[i]
      lines(c(xx,xx),c(y1,y2),lty="dotted")
      }
```

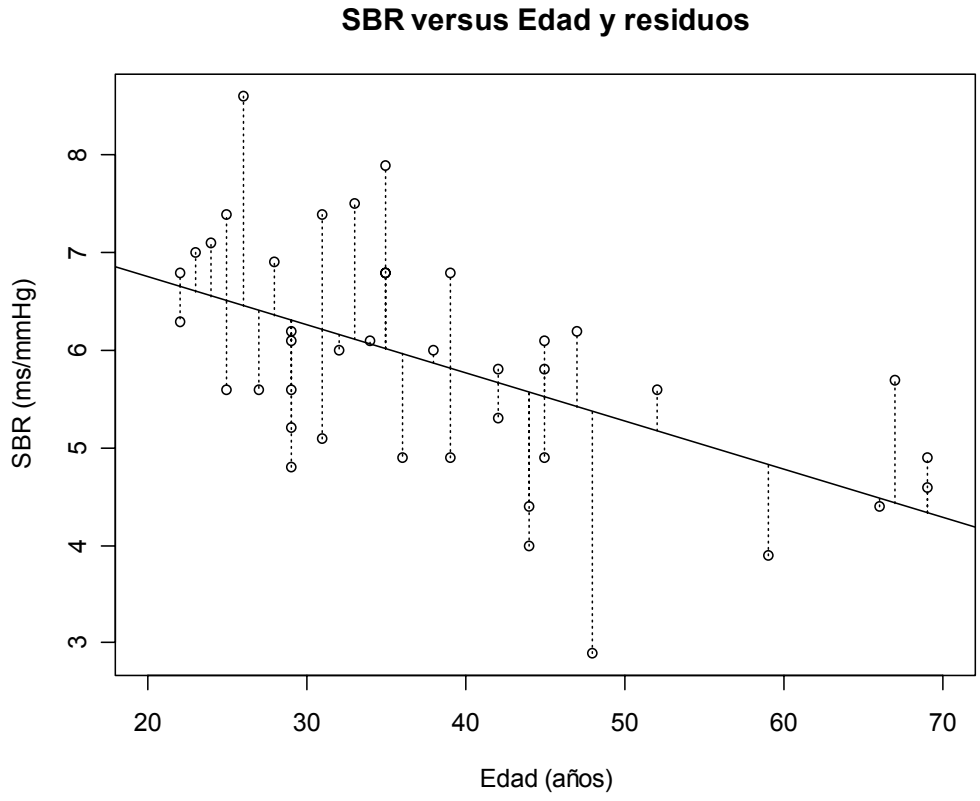

Figura 3: SBR versus edad mostrando en linea de puntos los residuos a la recta de regresión.

También podemos investigar las otras dos posibles asociaciones, la PAS versus la edad y la SBR versus la PAS. Para investigar dichas posibles asociaciones podemos hacer esto:

```
op <- par(mfrow = c(1, 2), pty = "m")

lm.PAS.Edad <- lm(PAS~Edad,data=PAS.SBR)
print(summary(lm.PAS.Edad))
plot(PAS.SBR$Edad,PAS.SBR$PAS,xlim=c(20,70),
     main="A",xlab="Edad (años)",ylab="PAS (mmHg)")
abline(lm.PAS.Edad)

lm.SBR.PAS <- lm(SBR~PAS,data=PAS.SBR)
print(summary(lm.SBR.PAS))
plot(PAS.SBR$PAS,PAS.SBR$SBR,
     main="B",xlab="PAS (mmHg)",ylab="SBR (ms/mmHg)")
abline(lm.SBR.PAS)
```

El resultado gráfico lo podemos apreciar en la figura 4, donde vemos que la PAS se incrementa con la edad, por otro lado la SBR decrementa con un un incremento de la Edad.

Los resultados del análisis de regresión para PAS versus Edad son:

```
Call:
lm(formula = PAS ~ Edad, data = PAS.SBR)
Residuals:
     Min       1Q   Median       3Q      Max
-10.9124  -3.6437  -0.2886   5.5290  10.6250
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  88.8701     2.8722   30.94   <2e-16 ***
Edad          1.0538     0.0711   14.82   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

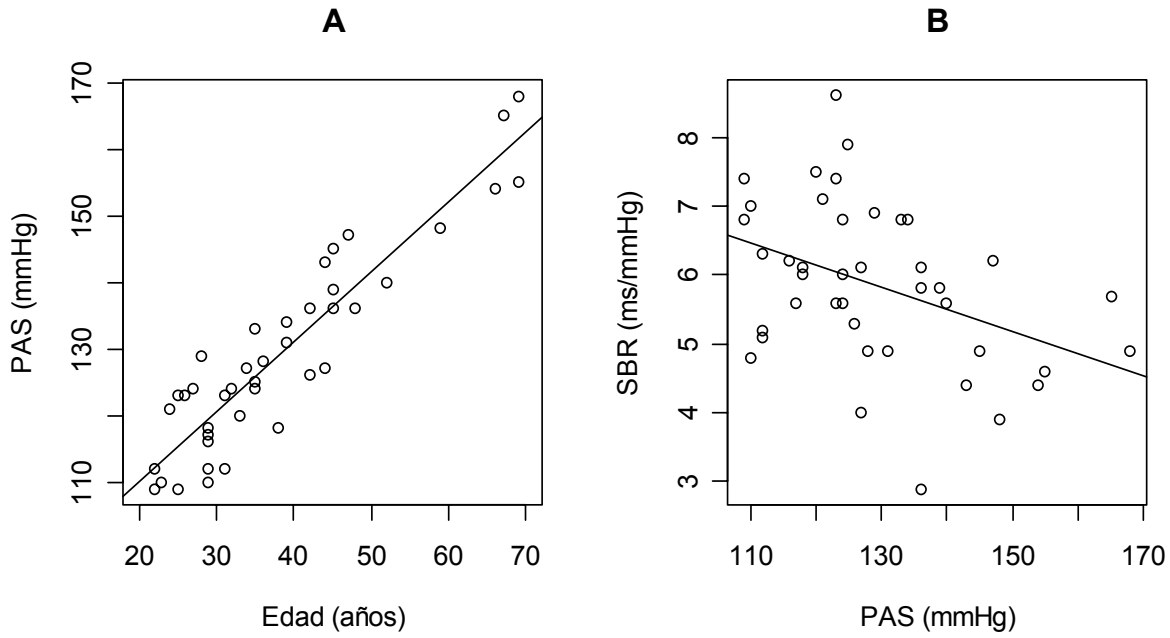

Figura 4: PAS versus edad (A), y SBR versus PAS (A), ambos con la correspondiente recta de regresión.

Podemos ver que la distribución de los residuos parece ser normal, luego que la estimación de los dos coeficientes es estadisticamente significativa para, finalmente la PAS se incrementa con la Edad a un ritmo de 1.054 mmHg.Año^{-1}, con un coeficiente de intersección de 88.9 mmHg.

Por otro lado el análisis de regresión de la SBR versus la PAS fue el siguiente:

```
Call:
lm(formula = SBR ~ PAS, data = PAS.SBR)
Residuals:
    Min      1Q  Median      3Q     Max
-2.7307 -0.6473 -0.0196  0.8942  2.5480
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.03862    1.49285   6.724 5.13e-08 ***
PAS         -0.03241    0.01148  -2.823  0.00745 **
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

En este caso los dos coeficientes son estadisticamente significativos, la SBR se decrementa con un incremento de la PAS con un ritmo de –0.0324 ms.mmHg^{-2}, y un valor de intersección de 10.04 ms.mmHg^{-1}.

Intervalos de confianza de la regresión

En algunas aplicaciones puede ser útil calcular el intervalo de confianza de la recta de regresión, de esta forma podemos verificar para una probabilidad dada (generalmente 95%), hasta donde se puede extender la recta de regresión. Los intervalos de confianza (IC) para la recta de regresión, se define como:

$$Y_{IC95\%,X_0} = \overline{Y} + \beta_1\left(X_0 - \overline{X}\right) \pm t_{n-2,1-0.95/2}\sqrt{\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{(n-1)S_X^2}} \quad \text{Ec. 2}$$

Por otro lado puede ser también conveniente calcular las bandas de predicción (BP), las cuales se definen como:

$$Y_{BP95\%,X_0} = \overline{Y} + \beta_1\left(X_0 - \overline{X}\right) \pm t_{n-2,1-0.95/2}\sqrt{1 + \frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{(n-1)S_X^2}} \quad \text{Ec. 3}$$

El siguiente código ejemplo en lenguaje R muestra la implementación de la función “regconf” para calcular y graficar los intervalos de confianza y de predicción:

```
regconf <- function(x,y,titulo,titx,tity,qq=0.975){
  plot(x,y,main=titulo,xlab=titx,ylab=tity)
  reg <- glm(y~x)
  abline(reg)
  dif <- y-reg$fitted.values
  Sxy <- sqrt(sum(dif*dif)/(length(dif)-2))
  Sx2 <- var(x)
  meanx <- mean(x)
  meany <- mean(y)
```

```
  n <- length(x)
  df <- n-2
  xs <- sort(x)
  tvalue <- qt(qq,df)
  confqqplus <- meany+reg$coefficients[2]*(xs-
meanx)+tvalue*Sxy*sqrt(1/n+((xs-meanx)*(xs-meanx)/((n-1)*Sx2)))
  confqqmin <- meany+reg$coefficients[2]*(xs-meanx)-
tvalue*Sxy*sqrt(1/n+((xs-meanx)*(xs-meanx)/((n-1)*Sx2)))
  lines(xs,confqqmin,lty="dashed")
  lines(xs,confqqplus,lty="dashed")
  confqqplus2 <- meany+reg$coefficients[2]*(xs-
meanx)+tvalue*Sxy*sqrt(1+1/n+((xs-meanx)*(xs-meanx)/((n-1)*Sx2)))
  confqqmin2 <- meany+reg$coefficients[2]*(xs-meanx)-
tvalue*Sxy*sqrt(1+1/n+((xs-meanx)*(xs-meanx)/((n-1)*Sx2)))
  lines(xs,confqqmin2,lty="dotted")
  lines(xs,confqqplus2,lty="dotted")
  }
```

La función “regconf” se puede invocar de la siguiente manera:

```
op <- par(mfrow = c(1, 1))
regconf(PAS.SBR$Edad,PAS.SBR$SBR,"SBR versus Edad con
IC95%","Edad (años)","SBR (ms/mmHg)")
```

La figura 5 muestra el resultado gráfico de la función “regconf”, note que 39 puntos de los 41 totales están dentro de las bandas de predicción del 95%, esto corresponde a un 95.1% de los puntos.

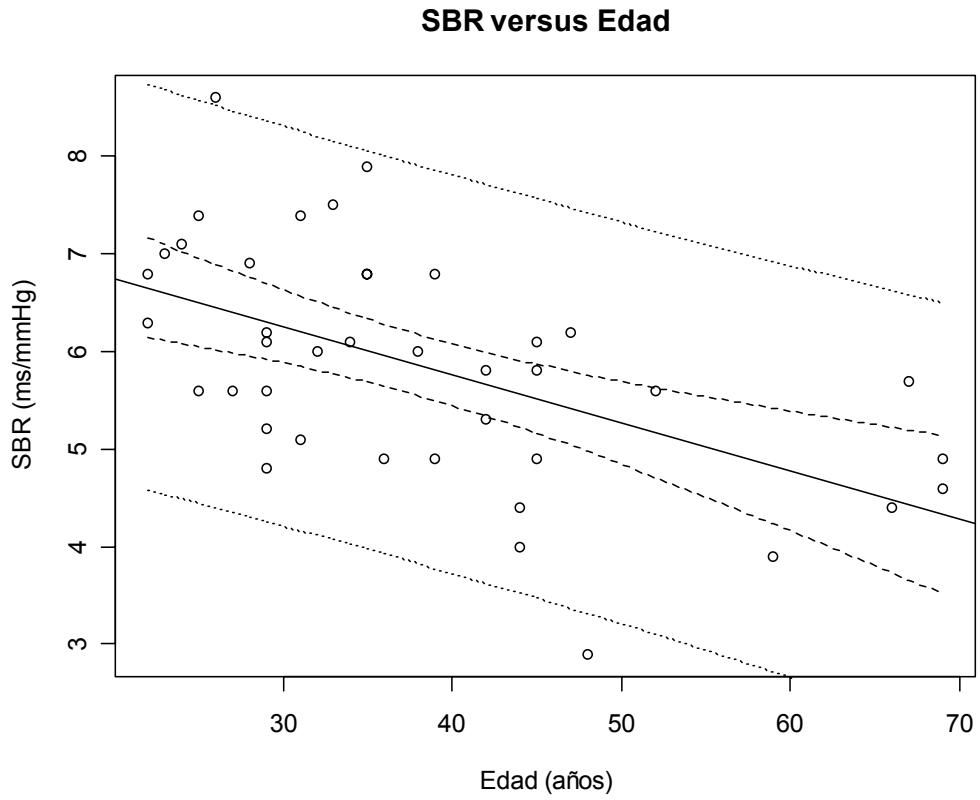

Figura 5: SBR versus Edad, con IC95% para la recta de regresión (lineas de rayas) y las bandas de predicción del 95% (lineas de puntos).

Correlación

El complento del análisis de regresión es el cálculo de la correlación, que se utiliza para cuantificar el grado de asociación de dos variables, note que en este caso no se considera a una variable dependiente y a la otra independiente, las dos tiene el mismo status. El coeficiente de correlación *r*, también denominado de Pearson, se define con la siguiente ecuación:

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} \quad \text{Ec. 4}$$

Note que *X* e *Y* pueden intervambiar su lugar y el resultado es el mismo. El coeficiente de correlación puede tener valores dentro del siguiente rango: $-1 \le r \le 1$, el signo indica la dirección de la asociación, por lo tanto las asociaciones más fuertes son –1 y 1, en el centro $r = 0$ indica falta de asociación. Otra medida derivada de *r* es el coeficiente de determinación, el cual simplemente se calcula como el cuadrado de *r* (r^2), por lo tanto el rango de valores es $0 \le r^2 \le 1$, el coeficiente de determinación multlipicado por 100 se puede interpretar como el porcentaje de pares de puntos que se pueden explicar con la recta de regresión.

El siguiente ejemplo muestra como calcular la correlación con la función "cor", para el caso de Edad versus SBR $r = -0.55$, por lo tanto el $r^2 = 0.30$, o sea explica el 30% de los casos:

```
> cor(PAS.SBR$Edad,PAS.SBR$SBR)
[1] -0.5503478
```

Es muy importante, y muchas veces omitido, el cálculo de significancia estadística del valor de *r*, para el caso de una distribución normal se puede calcular el estadístico *t* para un dado r con la siguiente ecuación:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{Ec. 5}$$

Para *gl=n-2*. Por lo tanto la hipótesis nula es que $r = 0$. La función "cor.test" calcula el coeficiente de correlación, el estadístico *t* y la *P* correspondiente:

```
> cor.test(PAS.SBR$Edad,PAS.SBR$SBR)
        Pearson's product-moment correlation
data:  PAS.SBR$Edad and PAS.SBR$SBR
t = -4.1164, df = 39, p-value = 0.0001931
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7337616 -0.2921652
sample estimates:
       cor
```

```
-0.5503478
```

Como podemos apreciar también calcula el intervalo de confianza del 95% para el valor de *r*, muy útil para comparar dos coeficientes de correlación. En el caso de no verificarse la distribución normal de los residuos, es posible calcular el coeficiente de correlación τ (tau) calculado con el método de Kendall, un método no-paramétrico:

```
> cor.test(PAS.SBR$Edad,PAS.SBR$SBR,method="kendall")
        Kendall's rank correlation tau
data:  PAS.SBR$Edad and PAS.SBR$SBR
z.tau = -3.6586, p-value = 0.0002536
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.3972351
```

Note que τ es menor que *r* en este caso, pero sin embargo está dentro del intervalo de confianza del 95% de *r*.

Análisis de la Covarianza

En los análisis anteriores pudimos verificar la asociación entre la Edad, la SBR y la PAS, pero no tuvimos en cuenta a que Grupo pertenecía cada sujeto. El método estadístico que permite incluir la variable Grupo, que tiene escala nominal, es el análisis de la covarianza [5]. El coeficiente estimado de la variable en escala nominal es el más importante en este análisis, por otro lado la variable en escala intervalar se denomina confundente. Supongamos que nuestra pregunta es: existe una asociación entre SBR y Edad ? Son diferentes para cada grupo ? para responder a esta pregunta la regresión tiene dos variables independientes: Edad y Grupo. En lenguaje R podemos calcularlo de la siguiente forma:

```
covar.SBR.Edad <- lm(SBR~Edad+Grupo,data=PAS.SBR)
print(summary(covar.SBR.Edad))
```

```
plot(PAS.SBR$Edad[PAS.SBR$Grupo=="C"],PAS.SBR$SBR[PAS.SBR$Grupo==
"C"],
     xlim=c(20,70),ylim=c(min(PAS.SBR$SBR)-
0.5,max(PAS.SBR$SBR)+0.5),
     main="SBR versus Edad covar Grupo",pch='C',
     xlab="Edad (años)",ylab="SBR (ms/mmHg)")
abline(covar.SBR.Edad$coefficients[1],covar.SBR.Edad$coefficients
[2])
points(PAS.SBR$Edad[PAS.SBR$Grupo=="P"],
     PAS.SBR$SBR[PAS.SBR$Grupo=="P"],pch='P')
abline(covar.SBR.Edad$coefficients[1]+covar.SBR.Edad$coefficients
[3],
     covar.SBR.Edad$coefficients[2],lty="dashed")
```

La figura 6 muestra graficamente el resultado del análisis de la covarianza, donde podemos ver dos rectas de regresión, corresponrientes a cada grupo (controles y pacientes), las dos rectas tienen la misma pendiente, y si la diferencia entre las rectas es estadisticamente significativa distinta de cero podemos decir que existe una diferencia entre los dos grupos.

El resultado numérico del análisis de la covarianza es el siguiente, como siempre verificamos la distribución normal de los residuos, luego vemos que la estimación de la pendiente es –0.04 ms.mmHg^{-1}.Año^{-1} (P<0.001), la estimación de GroupP es –1.34 ms.mmHg^{-1} (P<0.001), con lo cual se concluye que la SBR ajustada por la Edad en pacientes es 1.34 ms.mmHg^{-1} menor que en controles:

```
Call:
lm(formula = SBR ~ Edad + Grupo, data = PAS.SBR)
Residuals:
     Min       1Q   Median       3Q      Max
-1.68022 -0.47847 -0.03733  0.48288  1.79853
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.843286   0.377348  20.785  < 2e-16 ***
Edad        -0.040070   0.009506  -4.215 0.000148 ***
GrupoP      -1.343233   0.259345  -5.179 7.57e-06 ***
```

```
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

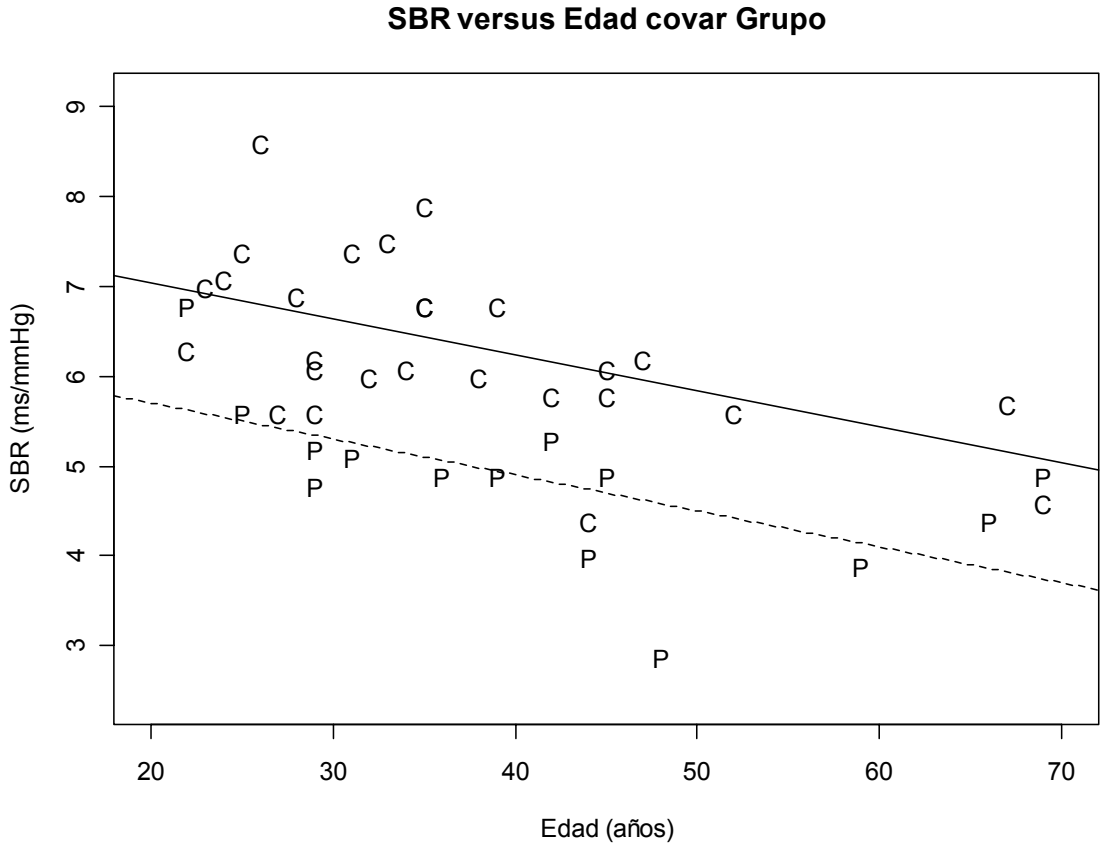

Figura 6: SBR versus Edad con Grupo como covariable, la regresión del Grupo de controles (C) se indica con una línea llena, mientras que la la del grupo de la pacientes (P) se indica con una línea de rayas.

La pendiente de la Edad versus SBR sin tener en cuenta el Grupo fue de –0.05 $ms.mmHg^{-1}.Año^{-1}$, y la intersección fue 7.75 $ms.mmHg^{-1}$; con el análisis de la covarianza las intersecciones fueron 7.84 $ms.mmHg^{-1}$ y 6.5 $ms.mmHg^{-1}$, para los grupos de controles y pacientes respectivamente.

Regresión múltiple

En los casos de más de una variable independiente la regresión se convierte en *múltiple*, con este método también es posible estudiar la interacción de dos variables calculando un coeficiente extra, por ejemplo:

$$Y = \beta_0 + \beta_1 Edad + \beta_2 PAS + \beta_3 Edad.PAS + \varepsilon \qquad \text{Ec. 6}$$

En lenguaje R se implementa de la siguiente forma:

```
lm.SBR.PAS.Edad <- lm(SBR~Edad+PAS+Edad*PAS,data=PAS.SBR)
print(summary(lm.SBR.PAS.Edad))
```

El resultado nos muestra que la SBR se decrementa a un ritmo de –0.23 ms.mmHg^{-1}.Año^{-1} (P=0.039), sin embargo los coeficientes estimados para la PAS y la interacción Edad:PAS no son significativos, podemos concluir que la SBR se decrementa significativamente con el incremento de la Edad controlado por la PAS y la interacción de la Edad y la PAS.

```
Call:
lm(formula = SBR ~ Edad + PAS + Edad * PAS, data = PAS.SBR)
Residuals:
     Min       1Q   Median       3Q      Max
-2.08159 -0.71276  0.02083  0.58461  2.02984
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.0639754  4.4991541   1.792   0.0813 .
Edad        -0.2293775  0.1073820  -2.136   0.0394 *
PAS          0.0153985  0.0380526   0.405   0.6881
Edad:PAS     0.0008936  0.0007230   1.236   0.2242
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
Residual standard error: 0.9583 on 37 degrees of freedom
Multiple R-Squared: 0.3883,     Adjusted R-squared: 0.3387
F-statistic: 7.828 on 3 and 37 DF,  p-value: 0.0003609
```

Como en todos los casos anteriores podemos verificar la normalidad de los residuos. El coeficiente de determinación múltiple, r^2_{mult}=0.39 en nuestro ejemplo, cuantifica la cercanía de los puntos al plano de regresión [6,7]; por otro lado el coeficiente de determinación ajustado tiene en cuenta la cantidad de coeficientes de la regresión múltiple [7], por lo tanto su valor es menor al r^2 múltiple, para nuestro ejemplo r^2_{adj}=0.34. Ambos coeficientes de determinación, junto con el estadístico *F* y el *P* asociado nos indican la validez de la regresión, en nuestro ejemplo se verifica con *P*<0.001. En la mayoría de los casos de regresión múltiple no es posible una representación gráfica, en algunos casos se puede graficar en forma parcial las asociaciones.

Comentarios finales

El análisis de regresión y la correlación son métodos utilizados para verificar y cuantificar la asociación entre dos o más variables. Es muy importante tener en cuenta que el análisis de regresión plantea un *modelo estadístico*, por lo tanto no es posible verificar *causalidad*, solamente podemos verificar *asociación*. Para poder verificar causalidad debemos utilizar *modelos determinísticos* [1,3]; no existen métodos estadísticos para analizar causa-efecto [1]. Como dijo John Wilder Tukey con su famosa frase el primer paso es formular correctamente el problema, por ejemplo en el estudio de la asociación entre la PAS y la Edad, la PAS es la variable dependiente y la Edad la independiente, formular el problema de esta forma nos muestra como la PAS aumenta con la Edad, si cambiamos el orden diríamos que la Edad aumenta con la PAS, esto no es correcto porque la Edad *no* depende la PAS ! La verificación de las condiciones anteriormente descriptas para poder aplicar el análisis de regresión es muy importante y muchas veces no tenidas en cuenta. Generalmente en los reportes científicos no se muestran los resultados del análisis de regresión en forma gráfica, a menos que muestre algo muy interesante, pero es muy aconsejable hacerlo para verificación personal de los investigadores.

Referencias

1. Castiglia V. Principios de Investigación Biomédica, 2da edición, 1998.
2. Dawson B, Trapp RG. Basic & Clinical Biostatistics, 3rd edition, 2001.

3. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied Regression Analysis and Other Multivariable Methods. Duxbury Press, 1998.
4. Venables WN, Smith DM and the R Development Core Team. An Introduction to R, Version 1.7.0, 2003.
5. Faraway J. Practical Regression and Anova using R. 2002.
6. Glantz SA, Slinker BK. Primer of Applied Regression and Analysis de Variance, McGraw-Hill, 1990.
7. Ryan TP. Modern Regression Analysis, John Wiley & sons, 1997.