CC-C-ES-18-273b
MAYO 2018

PATRICIO ROJAS

Estadísticas Descriptivas en R

Introducción

El objetivo de esta serie de notas técnicas es ayudarlo a analizar sus datos usando el software estadístico "R". El método que utilizaremos será describir el concepto de interés, aplicarlo a un caso sencillo, y luego ejercitarlo con un caso más complejo. Entremos en materia entonces.

Usualmente los directivos están interesados en entender mejor la relación entre distintas variables, por ejemplo el efecto que tiene la capacitación de los trabajadores en el servicio al cliente, o en la productividad. Previo a analizar la relación entre variables, es necesario conocer bien a cada una de ellas en forma independiente. Para ello son útiles las "estadísticas descriptivas", que son un conjunto de análisis estadísticos que nos ayudan a describir y entender mejor los datos o variables que deseamos analizar. Estos análisis pueden ser numéricos (mediante índices) o visuales (mediante diagramas). Algunos índices muy conocidos son el promedio, la mediana, el máximo, el mínimo, y la desviación estándar. Respecto a los diagramas, uno que se utiliza con bastante frecuencia es el histograma.

Si previo a utilizar "R" usted tiene sus datos en Excel, mi recomendación es que realice todos estos análisis en ambas herramientas para que se familiarice con las similitudes y diferencias. En esta comparación, los usuarios asiduos de Excel notarán que una de las principales diferencias entre esta herramienta y "R", es que en la segunda (i) los datos y (ii) los análisis van por "carriles" independientes. Esto, pues cuando en Excel usted quiere calcular un promedio, es cosa de que escoja la celda donde quiere que aparezca, aplique allí la función "promedio", indicando el rango de celdas al que se aplicará el cálculo. Asumiendo que los datos están en el mismo archivo, si luego de calcular el promedio usted guarda dicho archivo, la próxima vez que lo abra el promedio aparecerá en la misma celda en que usted lo calculó. Es decir, en Excel los cálculos pasan a ser parte del mismo

Nota Técnica preparada por Patricio Rojas E., profesor del ESE Business School de la Universidad de los Andes (Chile), para servir de base para la discusión en clase y no como ilustración de la gestión, adecuada o inadecuada, de una situación determinada.

archivo que contiene los datos. En "R" las cosas funcionan de otra forma. Cada vez que uno le aplica una función o comando a los datos, los resultados de dicha función no quedan asociados al archivo de datos. Simplemente aparecen en la pantalla del computador. Una forma simple de guardarlos es copiarlos y pegarlos en algún procesador de texto, por ejemplo el Block de notas, Word, o cualquiera que le sea cómodo. Otra forma es crearlos en "R" como un fichero independiente, tema que explicaremos más adelante. Por ahora, le anticipo que en "R" los ficheros pueden contener datos a analizar, o los resultados de correr un análisis específico sobre los datos de algún otro fichero.

Por el lado de las similitudes, tanto en Excel como en "R" debemos identificar a qué datos le vamos a aplicar la función o comando. En el caso de Excel lo hacemos identificando las filas y columnas donde se ubican los datos. En el caso de "R" indicamos el nombre del fichero donde están ubicados, y además le decimos cómo se llama la variable que deseamos utilizar.

Los comandos de estadísticas descriptivas en "R" son:

Comando	Descripción
summary(fichero)	Entrega varios índices para cada una de las variables del fichero identificados, entre ellas el máximo, mínimo, promedio, mediana, y la cantidad de datos faltantes (NA).
mean(fichero$variable)	Entrega el promedio de la variable de interés.
min(fichero$variable)	Entrega el valor mínimo de la variable de interés.
max(fichero$variable)	Entrega el valor máximo de la variable de interés.
sd(fichero$variable)	Entrega la desviación estándar de la variable de interés, que es una medida de la distribución y dispersión de los datos.
hist(fichero$variable)	Genera un histograma de la variable de interés, que es un gráfico de la distribución de los datos.

Además, para todos estos comandos es posible limitar el rango de valores al que se aplican. Por ejemplo, asumamos que usted tiene un fichero de nombre "datos" con información de sus trabajadores, y quiere saber la edad promedio de aquellos cuya edad se encuentra entre edad.min y edad.max. El comando en "R" se escribiría de la siguiente forma:

mean(datos$edad[datos$edad>=edad.min&datos$edad<=edad.max]).

En la práctica este comando le dice a R:

a) Quiero el promedio de los registros de la variable "edad", ubicada en el fichero "datos".
b) Cuando calcules dicho promedio, sólo ten en cuenta los registros en que la edad es mayor o igual a edad.min, y menor o igual a edad.max.

A continuación puede poner en práctica estos comandos con datos bajados de la red. Se trata de un archivo llamado "DatosPacientes.txt" que contiene los datos de edad, peso, y nivel de grasa de 25 personas. A este archivo lo llamaremos simplemente "pacientes" y lo cargaremos en "R" con el siguiente comando:

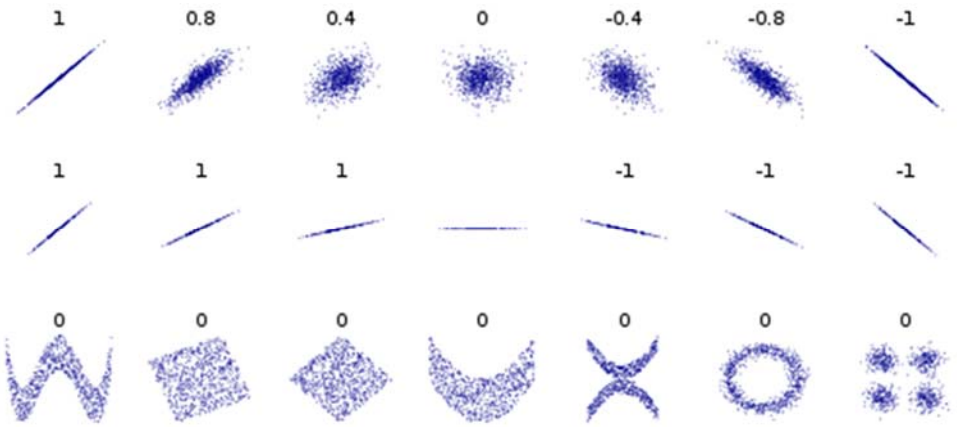
pacientes=read.table("http://crm.ese.cl/cmwkshop/DatosPacientes.txt",header = TRUE)

Tome nota de todo aquello que le llame la atención de este set de datos.

El Análisis de Correlación

Una vez que ya contamos con las estadísticas descriptivas para cada una de las variables que nos interesa analizar, podemos pasar a explorar cómo se relacionan unas con otras. Una forma de hacerlo es con el análisis de correlación, que nos da una medida de la correspondencia o proporcionalidad entre dos variables, indicando tanto el sentido como la fuerza de dicha relación.

El análisis de correlación trata de determinar qué tanto se parece la relación entre dos variables a una relación lineal, representada por una línea recta. La figura de más abajo dejará más clara la idea.



En la figura de arriba se aprecian una serie de números y gráficos x-y organizados en tres líneas. En la primera línea vemos, de izquierda a derecha, que los números van desde el 1 al -1, y que los gráficos x-y parten siendo un conjunto de puntos azules que asemejan una recta de pendiente positiva, para gradualmente convertirse en una nube de puntos, y termina como una recta de pendiente negativa.

El número es el coeficiente de correlación, que es positivo cuando la proporcionalidad es directa (si una variable aumenta, la otra también) y negativo cuando es inversa (si una variable aumenta, la otra disminuye). Es decir, el signo nos dice el sentido de la correlación, mientras que la magnitud del valor nos dice la fuerza. La segunda línea refuerza el concepto de la fuerza de la correlación. Cuando la relación se parece a una recta su valor absoluto siempre será igual al máximo, 1, sin importar la pendiente de la recta.

La tercera línea nos muestra la principal limitación del coeficiente de correlación. En cada uno de los casos el coeficiente de correlación tiene valor "0" pues la relación es distinta a una recta, sin embargo de los gráficos es evidente que en cada caso hay un patrón notorio, y por lo tanto hay una relación entre las variables. Los gráficos x-y son capaces de mostrar esta relación en términos visuales, mientras que el índice de correlación no puede. Esto sucede pues la relación es no-lineal. Dado esta complementariedad, cuando explore la relación entre dos variables siempre utilice tanto el análisis de correlación como el gráfico x-y.

Los comandos en "R" son:

Comando	Descripción
cor(fichero)	Entrega la matriz de correlación para todas las variables del fichero de interés. Si tiene un set de datos incompleto, aplique el comando cor(fichero , use="pairwise.complete.obs")
cor(fichero$variable1, fichero$variable2)	Entrega la correlación entre dos variables.
cor(fichero$variable1, fichero)	Entrega la correlación entre una variable, y el resto de las variables del fichero.
cor.test(fichero$variable1, fichero$variable2)	Realiza un test para determinar si el coeficiente de correlación entre dos variables es estadísticamente significativo, es decir evalúa si el valor obtenido es estadísticamente distinto de cero.
plot(fichero$variable1, fichero$variable2)	Genera un gráfico x-y. La primera variable va en el eje de las Xs y la segunda en el de las Ys.

A continuación puede poner en práctica estos comandos con datos que ya bajamos de la red, el fichero "pacientes". Tome nota de todo aquello que le llame la atención de este set de datos.

Testeando la "Normalidad" de los datos

En algunos casos podemos estar interesados en determinar qué tanto se asemeja la distribución de una variable a la distribución normal, también conocida como la campana de Gauss. Esta distribución describe apropiadamente muchos fenómenos físicos, biológicos, y sociales, y además tiene una serie de aplicaciones muy útiles en el análisis estadístico, algunas de las cuales abordaremos en las próximas notas técnicas. Recordemos algunas de sus propiedades más relevantes: (a) es simétrica respecto a [μ], el promedio, (b) la moda y la mediana son iguales al promedio, y siendo [σ] la desviación estándar (c) en el intervalo [$\mu-\sigma$, $\mu+\sigma$] se encuentra aproximadamente 2/3 de la distribución (68.26%), y en el intervalo [$\mu-2\sigma$, $\mu+2\sigma$] se encuentra el 95.44% de la distribución. "R" provee varias formas para determinar si la distribución de una variable corresponde a la distribución normal. Dos de las más usadas se describen a continuación. Aplíquelas a los datos que ya tenemos disponibles en el fichero "pacientes".

Comando	Descripción
shapiro.test(fichero$variable)	Ejecuta el test de Shapiro-Wilk. El p-value dice en qué grado es probable que la distribución corresponda a una normal o no. Si el valor es inferior a un umbral, por ejemplo 0.05, podemos rechazar que sea normal con una probabilidad de equivocarnos del 5%.
qqnorm(fichero$variable); qqline(fichero$variable)	Estos dos comandos generan un gráfico cuantil-cuantil de los datos, y la línea representativa de una distribución normal. La comparación visual entre ambas nos dice que tan "normales" son los datos.

Análisis Factorial

El análisis factorial es una técnica estadística avanzada que se utiliza para describir un conjunto de variables correlacionadas, en términos de una cantidad menor de variables latentes subyacentes a las que se llama factores. Por ejemplo, ocho variables correlacionadas podrían expresarse en términos de sólo dos variables latentes subyacentes. Las variables se modelan como una combinación lineal de los factores, más un error. Esta técnica se utiliza en muchas disciplinas, entre ellas la biología, la psicología, el marketing, las finanzas, y la gestión de operaciones, entre otras. Es una técnica especialmente útil cuando se dispone de un set de datos con muchas variables, y se cree que la variación de estas se debe a un pequeño conjunto de variables latentes. Esta técnica facilita en forma importante el análisis, pues una vez que se identifican los factores, el esfuerzo se concentra en ellos y no en el set completo de datos. En el ejemplo previo, en vez de analizar ocho variables, nos concentraríamos sólo en dos, los factores. Los comandos en "R" para esta técnica avanzada son:

Comando	Descripción
eigen_values <- eigen(cor(subset(fichero, select=c(var1, var2, vark))))	A partir de un fichero, toma una selección de variables (var1, var2 y vark en este ejemplo) y calcula la cantidad de factores, creando el archivo de nombre "eigen_values" con los cálculos.
eigen_values$values	Entrega los resultados del análisis previo, una serie de números. Hay tantos factores como números mayor que "1".
plot(eigen_values$values); abline(h=1, col="Red")	Genera un gráfico mostrando los factores a considerar, aquellos por sobre la línea roja.
factors_rot <- principal(subset(fichero, select = c(var1, var2, vark)), nfactors=k, rotate="varimax", scores = T)	Genera un análisis factorial sobre las variables identificadas del fichero, usando "k" factores, que guarda con nombre "factors_rot". Para usar este comando en "R" es necesario cargar y activar la librería "psych".
factors_rot$loadings	Para el análisis "factor rot", muestra el peso que tiene cada factor, en cada una de las variables. Además, muestra la porción de la variabilidad en los datos que explica cada factor.
factors_rot$scores	Para cada registro del análisis "factor_rot", muestra el valor que le corresponde de cada factor.
cor(factors_rot$scores[,posición], fichero$var_x)	Determina la correlación entre uno de los factores determinados, en este caso el ubicado en una cierta posición (de 1 a "k"), y la variable "var_x" del fichero seleccionado.

Comandos adicionales

Finalmente, algunos comandos que pueden serle de utilidad son:

- ls(): entrega un listado de todos los ficheros cargados en memoria.
- rm(fichero-1, fichero-k): borra de memoria los ficheros identificados, en este caso "fichero-1" y "fichero-k". Esto es útil para mantener "ordenado" en entorno de trabajo en "R".
- fichero[,columna]<-as.numeric(fichero[,columna]): de ser factible, transforma en valores numéricos los datos ubicados en la columna seleccionada.
- fichero[fila,columna]<-x: Ingresa el valor x en la ubicación escogida.
- pairs(fichero): genera gráficos x-y para cada una de los pares de variables contenidos en el fichero.
- getwd(): indica el directorio de trabajo que está utilizando "R".
- setwd(directorio): define el directorio en que usted desea que "R" trabaje.