Notas del curso Análisis Avanzado de Datos con R

Omar Rojas orojas@up.edu.mx
Universidad Panamericana Campus Guadalajara

6 de febrero de 2015

Resumen

Estas notas son material auxiliar para la materia *Análisis Avanzado de Datos* de la Especialidad en Optimización de Procesos Productivos de la Facultad de Ingenierías de la Universidad Panamericana Campus Guadalajara.

1. Introducción a la probabilidad y estadística univariada con R

1.1. Breve introducción a R

Ver archivo `Intro.R`

A continuación veremos el capítulo 2 de libro de G. Jay Kerns, *Introduction to Probability and Statistics Using R.*

Se pueden hacer cálculos de operaciones básicas como sigue:

```
> 2+3
```

```
[1] 5
```

```
> 10-7
```

```
[1] 3
```

Todas las variables (escalares, vectores, matrices, etc.) creadas en R se llaman objetos.

```
> x <- 2*3
> x
```

```
[1] 6
```

Existen varias estructuras posibles de objetos en R, incluyendo escalares, vectores, matrices, arreglos, panel de datos, tablas y listas.

```
> my.vector <- c(8, 6, 9, 10, 5)
```

Ejercicio 1. Despliega los siguientes elementos del vector `my.vector`

1. Primer elemento
2. último elemento
3. elementos del segundo al cuarto
4. normaliza el vector
5. suma el vector a si mismo
6. multiplica el vector por si mismo
7. calcula la media y desviación estándar del vector
8. agrega al final del vector el número -1

Ejercicio 2. Genera los siguientes vectores

1. vector de dimensión 20 con puros 1
2. vector de 100 elementos $x \sim N(0, 1)$
3. vector con elementos del 1 al 20

En contraste con un vector, una lista contiene elementos de diferentes tipos, que pueden ser numéricos y de caracteres.

```
> my.list <- list(name="Fred", wife="Mary", my.vector)
> my.list

£name
[1] "Fred"

£wife
[1] "Mary"

[[3]]
[1]  8  6  9 10  5

> my.list[[1]] == my.list$name

[1] TRUE
```

Data frames

```
> Year <- c(1800, 1850, 1900, 1950, 2000)
> Carbon <- c(8, 54, 534, 1630, 6611)
> plot(Carbon  ~ Year, pch=16)
> fossilfuel <- data.frame(year=Year, carbon=Carbon)
> fossilfuel
```

```
  year carbon
1 1800      8
2 1850     54
3 1900    534
4 1950   1630
5 2000   6611
```

Ejercicio 3. Problemas introductorios del uso de `R`

1. Utiliza el sistema de ayuda para encontrar información en `R` sobre la *media* (mean) y la *mediana* (median).
2. Obtén una lista de todas las funciones en `R` que contienen los caracteres *test*
3. Crea un vector *info* que contenga tu edad, altura (en centímetros) y número de teléfono.
4. Crea una matriz *Ident* definida como una matrix 3×3
5. Guarda tu trabajo de esta sesión en el archivo *1str.txt*

Ejercicio 4. Ejercicios de cálcuos numéricos. Usa `R` para calcular lo siguiente

1. $|2^3 - 3^2|$
2. e^e
3. $(2{,}3)^8 + \ln(7{,}5) - \cos(\pi/\sqrt{2})$
4. Let $A = \begin{pmatrix} 1 & 2 & 3 & 2 \\ 2 & 1 & 6 & 4 \\ 4 & 7 & 2 & 5 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 3 & 5 & 2 \\ 0 & 1 & 3 & 4 \\ 2 & 4 & 7 & 3 \\ 1 & 5 & 1 & 2 \end{pmatrix}$. Encuentra AB^{-1} y BA^T.

Esto es sólo un vistazo general. Para más detalles se recomiendan los siguientes libros o manuales:

1. *Nivel introductorio:*
   a) *Peter Dalgaard,* Introductory Statistics with R*, 2nd. ed., Springer, 2008*
   b) *W. J. Owen,* The R Guide*, 2007*
   c) *W. N. Venables, D. M. Smith and the R Development Core Team,* An Introduction to R*, 2009.*
   d) *G. Jay Kerns,* Introduction to Probability and Statistics Using R*, 2011*
2. *Nivel avanzado:*
   a) *John Maindonald and W. John Braun,* Data Analysis and Graphics Using R. An Example-Based Approach*, 3rd. ed., Cambridge University Press, 2003.*
   b) *Norman Matloff,* The Art of R Programming. A Tour of Statistical Software Design*, No Starch Press, 2011*

3. Análisis multivariado

a) *Brian Everitt and Torsten Hothorn,* An Introduction to Applied Multivariate Analysis with R*, Springer, 2011*

b) *Gaston Sanchez,* PLS Path Modeling with R*,*

1.2. Tipos de datos

Ejercicio 5. Investiga la clasificación de datos de acuerdo a su forma de medición.

Ejemplo 1. Precipitación de lluvia anual para algunas ciudades de E.U.A

```
> head(precip)
```

```
    Mobile      Juneau     Phoenix Little Rock Los Angeles  Sacramento
      67.0        54.7         7.0        48.5        14.0        17.2
```

1.3. Presentación gráfica de datos

```
> library(DAAG)
```

Ejercicio 6. Ejercicios de presentación de datos. Utilizando el conjunto de datos `stack.loss` disponible dentro de `R`, realiza lo siguiente:

1. Calcula la media, varianza, y el resumen de 5 números de la variable `stack.loss`
2. Crea un histograma, diagrama de caja y bigote, y gráfica de probabilidad normal para la variabla `stack.loss`. Parecen apropiadas las suposiciones de normalidad para esta muestra?

Ejercicio 7. En orden alfabético, los seis apellidos más comunes en Estados Unidos son Brown, Davis, Johnson, Jones, Smith y Williams. Suponga que una muestra de 50 personas con uno de estos apellidos proporcionó los datos siguientes Resume los datos mediante la elaboración de lo siguiente:

Brown	Williams	Williams	Williams	Brown
Smith	Jones	Smith	Johnson	Smith
Davis	Smith	Brown	Williams	Johnson
Johnson	Smith	Smith	Johnson	Brown
Williams	Davis	Johnson	Williams	Johnson
Williams	Johnson	Jones	Smith	Brown
Johnson	Smith	Smith	Brown	Jones
Jones	Jones	Smith	Smith	Davis
Davis	Jones	Williams	Davis	Smith
Jones	Johnson	Brown	Johnson	Davis

1. Distribuciones de frecuencia, frecuencia relativa y frecuencia porcentual
2. Una gráfica de barras

3. Una gráfica circular

4. Con base en estos datos, cuáles son los dos apellidos más comunes

Ejercicio 8. Comenta la diferencia entre los siguientes conceptos:

1. Estadística como hechos numéricos y estadística como ciencia

2. Estadística descriptiva e inferencial

3. Datos cualitativos y cuantitativos

Ejercicio 9. Ejercicios de simuación y distribuciones de probabilidad en `R`

1. Simula 20 observaciones de la distribución Binomial con $n = 15$ y $p = 0{,}2$

2. Encuentra el percentil 20th de la distribución Gamma con $\alpha = 2$ y $\theta = 10$

3. Encuentra $P(T > 2)$ para $T \sim t_8$

4. Grafica la función de distribución de Poisson con $\lambda = 4$ sobre el rango $x = 0, 1, \ldots, 15$

5. Simula 100 observaciones de la distribuciíon normal con $\mu = 50$ y $\sigma = 4$. Grafica la cdf $F_n(x)$ empírica para esta muestra

6. Simula 25, 250, 2500 volados de una moneda sin cargar, donde los resultados posibles son *cara* o *cruz*

7. Calcula la probabilidad de cada una de las posibles manos de poker.

Ejercicio 10. Un informe proporciona información sobre tecnología en el hogar y su uso. Los datos siguientes registran las horas de uso de computadoras personales durante una semana para una muestra de 50 persona Resume los datos mediante la elaboración de lo siguiente:

4.1	3.1	4.1	10.8	7.3
1.5	4.8	4.1	2.8	6.1
10.4	2.0	8.8	9.5	5.7
5.9	14.8	5.6	12.9	5.9
3.4	5.4	4.3	12.1	4.7
5.7	4.2	3.3	5.7	3.9
1.6	3.9	7.1	0.7	3.7
6.1	4.1	10.3	4.0	3.1
3.0	11.1	6.2	9.2	6.1
3.7	3.5	7.6	4.4	3.1

1. Una distribución de frecuencia (usa ancho de clase de 3 horas), frecuencia relativa, porcentual, acumulativa, acumulativa relativa y porcentual

2. Un histograma y una ojiva en la misma gráfica

3. Comenta qué indican los datos sobre el uso de computadoras personales en casa

Ejercicio 11. En un medio maratón se registraron 1228 corredores. La competencia se celebró en seis grupos de edades. Los datos siguientes muestras las edades de 40 individuos que participaron en la carrera.

49	33	40	37	56
44	46	57	55	32
50	52	43	64	40
46	24	30	37	43
31	43	50	36	61
27	44	35	31	43
52	43	66	31	50
72	26	59	21	47

1. Elabora un diagrama de tallo y hoja
2. Qué grupo de edad tuvo el mayor número de corredores?
3. Qué edad se registró con mayor frecuencia?

Ejercicio 12. Las siguientes 20 observaciones son para dos variables cuantitativas x y y

Observación	x	y	Observación	x	y
1	-22	22	11	-37	48
2	-33	49	12	34	-29
3	2	8	13	9	-18
4	29	-16	14	-33	31
5	-13	10	15	20	-16
6	21	-28	16	-3	14
7	-13	27	17	-15	18
8	-23	35	18	12	17
9	14	-5	19	-20	11
10	3	-3	20	-7	-22

1. Elabora un diagrama de dispersión para la relación entre x y y
2. Cuál es la relación, si existe, entre x y y?

2. Estadística descriptiva

Ejercicio 13. A continuación se presentan las ventas anuales, en millones de dólares, de 10 compañías farmacéuticas

8,408	608	10,498	3,653	1,872
1,374	14,138	7,478	5,794	4,019

Calcula o elabora lo que se pide en cada inciso:

1. (2 punts) Diagrama de caja y bigote
2. (1 punto) Media
3. (2 punts) Desviación estándar
4. (2 punts) Sesgo

Ejercicio 14. A continuación se presentan cinco observaciones tomadas para dos variables

x_i	4	6	11	3	16
y_i	50	50	40	60	30

1. (2 punts) Elabora un diagrama de dispersión de estos datos
2. (1 punts) Calcula la covarianza muestral
3. (1 punts) Determina e interpreta el coeficiente de correlación muestral

Ejercicio 15. Supón que tienes un espacio muestral con cinco resultados experimentales igualmente probables E_1, E_2, E_3, E_4, E_5. Sea

$$A = \{E_1, E_2\}, \qquad B = \{E_3, E_4\}, \qquad C = \{E_2, E_3, E_5\}$$

1. (1 puntos) Calcula $P(A), P(B)$ y $P < U + 00A9 >$
2. (1 puntos) Encuentra $P(A \cup B)$. Son A y B mutuamente excluyentes?

Ejercicio 16. Considera el experimento de arrojar un par de dados. Supón que te interesa la suma de los valores de las caras mostradas en los dados

1. (1 puntos) Elabora una lista de los puntos muestrales. Cuántos puntos de la muestra son posibles?
2. (1 puntos) Cuál es la probabilidad de obtener un valor de 7?

3. Ejercicios de Variable Aleatoria

Ejercicio 17. El empleado de un almacén regresa tres cascos de seguridad al azar a tres trabajadores de un taller siderúrgido que ya los había probado. Si Smith, Jones y Brown, en ese orden, reciben uno de los tres cascos, lista los puntos muestrales para los posibles órdenes de regreso de los cascos, y encuentra el valor x de la variable aleatoria X que representa el número de asociaciones correctas. Da la distribución de probabilidad y la distribución acumulada.

Ejercicio 18. Una variable aleatoria X tiene una media $\mu = 8$, una varianza $\sigma^2 = 9$, y distribución de probabilidad desconocida. Encuentra $P(-4 < X < 20)$

Ejercicio 19. Se selecciona un foco al azar de una caja que contiene focos de 40, 60, 75 y 100 watts. Cuál es la probabilidad de sacar un foco de 60? y de 75?

Ejercicio 20. Un embarque de 8 computadoras similares para una tienda de electr<U+009C>nicos contiene 3 que están defectuosas. Si una empresa hace una compra al azar de dos de estas computadoras, encuentra la distribución de probabilidad para el número de defectuosas

Ejercicio 21. Suponga que el error en la temperatura de reacción, en grados centígrados, para un experimento de laboratorio controlado, es una variable aleatoria continua X, que tiene la función de densidad de probabilidad

$$f(x) = \frac{x^2}{3}, \quad -1 < x < 2.$$

Encuentra $P(0 < X \leq 1)$

Ejercicio 22. Un inspector de calidad muestrea un lote que contiene 7 componentes; el lote contiene 4 componentes buenos y 3 defectuosos. El inspector toma una muestra de 3 componentes. Encuentra el valor esperado del nḿuero de componentes buenos en esta muestra.

Ejercicio 23. En un proceso de ensamble, se seleccionan tres artículos al azar, se inspeccionan y se clasifican como defectuosos o no defectuosos. Cuál es la probabilidad de encontrar dos defectuosos, si se produce un 25 % de defectuosos?

Ejercicio 24. Se sabe que en cierto proceso de fabricación, en promedio, uno de cada 100 artículos está defectuoso. Cuál es la probabilidad de que el quinto artículo que se inspecciona sea el primer defectuoso que se encuentra?

Ejercicio 25. El número promedio de camiones que llega a un CEDIS es 10. Las instalaciones pueden manejar a lo más 15 camiones por día. Cuál es la probabilidad de que en un día dado los camiones se tengan que regresar?

Ejercicio 26. Una empresa de material eléctrico fabrica focos que tienen una duración antes de fundirse, que se distribuye normalmente con $\mu = 800$ horas y $\sigma = 40$ horas. Encuentra la probabilidad de que un foco se funda entre 778 y 834 horas.

Ejercicio 27. En un proceso industrial el diámetro de un cojinete es una parte componente importante. El comprador establece que las especificaciones en el diámetro sean 3 ± 0.01 cm. La implicación es que no se aceptará ninguna parte que quede fuera de estas especificaciones. Se sabe que en el proceso el diámetro de un cojinete tiene una distribución normal con $\mu = 3$ y $\sigma = 0.005$. En promedio, cuántos cojinetes fabricados se descartarán?

Ejercicio 28. Suponga que un sistema contiene cierto tipo de componente cuyo tiempo de operación antes del fallo, en años, está dado por T. Su tiempo medio de operación antes del fallo es de 5 años. Si se instalan 5 de estos componentes en diferentes sistemas, cuál es la probabilidad de que al menos dos aún funcionen al final de 8 años?

Ejercicio 29. Suponga que las llamadas telef<U+009C>nicas que llegan a un conmutador particular siguen un proceso de Poisson con un promedio de 5 llamadas entrantes por minuto. Cuál es la probabilidad de que transcurra a lo más un minuto hasta que lleguen 2 llamadas al conmutador?

4. Estadística Inferencial

4.1. Distribuciones de Muestreo

A menudo, los parámetros*, como la media* (μ) *o la desviación estándar* (σ)*, son desconocidos y por lo tanto estimados utilizando* estadísticos *calculados usando un muestreo aleatorio tomado de la población de interés.*

Definición 1. Una *población* consiste en la totalidad de las observaciones en las que estamos interesados

Cada observación en una población es un valor de una v.a. X que tiene alguna distribución de probabilidad $f(x)$.

Definición 2. Una *muestra* es un subconjunto de una población

Cualquier procedimiento de muestreo que produzca inferencias que sobreestimen, o subestimen, de forma consistente alguna característica de la población se dice que está sesgado*. Es deseable elegir una* muestra aleatoria*, de acuerdo a la siguiente definición*

Definición 3. Sean $X_1, X_2, \ldots, X_n$ v.a. i.i.d. Estas constituyen una *muestra aleatoria* de tamaño n de la población $f(x)$ y escribimos su distribución de probabilidad conjunta como

$$f(x_1, x_2, \ldots, x_n) = f(x_1)\ f(x_2) \cdots f(x_n)$$

Nuestro principal propósito al seleccionar muestras aleatorias consiste en obtener información acerca de los pará metros desconocidos de la población.

Definición 4. Cualquier función de las variables aleatorias que forman una muestra aleatoria se llama *estadístico*

Ejemplo 2. Si $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de tamanõ n, entonces la media muestra $\bar{X}$ y la varianza muestral S^2 son estadísticos.

Puesto que un estadístico es una v.a., tiene una pdf. A la pdf de un estadístico se le llama distribución de muestreo*.*

4.1.1. Distribuciones muestrales de medias

Para una muestra aleatoria de tamaño n tomada de una $N(\mu, \sigma^2)$ tenemos que $\bar{X} = \frac{1}{n}\sum_i X_i$ tiene una distribución normal con media $\mu_{\bar{X}} = \mu$ y varianza $\sigma^2_{\bar{X}} = \sigma^2/n$. En general, si tomamos una muestra aleatoria de una población con distribución desconocida, la distribución de $\bar{X}$ será normal con los pará metros anteriores, siempre que la muestra sea grande ($n > 30$)*, de acuerdo al siguiente*

Teorema 1 (Teorema del Límite Central)**.** Si $\bar{X}$ es la media de una muestra aleatoria de tamaño n tomada de una población con media μ y varianza finita σ^2, entonces la forma límite de la distribución de

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

conforme $n \to \infty$, es la distribución normal estándar $N(z; 0, 1)$

Ejemplo 3. Una empresa de material eléctrico fabrica focos que tienen una duración aproximadamente de forma normal, con $\mu = 800$ y con una desviación estándar de 40 horas. Encuentra la probabilidad de que una muestra aleatoria de 16 focos tenga una vida promedio de menos de 780 horas

```
> Z <- function(xbar, mu, sigma, n) (xbar-mu)/(sigma/ sqrt(n))
> 100*pnorm(Z(780, 800, 40, 16))
[1] 2.275013
```

4.2. Estimadores

Definición 5. Una *estimación puntual* de un parámero poblacional θ es el valor numérico particular $\hat{\theta}$ de un estadístico $\hat{\Theta}$

Ejemplo 4. Considera una v.a. $X \sim N(\mu, \sigma^2)$, con μ desconocida. Entonces $\bar{x}$ es un estimador puntual de μ, i.e., $\hat{\mu} = \bar{X}$. Si tenemos la muestra dada por $(25, 30, 29, 31)$, entonces $\bar{x} = 28{,}75$

```
> X <- c(25, 30, 29, 31)
> mean(X)
[1] 28.75
```

Con frecuencia es necesario estimar, con sus respectivos estimadores:

1. *Para μ, la estimación es $\hat{\mu} = \bar{x}$, la media muestral*
2. $\hat{\sigma}^2 = s^2$
3. *Para p, $\hat{p} = x/n$*
4. $\hat{\mu}_1 - \hat{\mu}_2 = \bar{x}_1 - \bar{x}_2$
5. $\hat{p}_1 - \hat{p}_2$

4.2.1. Propiedades de los estimadores

Definición 6. El estimador puntual $\hat{\Theta}$ es un *estimador insesgado* del parámetro θ si

$$\mathbb{E}(\hat{\Theta}) = \theta.$$

Si el estimador no es insesgado, entonces, a la diferencia $\mathbb{E}(\hat{\Theta}) - \theta$ se le llama el *sesgo* del estimador $\hat{\Theta}$

Ejercicio 30. Muestra que S^2 es un estimador insesgado del parámetro σ^2

Si $\hat{\Theta}_1$ y $\hat{\Theta}_2$ son dos estimadores insesgados del mismo parámetro poblacional θ, y si $\sigma^2_{\hat{\theta}_1} < \sigma^2_{\hat{\theta}_2}$, entonces se dice que $\hat{\Theta}_1$ es un estimador más eficaz *que $\hat{\Theta}_2$*

Definición 7. Si consideramos todos los posibles estimadores insesgados de algún parámero θ, el de menor varianza se llama *estimador más eficaz* de θ

Definición 8. El *error estándar* de un estimador $\hat{\Theta}$ es su desviación estándar, dada por $\sigma_{\hat{\Theta}} = \sqrt{V(\hat{\Theta})}$. Si el error estándar incluye parámetros desconocidos que pueden estimarse, entonces la sustitución de dichos valores en $\sigma_{\hat{\Theta}}$ produce un *error estándar estimado*, denotado por $\hat{\sigma}_{\hat{\Theta}}$

En particular, tenemos que para un muestreo de una distribución normal, $\bar{X} \sim N(\mu, \sigma^2/n)$, por lo que el error estándar de $\bar{X}$ es

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

Si σ es desconocida, entonces

$$\hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}}.$$

Ejemplo 5. Considera lo siguiente

```
library(DAAG)
data(pair65)
dif <- pair65$heated - pair65$ambient
mean(dif)
sd(dif)
SEM <- function(s, n) s/sqrt(n)
SEM(sd(dif), length(dif))
```

Nota 1. Para estimaciones puntuales de parámetros, uno de los mejores métodos es el de máxima verosimilitud. Sin embargo, hay ocasiones en las que es preferible encontrar un intervalo dentro del cual se espera encontrar el parámetro, como veremos a continuación.

4.2.2. Estimación por intervalo

La estimación por intervalo de un parámetro poblacional θ es un intervalo de la forma $\hat{\theta}_L < \theta < \hat{\theta}_U$, donde $\hat{\theta}_L$ y $\hat{\theta}_U$ dependen del valor del estadístico $\hat{\Theta}$ para una muestra específica, y también de la distribución de muestreo de $\hat{\Theta}$.

El objetivo es encontrar las v.a. $\hat{\Theta}_L$ y $\hat{\Theta}_U$ tales que

$$\mathbb{P}(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha,$$

para $0 < \alpha < 1$, con lo cual tenemos una probabilidad de $1 - \alpha$ de seleccionar una v.a. que produzca un intervalo que contenga θ.

4.2.3. Estimación de la media

Sabemos que $\bar{X} \sim N(\mu, \sigma^2/n)$. Entonces

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

donde $Z = \frac{\bar{X}-\mu}{\sigma_{\bar{X}}}$. Entonces

$$\mathbb{P}(\bar{X} - \sigma_{\bar{X}} z_{\alpha/2} < \mu < \bar{X} + \sigma_{\bar{X}} z_{\alpha/2}) = 1 - \alpha,$$

de donde se obtiene el intervalo de confianza de $(1-\alpha)100\,\%$, dado por $\bar{X} - \sigma_{\bar{X}} z_{\alpha/2} < \mu < \bar{X} + \sigma_{\bar{X}} z_{\alpha/2}$

Ejemplo 6. Se encuentra que la concentración promedio de zinc que se obtiene a partir de una muestra de mediciones de zinc en 36 sitios diferentes es 2.6 gramos por mililitro. Encuentra los intervalos de confianza de 90 %, 95 % y 99 % para la concentración media de zinc en el río. Supón que $\sigma = 0{,}3$

```
simple.z.test = function(xbar, sigma, conf.level, n){
alfa = 1 - conf.level
Z = qnorm(1-alfa/2)
SE = sigma/sqrt(n)
xbar + c(-Z*SE,Z*SE)
}
> simple.z.test(2.6, 0.3, .95, 36)
[1] 2.502002 2.697998
```

Ejercicio 31. Una empresa de material eléctrico fabrica focos que tienen una duración aproximadamente de forma normal, con una desviación estándar de 40 horas. Si una muestra de 30 focos tiene una duración promedio de 780 horas, encuentra un intervalo de confianza de 95 % para la media de la población de todos los focos que produce esta empresa.

Podemos calcular el tamaño necesario de una muestra para asegurarnos de que el error al estimar μ sea menor que una cantidad específica ϵ, dado que $\epsilon = z_{\alpha/2}\ \sigma/\sqrt{n}$

Ejemplo 7. De qué tamaño se necesita una muestra del Ejercicio 6 si deseamos tener 95 % de confianza de que nuestra estimación de μ difiera por menos de 0.05?

```
alfa = 0.05
sigma = 0.3
Z = qnorm(1-alfa/2)
error = 0.05
n = ((Z*sigma)/error)^2
> n
[1] 138.2925
```

Qué hacer en el caso en que la desviación estándar poblacional sea desconocida? Si $\bar{x}$ y s son la media y la desviación estándar de una muestra aleatoria de una población con varianza σ^2 desconocida, un intervalo de confianza de $(1-\alpha)100\,\%$ para μ es

$$\bar{x} - \hat{\sigma}_{\bar{X}} t_{\alpha/2} < \mu < \bar{x} + \hat{\sigma}_{\bar{X}} t_{\alpha/2},$$

donde $\hat{\sigma}_{\bar{X}} = s/\sqrt{n}$ y $t_{\alpha/2}$ es el valor t con $\nu = n-1$ grados de libertad que deja un área de $\alpha/2$ a la derecha de la gráfica de distribución t-Student.

Ejemplo 8. Una muestra aleatoria de 10 barras de chocolate energético de cierta marca tiene, en promedio, 230 calorías con una desviación estándar de 15 calorías. Construye un intervalo de confianza de 99 % para el contenido medio de calorías real de esta marca de barras de chocolate energético. Supón que la distribución de las calorías es aproximadamente normal.

```
simple.t.test = function(xbar, s, conf.level, n){
alfa = 1 - conf.level
t = qt(1 - alfa / 2, n-1 )
SE = s/sqrt(n)
xbar + c(-t*SE,t*SE)
```

```
}
> simple.t.test(230, 15, 0.99,10)
[1] 214.5847 245.4153
```

4.2.4. Intervalos de predicción

Algunas veces, aparte de la media de la población, quizás el experimentador esté interesado en predecir los posibles valores de una observación futura. Entonces, para una distribución normal de mediciones con media desconocida μ y varianza conocida σ^2, un intervalo de predicción *de* $(1-\alpha)100\,\%$ *de una observación futura x_0 es*

$$\bar{x} - z_{\alpha/2}\sigma\sqrt{1+1/n} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1+1/n}$$

Ejemplo 9. A causa de la disminución de las tasas de interés, un Banco recibió muchas solicitudes para hipoteca. Una muestra reciente de 50 créditos hipotecarios resultó en un promedio de \$257,300. Supón que la desviación estándar de la población es de \$25,000. Si el siguiente cliente llamó para una solicitud de crédito hipotecario, encuentra un intervalo de predicción de 95 % para la cantidad del crédito de este cliente

Cuando la media μ y varianza poblacional σ^2 son desconocidas, se puede calcular un intervalo de predicción *de la siguiente manera*

$$\bar{x} - t_{\alpha/2}s\sqrt{1+1/n} < x_0 < \bar{x} + t_{\alpha/2}s\sqrt{1+1/n}$$

Ejercicio 32. Un inspector de alimentos midió aleatoriamente 30 paquetes de carne de res 95 % sin grasa. La muestra resultó en una media de 96.2 % con la desviación estándar muestral de 0.08 %. Encuentra un intervalo de predicción de 99 % para un paquete nuevo. Suponga normalidad.

Nota 2. Una metodología para la detección de valores extremos implica la regla de que una observación es un valor extremo si cae fuera del intervalo de predicción calculado sin incluir la observación cuestionable en la muestra

4.2.5. Límites de tolerancia

Definición 9. Para una distribución normal de mediciones con media μ y desviación estándar σ, ambas desconocidas, los *límites de tolerancia* están dados por

$$\bar{x} \pm ks,$$

donde k se determina de manera que se pueda asegurar con una confianza de $(1-\gamma)100\,\%$ que los límites dados contienen al menos la proporción $1-\alpha$ de las mediciones.

Ejemplo 10. Una máquina produce piezas de metal que tienen forma cilíndica. Se toma una muestra de tales piezas y se encuentra que los diámetros son 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01, 1.03 centímetros. Encuentra los límites de tolerancia de 99 % que contendrán 95 % de las piezas de metal que produce la máquina. Suponga una distribución aproximadamente normal.

```
library("tolerance")
mq <- c(1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01, 1.03)
normtol.int(x = mq, alpha = 0.05, P = 0.99, side = 2)
alpha   P    x.bar 2-sided.lower 2-sided.upper
1  0.05 0.99 1.005556   0.8905689     1.120542
```

4.3. Pruebas de Hipótesis

Definición 10. Una *hipótesis estadística es una aseveración o conjetura con respecto a una o más poblaciones*

Un procedimiento debe hacerse con la noción de la probabilidad de una conclusión errónea*. El rechazo de una hipótesis simplemente implica que la evidencia de la muestra la refuta. Por otro lado, el rechazo significa que hay una pequeña probabilidad de obtener la información muestra observada cuando, de hecho, la hipótesis es verdadera.*

La estructura de la prueba de hipóte<U+00E9>is se formulará usando el tt́ermino hipótesis nula*, el cual se refiere a cualquier hipótesis que deseamos probar y se denota H_0. El rechazo de H_0 conduce a la aceptación de una* hipótesis alternativa*, denotada por H_a. Esta representa, por lo general, la pregunta que debe responderse o la teoría que debe probarse.*

4.3.1. Errores de prueba

Existen dos tipos de errores al hacer una prueba de hipótesis estadística

Definición 11. El rechazo de la hipótesis nula cuando es verdadera se llama *error tipo I*. La probabilidad de cometer un error tipo I, también llamada *nivel de significancia*, se denota por α.

Definición 12. El rechazo de la hipótesis nula cuando es falsa se llama *error tipo II*. La probabilidad de cometer un error tipo II se denota por β.

4.3.2. Prueba con respecto a una sola media, varianza conocida

Considera un experimento con $X_1, X_2, \ldots, X_n$, que representan una muestra aleatoria de una distribución $N(\mu, \sigma^2)$. Consideremos la hipótesis

$$\begin{aligned} H_0: &\quad \mu = \mu_0 \\ H_a: &\quad \mu \neq \mu_0 \end{aligned}$$

La prueba es equivalente a calcular un intervalo de confianza de $(1-\alpha)100\,\%$ sobre μ y rechazar H_0, si μ_0 no está dentro del intervalo de confianza, de donde tenemos

$$\bar{x} - \sigma_{\bar{X}} z_{\alpha/2} < \mu_0 < \bar{x} + \sigma_{\bar{X}} z_{\alpha/2}.$$

Las formas usuales de realizar las pruebas son mediante el estadístico z y los p- values, como sigue: Considera el estadístico de prueba z_0 dado por

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2} \quad o \quad z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}.$$

Si $-z_{\alpha/2} < z_0 < z_{\alpha/2}$, no se rechaza H_0.

Ejemplo 11. Un fabricante de equipo deportivo desarrlló un nuevo sedal para pesca sintético que afirma que tiene una resistencia media a la rotura de 8 kilogramos con σ =0.5 kg. Prueba la hipótesis de que $\mu = 8$ kg contra la alternativa de que $\mu \neq 8$, si se prueba una muestra aleatoria de 50 sedales y se encuentra que tiene resistencia media a la rotura de 7.8 kg. Utiliza un nivel de significancia de 0.01

```
simple.z.test.two.tail = function(xbar, mu0, sigma, conf.level, n){
alfa = 1 - conf.level
z0 = (xbar - mu0)/(sigma / sqrt(n))
zalpha = qnorm(1-alfa/2)
c(-zalpha, z0, zalpha)
}
simple.z.test.two.tail(7.8, 8, 0.5, .99, 50)
```

Definición 13. El *valor P* o *p-value* es el nivel de significancia más bajo que llevaría al rechazo de la hipótesis nula H_0 con los datos dados.

Ejemplo 12. Para el ejercicio 11, tenemos

```
> 2*pnorm(-z_0)
[1] 0.0046548
```

que nos permite rechazar H_0 en un nivel de significancia de 0.01

Ejemplo 13. Una muestra aleatoria de 100 muertes registradas en EUA el año pasado mostró una vida promedio de 71.8 años. Suponiendo $\sigma^2 = 8.9$ años, esto parece indicar que la vida media actual es mayor que 70 años?

Ejercicio 33. Se comparan las resistencia de dos clases de hilo. Cincuenta piezas de cada clase se prueban bajo condiciones similares. La marca A tiene una resistencia a la tensión promedio de 78.3 kg con una desviación estándar de 5.6 kg; en tanto que la marca B tiene una resistencia a la tensión promedio de 87.2 kg con una desviación estándar de 6.3 kg. Construye un intervalo de confianza de 95 % para la diferencia de las medias poblacionales.

```
dif.z.test = function(x1, x2, sigma1, sigma2, conf.level, n1, n2){
alfa = 1 - conf.level
Z = qnorm(1 - alfa / 2)
SE = sqrt(sigma1^2/n1+sigma2^2/n2)
(x1-x2) + c(-Z*SE,Z*SE)
}
dif.z.test(87.2,78.3, 6.3, 5.6, 0.95, 50, 50)
```

Los siguientes ejercicios están basados en el libro: Data Analysis and Graphics Using R. An Example-Bases Approach*, de John Maindonald y W. John Braun, Cabridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Third Edition, 2003*

Ejercicio 34. Contesta 5 de los 21 ejercicios del capítulo 1 *A brief introduction to* R

Ejercicio 35. Contesta 5 de los 14 ejercicios del capítulo 2 *Styles of data analysis*

Ejercicio 36. Contesta 5 de los 13 ejercicios del capítulo 3 *Statistical models*

5. Análisis Multivariado

La principal dificultad o limitación de las investigaciones en el ámbito socioeconómico, nacen de la complejidad inherente a los sistemas sociales, que es el resultado del gran nḿuero de factores de influencia en el comportamiento de estos sistemas, así como de la gran cantidad de interpelaciones entre estos factores, a través de los cuales se explica el comportamiento o funcionamiento de los mismos. La medición de los fenómenos sociales o económisos no es una tarea simple, precisamente en función de la complejidad mencionada. Ello obliga a la utilización de técnicas o mt́odos estadísticos capaces de sintetizar estos fenḿenos, destacando los factores verdaderamente relevantes o significativos, estimando su capacidad de influencia, midiénrola, en definitiva, de la manera más aproximada posible, que facilite su interpretación y suministre racionalidad a los procesos que de tal interpretación deriven.

Definición 14 (Análisis multivariado)**.** Es el conjunto de técnicas estadísticas que de forma simultánea miden, explican y predicen todas las relaciones existentes entre los elementos que conforman una tabla de datos, proporcionando un resultado que deber ser interpretado minuciosamente por el analista.

Se ha dicho que aunque es fácil reunir datos, es mucho más difícil reunir información. Los métodos multivariados pueden ayudar a determinar si existe información en los datos y también pueden ayudar a resumir esa información.

Ejemplos de datos multivariados

1. *Un agricultor que cultiva trigo podría interesarse en otras variedades, así como da/ nos por insectos y sequías.*

2. *Relaciones entre comportamientos humanos*

A menudo, el objetivo primario de los análisis multivariados es resumir grandes cantidades de datos por medio de relativamente pocos parámetros. El tema subyacente de muchas técnicas multivariadas es la simplificación.

Distinción entre los métodos

1. Técnicas dirigidas por las variables*: se enfocan primordialmente en las relaciones que podrían existir entre las variables respuesta que se están midiendo.*

2. Técnicas dirigidas por los individuos*: se interesan principalmente en las relaciones que podrían existir entre las unidades experimentales o individuos que se están midiendo.*

5.1. Técnicas del análisis multivariable

Se pueden catalogar en dos:

1. *De análisis de dependencia (o de relación): técnicas aplicables cuando una o varias variables dependientes van a ser explicadas por un conjunto de variables independientes que actúan como predictoras.*

2. *De análisis de interdependencia (o de comparación): técnicas que otorgan la misma consideración a todas las variables objeto de estudio, sin distinguir entre dependientes e independientes, y que tienen como fin descubrir las interpelaciones y, en definitiva, la estructura subyacente en ellas. Son, por tanto, técnicas de clasificación.*

5.2. Descripción de técnicas principales

5.2.1. Técnicas de análisis de dependencia

Análisis univariable y multivariable de la varianza y de la covarianza *El* análisis de la varianza *o ANOVA es una técnica caracterizada por el empleo de una variable dependiente de carácter métrico y varias independientes no métricas que actúan como predictoras, i.e.,*

$$y = f(x_1, x_2, \ldots, x_m)$$

donde y es métrica y x_i no métricas.

Se utiliza para analizar con la variable dependiente si diversas muestras proceden de poblaciones con igual media. En función de los valores presentados por las variables independientes se distinguirán una serie de grupos. El ANOVA mide la significancia estadística de las diferencias entre las medias que la variable dependiente presenta en los distintos grupos.

El análisis multivariable de la varianza *(MANOVA) es una extensión del ANOVA, que se aplica a una combinación de variables dependientes relacionadas entre si, i.e.,*

$$g(y_1, y_2, \ldots, y_n) = f(x_1, x_2, \ldots, x_m)$$

donde y_i métricas, x_j no métricas.

Análisis discriminante múltiple *Es una técnica de clasificación, ya que permite agrupar a los elementos de una muestra en dos o más categorías diferentes.*

Regresión lineal múltiple *Permite analizar la relación que existe entre una variable dependiente métrica y varias variables independientes también métricas, i.e.,*

$$y = f(x_1, x_2, \ldots, x_m)$$

donde y métrica y x_i métricas.

Un análisis de regresión pretende determinar la combinación lineal de variables independientes cuyos cambios son los mejores predictores de los cambios experimentados por la variable dependiente.

Análisis conjunto *Es una técnica que se emplea para entender cómo conforman los individuos sus preferencias hacia objetos.*

Segmentación jerárquica *Tiene por objeto distinguir grupos de elemntos homogéneos en una población.*

Ecuaciones estructurales *Permite analizar varias relaciones de dependencia que se presentan simultáneamente.*

5.2.2. Técnicas de análisis de interdependencia

Análisis Factorial *Parecido al PCA.*

Análisis de Componentes Principales *El análiso de componentes principales (PCA) trata de explicar la estructura de varianza-covarianza de un conjunto de variables a través de unas cuantas combinaciones lineales de estas variables. Sus objetivos generales son: reducción de datos e interpretación. PCA a menudo es más efectivo para resumir la variablidad en un conjunto de variables cuando estas están altamente correlacionadas. Además, PCA es normalmente un paso intermedio en el análisis de datos dado que las nuevas variables creadas (predictores) pueden ser usadas en un análisis subsecuente como regresión multivariada o análisis de clusters. Es importante destacar que las nuevas variables producidas estarán no correlacionadas.*

Lo que se espera del PCA es que los primeros pocos componentes explicarán una proporción substancial de la variación en las variables originales y puedan, por consiguiente, ser usados para proveer un resumen conveniente de menor dimensión de estas variables que pueda ser útil en sí mismas o como entrada para otros análisis.

PCA es usado para determinar la estructura de un conjunto de datos multivariados. Específicamente, los propósitos de PCA son

1. *determinar la dimensionalidad* efectiva *del espacio de variables*
2. *encontrar combinaciones lineales de las variables originales las cuales dan cuenta de la mayoría de la variación en el sistema multivariado*
3. *visualizar las relaciones entre las observaciones y las variables*
4. *determinar variables derivadas las cuales contienen la información multivariada esencial como una primera etapa para otros análisis*
5. *identificar outliers multivariados*
6. *determinar si hay una estructura poblacional presente*
7. *examinar la relación entre las principales variables y las variables explicativas, definidas interna o externamente*
8. *explicar la variación residual después de ajustar las variables condicionantes.*

PCA no es una técnica de modelación, pero es a menudo usada en el proceso de modelación para aprender sobre la estructura interna de un conjunto de datos.

Análisis de correspondencias *Es una herramienta de análisis de la interdependencia que, a semejanza del análisis factorial, busca la reducción de datos a un número pequeño de dimensiones; pero, a diferencia de aquel, permite trabajar con variables no métricas y estudiar relaciones no lineales.*

Análisis de cluster *Técnica cuyo fin es clasificar objetos en función de ciertas características; formar grupos con ellos de modo que las diferencias entre los contenidos dentro de un grupo determinado sean mínimas y las existentes respecto a los objetos de los restantes grupos, máximas.*

Escalamiento multidimensional *Tiene como fin elaborar una representación gráfica que permita conocer la imagen que los individuos se crean de un conjunto de objetos por posicionamiento de cada uno en relación a los demás.*