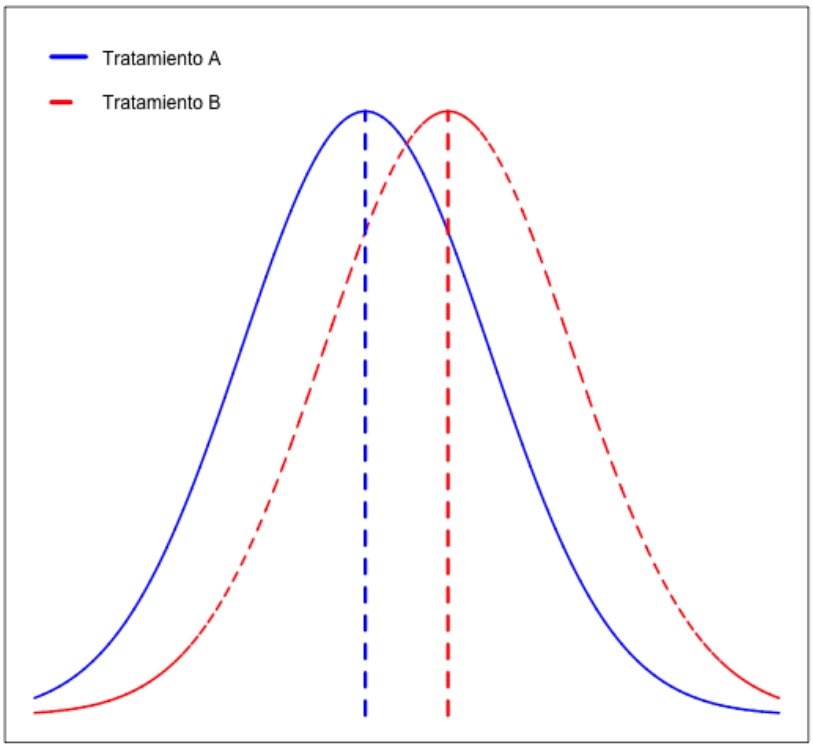
METODOLOGÍA ESTADÍSTICA

P. Saavedra



Índice general

Capítulo 1

Bases Probabilísticas

1.1. Fenómenos aleatorios

La mayor parte de los fenómenos que se observan en los campos de la sanidad y bromatología son de naturaleza aleatoria. Así por ejemplo, el efecto que puede producir una intervención terapéutica en un ser vivo a menudo es impredecible. Ello significa que las predicciones de tales fenómenos sólo podrían hacerse en términos de probabilidades.

En el contexto de la observación de un *fenómeno aleatorio*, un *suceso* es *cualquier conjunto de posibilidades que pueda presentarse.*

Ejemplo 1.1. Supóngase que al aplicar un cierto tratamiento terapéutico a un paciente, los posibles resultados que pueden darse se resumen en que el *paciente responda o no favorablemente al tratamiento*. En este caso, '*responder favorablemente'* es un suceso. Si el tratamiento se aplica a diez pacientes un suceso sería *'a lo sumo seis pacientes responden al tratamiento*'. Nótese que este suceso está formado por siete resultados elementales.

Para cualquier suceso A, su contrario es otro suceso que se expresa por A^C y que consiste en que no ocurra A (que ocurran todas aquellas posibilidades que son contrarias al suceso A). En el ejemplo 1.1, el contrario del suceso *'a lo sumo seis pacientes responden al tratamiento*' es '*al menos siete pacientes responden al tratamiento*'.

Para dos sucesos A y B consideraremos las siguientes operaciones:

- *Unión de sucesos.* Por $A \cup B$ representamos un suceso que consiste en que *ocura al menos uno de los dos sucesos* (sin perjuicio de que ocurran ambos a la vez).

- *Intersección de sucesos.* Por $A \cap B$ se representa al suceso que consiste en que *ocurran A y B simultáneamente.*

Decimos finalmente que dos sucesos A y B son incompatibles cuando es imposible que ocurran simultáneamente; esto es: si $A \cap B = \emptyset$, donde el símbolo $\emptyset$ designa al *suceso imposible.* Por último, el suceso seguro es el que contiene todos los posibles resultados y los designamos por Ω.

1.2. Concepto de probabilidad

1.2.1. Nociones básicas de probabilidad

La probabilidad de un suceso A es una *medida* $\Pr(A)$ de la verosimilitud del suceso o de la posibilidad de que éste ocurra. Normalmente se expresa en una escala comprendida entre 0 y 1, aunque puede expresarse también en escala porcentual (0-100). Cualquier medida de probabilidad debe satisfacer los siguientes axiomas:

1. $\Pr(\Omega) = 1$ *(el suceso seguro tiene la máxima probabilidad)*
2. Si $A \cap B = \emptyset$, entonces $\Pr(A \cup B) = \Pr(A) + \Pr(B)$

De los axiomas anteriores se deducen las siguientes propiedades:

1. $\Pr(\emptyset)$ *(el suceso imposible tiene la probabilidad mínima)*
2. Para cualquier suceso A, $\Pr\left(A^C\right) = 1 - \Pr(A)$
3. Para los sucesos A y B, se satisface que: $\Pr(B) = \Pr(A \cap B) + \Pr\left(A^C \cap B\right)$

Esta última propiedad se puede expresar en los siguiente términos: un suceso B puede ocurrir conjuntamente con otro suceso A ó con su contrario (A^C)

En determinadas circunstancias, la observación de un suceso A puede tener un valor predictivo para la observación de otro suceso B. Por ejemplo, la presencia de obesidad en un sujeto (A) puede ser predictiva de hipertensión arterial (B). En tal caso, la probabilidad de que ocurra B dependería que que A ocurriera o no. Esta idea conduce al concepto de probabilidad condicional, la cual se define por:

$$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

De la definición anterior se sigue la siguiente regla multiplicativa:

$$\Pr(A \cap B) = \Pr(A) \Pr(B \mid A)$$

Ejemplo 1.2. Un inspector desea estimar el número de restaurantes que incumplen una cierta normativa legal. Del total de los N restaurantes, x incumplen tal normativa. Supóngase ahora que pretende obtener la probabilidad de que al seleccionar dos restaurants al azar, ambos incumplan la referida normativa. Para ello deben definirse los siguientes sucesos: $F_1 =$ *'el primer restaurante elegido incumple la normativa'* y $F_2 =$ *'el segundo restaurante incumple la normativa'*. Se tiene entonces:

$$\Pr(F_1 \cap F_2) = \Pr(F_1) \Pr(F_2 \mid F_1) = \frac{x}{N} \times \frac{x-1}{N-1}$$

1.2.2. Independencia de sucesos

El hecho de que $\Pr(B \mid A) = \Pr(B)$ supone que la posibilidad de que ocurra el suceso B no está afectada por el hecho de que A ocurra o deje de ocurrir. En este caso se dice que el suceso B *es independiente de* A. Es interesante comprobar que cuando un suceso B es independiente de otro suceso A, necesariamente A lo es también de B. En efecto:

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(A) \Pr(B \mid A)}{\Pr(B)} = \frac{\Pr(A) \Pr(B)}{\Pr(B)} = \Pr(A)$$

Al ser $P(A \mid B) = \Pr(A)$, el suceso A es independiente de B por definición. De acuerdo con este resultado, podemos enunciar entonces la siguiente propiedad: *los sucesos A y B son independientes si y sólamente si:*

$$\Pr(A \cap B) = \Pr(A) \Pr(B)$$

Nótese que si un suceso B es independiente de otro suceso A, también lo es de su contrario. Para probar esta afirmación, supondremos que B es independiente de A y evaluamos la probabilidad condicional $\Pr\left(B \mid A^C\right)$.

$$\Pr\left(B \mid A^C\right) = \frac{\Pr\left(A^C \cap B\right)}{\Pr\left(A^C\right)} = \frac{\Pr(B) - \Pr(A \cap B)}{\Pr\left(A^C\right)} = \frac{\Pr(B) - \Pr(A) \Pr(B)}{\Pr\left(A^C\right)} =$$

$$\frac{\Pr(B)(1-\Pr(A))}{\Pr(A^C)} = \frac{\Pr(B)\Pr\left(A^C\right)}{\Pr(A^C)} = \Pr(B)$$

1.2.3. Riesgo relativo

El *riesgo relativo* (RR) o la *razón de tasas* es una medida de asociación entre dos sucesos. Más concretamente, supongamos que nos planteamos si un suceso A se asocia con otro B (por ejemplo, ¿se asocia la obesidad con la hipertensión arterial?). La cuestión se podría formular en otros términos, a saber: ¿es más probable que ocurra B presencia de A que en la de su contrario A^C?. Esto último lleva a comparar las probabilidades condicionales $\Pr(B \mid A)$ con $\Pr\left(B \mid A^C\right)$, lo cual conduce a la definición de riesgo relativo (RR) en los siguientes términos:

$$RR = \frac{\Pr(B \mid A)}{\Pr(B \mid A^C)}$$

Nótese que si el suceso B es independiente de A (y por tanto, independiente de su contrario), el valor del riesgo relativo es la unidad y recíprocamente, si $RR = 1$, los sucesos son independientes. El lector debería examinar esta cuestión detenidamente.

En aquellos casos en los que $RR > 1$ hay asociación positiva (en los estudios epidemiológicos suele decirse que A es factor de riesgo de B), mientras que cuando $RR < 1$ la asociación se dice inversa (a menudo, A es un factor protector de B).

Ejemplo 1.3. En la tabla 1.1 se resumen las tasas de hipertensión arterial (HTA) en los subgrupos de una cierta población caracterizados por la presencia o ausencia de obesidad ($IMC > 30Kg/m^2$).

	Obesidad	
	Si ($N = 327$)	No ($N = 693$)
Hipertensión arterial, OMS, n, (%)	170 (52.0)	150 (21.6)
Edad media, años	51.3	46.5

Tabla 1.1: Prevalencias de hipertensión arterial y edad media en los grupos determinados por la presencia o no de obesidad

De los datos anteriores puede estimarse que la probabilidad condicionada de que un obeso sea hipertenso es del 52.0 %, mientras que en los no obesos, esta probabilidad baja al 21.6 %. De esta forma, el riesgo relativo es:

$$RR = \frac{\Pr(HTA \mid OB)}{\Pr(HTA \mid OB^C)} \approx 2{,}40$$

En este punto cabe advertir que la asociación detectada no supone necesariamente la existencia de una relación de causalidad. Nótese que la edad media en el grupo de personas obesas supera en casi cinco años a la media en el de los no obesos. Al menos parcialmente, la mayor prevalencia de HTA en el grupo de obesos podría atribuirse al hecho de que éstos tienen mayor edad.

1.2.4. Odd-ratio

Una medida más propia de asociación entre dos sucesos A y B es la *odd-ratio* (OR), la cual se define por:

$$OR = \frac{\Pr(B \mid A)\Pr\left(B^C \mid A^C\right)}{\Pr(B \mid A^C)\Pr(B^C \mid A)}$$

Una propiedad atractiva de la OR que puede verificarse fácilmente es:

$$OR = \frac{\Pr(A \cap B)\Pr\left(A^C \cap B^C\right)}{\Pr(A^C \cap B)\Pr(A \cap B^C)}$$

De lo anterior se deduce que la OR puede expresarse también en la forma:

$$OR = \frac{\Pr(A \mid B)\Pr\left(A^C \mid B^C\right)}{\Pr(A^C \mid B)\Pr(A \mid B^C)}$$

Puede comprobarse también que la independiencia entre los sucesos A y B es equivalente a $OR = 1$.

La OR que corresponde a la asociación de hipertensión y obesidad del ejemplo 1.3 es:

$$OR = \frac{\Pr(HTA \mid OB)\Pr\left(HTA^C \mid OB^C\right)}{\Pr(HTA \mid OB^C)\Pr(HTA^C \mid OB)} = \frac{0{,}520 \times 0{,}784}{0{,}216 \times 0{,}480} \approx 3{,}92$$

Hacemos ahora un estudio comparativo del RR y la OR en el contexto de la evaluación de la asociación entre un factor de exposición F y una enfermedad E. Para tal fin consideramos dos escenarios (diseños) alternativos:

- *Estudio de cohortes*. Considérese una población de *sujetos sanos* en un instante determinado clasificada en dos cohortes determinadas por la exposición o no a un

cierto factor F. Supóngase ahora que a lo largo de un periodo de seguimiento se observan las incidencias de una enfermedad E en cada una de las dos cohortes (F y F^C). El riesgo relativo definido por $RR = \Pr(E \mid F) / \Pr\left(E \mid F^C\right)$ tiene la interpretación simple vista anteriormente. La odd-ratio se puede expresar en la forma:

$$OR = \frac{\Pr(E \mid F) / \Pr\left(E^C \mid F\right)}{\Pr(E \mid F^C) / \Pr(E^C \mid F^C)}$$

El numerador expresa *cuanto más probable es enfermar que no enfermar en los sujetos con la exposición* F mientras que el denominador representa la misma relación pero entre los sujetos no expuestos (F^C).

Nótese que todas las probabilidades que intervienen en ambas definiciones son estimables en este tipo de estudio y por tanto, los son el RR y la OR.

- *Estudio de caso-control.* Este es un escenario en el que inicialmente los sujetos de la población están clasificados en enfermos (E) y sanos (E^C) y se quiere investigar retrospectivamente los posibles factores que han podido causar la enfermedad. En este estudio no es posible estimar las probabilidades de E en los grupos F y F^C y por tanto, no es posible estimar el riesgo relativo. Sin embargo, la OR puede estimarse por:

$$OR = \frac{\Pr(F \mid E) / \Pr\left(F^C \mid E\right)}{\Pr(F \mid E^C) / \Pr(F^C \mid E^C)}$$

Ejemplo 1.4. Logroscino *et al* llevaron a efecto un estudio de caso-control para evaluar la asociación entre las dietas antioxidantes y la enfermedad de Parkinson. En el estudio incluyeron 110 pacientes parkinsonianos y 287 sujetos controles.

	Casos (E) $n = 110$	Controles (E^C) $n = 287$	OR
Factor F: Calorías $> 1353, \%$	66.4	41.8	2.75

Figura 1.1:

los datos significan que $\Pr(F \mid E) = 0{,}664$ y $\Pr\left(F \mid E^C\right) = 0{,}418$ (en realidad, estos valores son estimaciones de las referidas probabilidades). Se tiene entonces:

$$OR = \frac{0{,}664 \times 0{,}582}{0{,}418 \times 0{,}336} \approx 2{,}75$$

Nótese que de estos datos no es posible estimar las probabilidades $\Pr(E \mid F)$ y $\Pr\left(E \mid F^C\right)$, y de ahí, el riesgo relativo. Para que tales probabilidades pudieran determinarse se requeriría conocer la prevalencia de la enfermedad; esto es: $\Pr(E)$. En tal caso, mediante la fórmula de Bayes se obtiene:

$$\Pr(E \mid F) = \frac{\Pr(F \mid E)\Pr(E)}{\Pr(F \mid E)\Pr(E) + \Pr(F \mid E^C)\Pr(E^C)}$$

Análogamente se obtendría $\Pr\left(E \mid F^C\right)$.

1.3. Variables aleatorias y su distribución de probabilidad

1.3.1. Concepto de variable aleatoria

Una variable aleatoria es cualquier magnitud X cuyo valor depende del azar.

Ejemplo 1.5. Un inspector quiere estimar el número total de restaurantes de una población que incumplen una cierta normativa. Para ello selecciona una muestra aleatoria de tamaño determinado. El *número de restaurantes que encuentra en la muestra que incumplen la referida normativa* es una variable aleatoria.

.

Ejemplo 1.6. El nivel de triglicéridos (TG) de un sujeto seleccionado aleatoriamente de una población es una variable aleatoria.

.

Ejemplo 1.7. El tiempo que sobrevive un paciente después del diagnóstico de una cierta enfermedad es también una variable aleatoria.

1.3.2. Distribuciones de probabilidad

La distribución de probabilidad de una variable aleatoria es una medida cuya finalidad es predecir valores de la variable. Para su especificación distinguiremos entre *variables aleatorias discretas* y *variables aleatorias absolutamente continuas*.

1.3.2.1. Distribuciones discretas

Una variable aleatoria X es discreta si el conjunto de posibles valores que puede tomar es *finito* o *numerable*. En tal caso, la distribución de probabilidad se especifica mediante la función $\Pr(X = t)$.

Ejemplo 1.8. Supóngase que cuando se aplica un tratamiento a los elementos de una cierta población, los posibles sucesos que pueden presentarse son: *repuesta favorable* (F) y *respuesta desfavorable* (D). En tal escenario se define la *tasa de respuestas favorables* como la probabilidad de que ocurra el suceso F; esto es : $\pi = \Pr(F)$. Supongase ahora que el tratamiento se aplica consecutivamente a tres elementos de la población de estudio y se define $X = n^o$ *total de respuestas favorables*. Obviamente X es una variable aleatoria que satisface $X \in \{0, 1, 2, 3\}$ y de esta forma es discreta. Su distribución de probabilidad se resume en la siguiente tabla:

	DDD	$DDF \quad DFD \quad FDD$	$DFF \quad FDF \quad FFD$	FFF
X	0	1	2	3
$\Pr(X = t)$	$(1-\pi)^3$	$3\pi(1-\pi)^2$	$3\pi^2(1-\pi)$	π^3

1.3.2.2. Distribuciones absolutamente continuas

Una variable aleatoria X se dice que es absolutamente continua cuando su distribución de probabilidad puede especificarse por una función $f(t)$ no negativa y tal que:

$$\Pr(a \leq X \leq b) = \int_a^b f(t)\, dt$$

La expresión anterior indica que la probabilidad de que la variable aleatoria X tome un valor comprendido entre a y b es igual al área comprendida bajo la gráfica de la función de densidad entre los valores de abscisa a y b.

Ejemplo 1.9. En la figura que sigue se muestra la función de densidad de probabilidad de la variable aleatoria TG = nivel de triglicéridos de un sujeto seleccionado aleatoriamente de una cierta población.

Nótese que el área marcada bajo la gráfica de la función y sobre el intervalo $[80\,;\,120]$ es la probabilidad de que el nivel resultante de TG sea un valor perteneciente a ese intervalo; esto es:

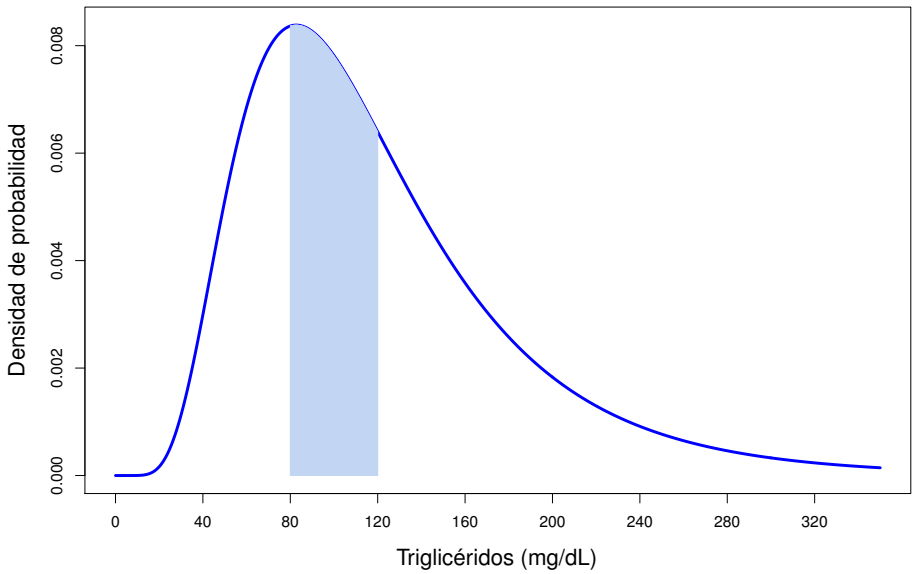


Figura 1.2: Densidad de probabilidad del nivel de TG de un sujeto seleccionado aleatoriamente de una cierta población

$$P\left(80 \leq TG \leq 120\right) = \int_{80}^{120} f\left(t\right) dt$$

1.3.3. Esperanza matemática

Si hicieramos una analogía entre distribución de masas y distribución de probabilidad, el concepto de esperanza de una variable aleatoria coincidiría con el de centro de gravedad de su distribución de probabilidad. Teniendo en cuenta la definición de centro de masas o de gravedad, definimos la esperanza de una variable aleatoria discreta X por:

$$E\left[X\right] = \sum_{t} t \cdot \Pr\left(X = t\right)$$

Para la distribución del ejemplo 1 queda:

$$E\left[X\right] = 3\pi\left(1-\pi\right)^2 + 6\pi^2\left(1-\pi\right) + 3\pi^3 = 3\pi$$

Para el caso $\pi = 0{,}7$, el valor esperado de la variable es $E\left[X\right] = 2{,}1$. Este resultado

es de fácil interpretación, a saber: cabe esperar que alrededor del 70 % del los tres sujetos que reciben el tratamiento presenten una respuesta favarable ($3 \times 0{,}7$).

Cuando la variable aleatoria X es continua y tiene función de densidad $f(t)$, su esperanza se define por:

$$E[X] = \int t \cdot f(t) \cdot dt$$

1.3.4. Desviación estándar

Para cualquier variable aleatoria X cuya esperanza es $E[X] = \mu$, la *desviación estándar* es una medida de dispersión de su distribución de probabilidad, la cual se define por:

$$\mathrm{sd}(X) = E\left[(X-\mu)^2\right]^{1/2}$$

En la figura se muestran simultáneamente tres funciones de densidad de probabilidad con esperanza común $\mu = 80$, pero diferentes desviaciones estándar, a saber: 3, 5 y 8.

1.3.5. Distribución binomial

Considérese un experimento aleatorio en el que un cierto evento tiene probabilidad π de ocurrir. Supóngase el experimento de repite n veces en las mismas condiciones siendo los sucesivos resultados independientes, y sea X el número de veces que ocurre el evento de estudio. En tales condiciones se dice que X es una variable aleatoria con distribución de probabilidad *binomial* de parámetros n y π ($X \cong b(n, \pi)$). En tal caso ocurre:

- $E[X] = n\pi$
- $\mathrm{sd}(X) = \sqrt{n\pi(1-\pi)}$

1.3.6. Distribución normal o gausiana

Una variable aleatoria continua X se dice que tiene distribución de probabilidad normal o gausiana con esperanza μ y desviación estándar σ ($X \cong N(\mu, \sigma)$) si su función de densidad tiene la forma:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Los parámetros μ y σ corresponden a la esperanza y desviación estándar respectivamente. Las tres funciones de densidad que aparecen superpuestas en la figura 1.2 corresponden a distribuciones normales con esperanza $\mu = 80$ y desviaciones estándar 3, 5 y 8. Nótese que la función de densidad es simétrica respecto de su esperanza (centro de gravedad).

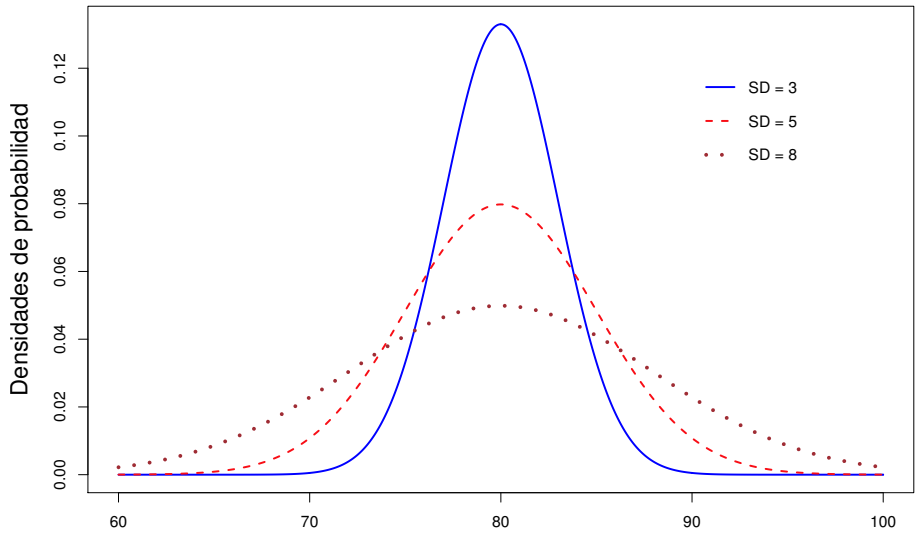


Figura 1.3: Densidades de probabilidad de a distribución normal de esperanza 80 y diferentes desviaciones estándar

Una propiedad que permite interpretar datos generados por una distribución $N(\mu, \sigma)$ es la siguiente: si $X \cong N(\mu, \sigma)$, entonces:

$$\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0{,}95$$

La distribución $N(0, 1)$ recibe el nombre de distribución normal estándar.

1.3.7. Distribución *ji*-cuadrado con un grado de libertad

Sea Z una variable aleatoria con distribución normal estándar ($Z \cong N(0,1)$). Por definición, la variable aleatoria $Y = Z^2$ tiene distribución de probabilidad *ji*-cuadrado con un grado de libertad ($Y \cong \chi^2(1)$). Su función de densidad de probabilidad es:

$$f(x) = \frac{\exp(-x/2)}{\sqrt{2\pi x}}$$

En la figura 1.3 se representan las funciones de densidad de las distribuciones normal estándar y *ji*-cuadrado con un grado de libertad.

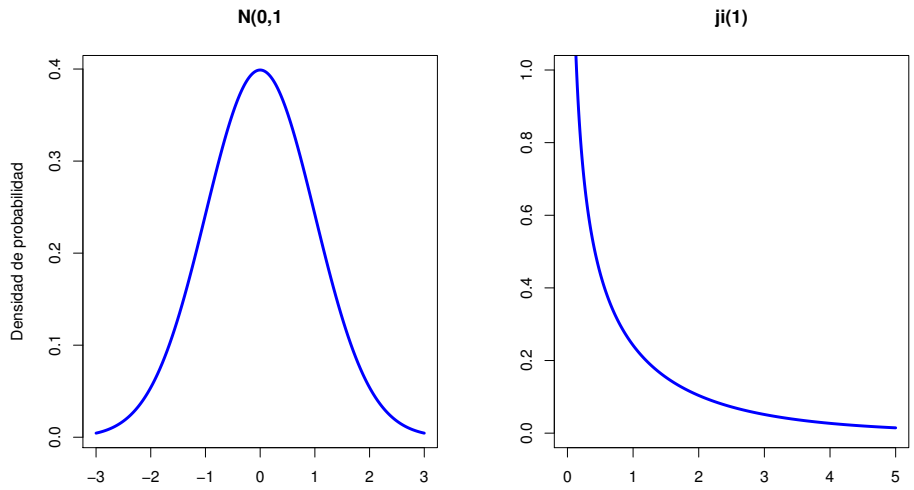


Figura 1.4: Densidades de probabilidad de las distribuciones normal estándar y ji-cuadrado

1.3.8. Cuantiles

Para cualquier variable aleatoria X, el cuantil p $(0 < p < 1)$ se define como la cantidad x_p tal que:

$$\Pr(X \leq x_p) = p$$

El cuantil p de la distribución $N(0,1)$ lo representaremos por z_p mientras que el de la distribución $\chi^2(1)$, por $\chi^2_p(1)$. Esto es:

- Si $Z \cong N(0,1)$, entonces $\Pr(Z \leq z_p) = 0$.

	$N(0,1)$	$\chi^2(1)$			$N(0,1)$	$\chi^2(1)$
p	z_p	$\chi^2_p(1)$		p	z_p	$\chi^2_p(1)$
0.800	0.8416	1.6424		0.805	0.8596	1.6794
0.810	0.8779	1.7176		0.815	0.8965	1.7570
0.820	0.9154	1.7976		0.825	0.9346	1.8396
0.830	0.9542	1.8829		0.835	0.9741	1.9278
0.840	0.9945	1.9742		0.845	1.0152	2.0223
0.850	1.0364	2.0723		0.855	1.0581	2.2141
0.860	1.0803	2.1780		0.865	1.1031	2.2340
0.870	1.1264	2.2925		0.875	1.1503	2.3535
0.880	1.1750	2.4173		0.885	1.2004	2.4841
0.890	1.2265	2.5542		0.895	1.2536	2.6279
0.900	1.2816	2.7055		0.905	1.3106	2.7875
0.910	1.3408	2.8744		0.915	1.3722	2.9666
0.920	1.4051	3.0649		0.925	1.4395	3.1701
0.930	1.4758	3.2830		0.935	1.5141	3.4050
0.940	1.5548	3.5374		0.945	1.5982	3.6821
0.950	1.6449	3.8415		0.955	1.6954	4.0186
0.960	1.7507	4.2179		0.965	1.8119	4.4452
0.970	1.8808	4.7093		0.975	1.9600	5.0239
0.980	2.0537	4.4119		0.985	2.1701	5.9165
0.990	2.3263	6.6349		0.995	2.5758	7.8794

Tabla 1.2: Cuantiles para las distribuciones normal estándar y ji-cuadrado con 1 grado de libertad

- Si $Y \cong \chi^2(1)$, entonces $\Pr\left(Y \leq \chi^2_p(1)\right) = p$.

En la tabla 1.2 se muestran los cuantiles de las distribuciones normal estándar y $\chi^2(1)$.

Capítulo 2

Bases de estadística inferencial

2.1. Estimación puntual

Haremos una aproximación heurística al concepto de estimador puntual a través de un estudio de simulación en el que se considera un parámetro desconocido π que representa la tasa de respuestas favorables de un cierto tratamiento. El estimador natural de este parámetro es la proporción de respuestas favorables $\hat{\pi}_n$ observadas en una muestra aleatoria de n pacientes que han recibido el tratamiento de estudio. Obviamente el valor de $\hat{\pi}_n$ depende de la muestra aleatoriamente seleccionada lo que significa que tiene naturaleza de *variable aleatoria*. Resumimos ahora el algoritmo de simulación en los siguientes pasos:

1. Supóngase que la tasa real de respuestas es $\pi = 0{,}70$. Ello significa que cada vez que se aplica el tratamiento a un paciente, la probabilidad de que éste presente una respuesta favorable es del 70 %.

2. Se aplica el tratamiento a una muestra aleatoria de n pacientes y se calcula la proporción de respuestas favorables $\hat{\pi}_n$. Se considerarán los tamaños $n = 20$, 100, 300 y 1000.

3. El paso anterior se repite 20,000 veces, lo que supone disponer de 20,000 observaciones de la variable aleatoria $\hat{\pi}_n$.

4. Representamos finalmente las 20,000 observaciones de $\hat{\pi}_n$ mediante un histograma para cada uno de los tamaños muestrales n que se han considerado. Los resultados se muestran en la figura 2.1.

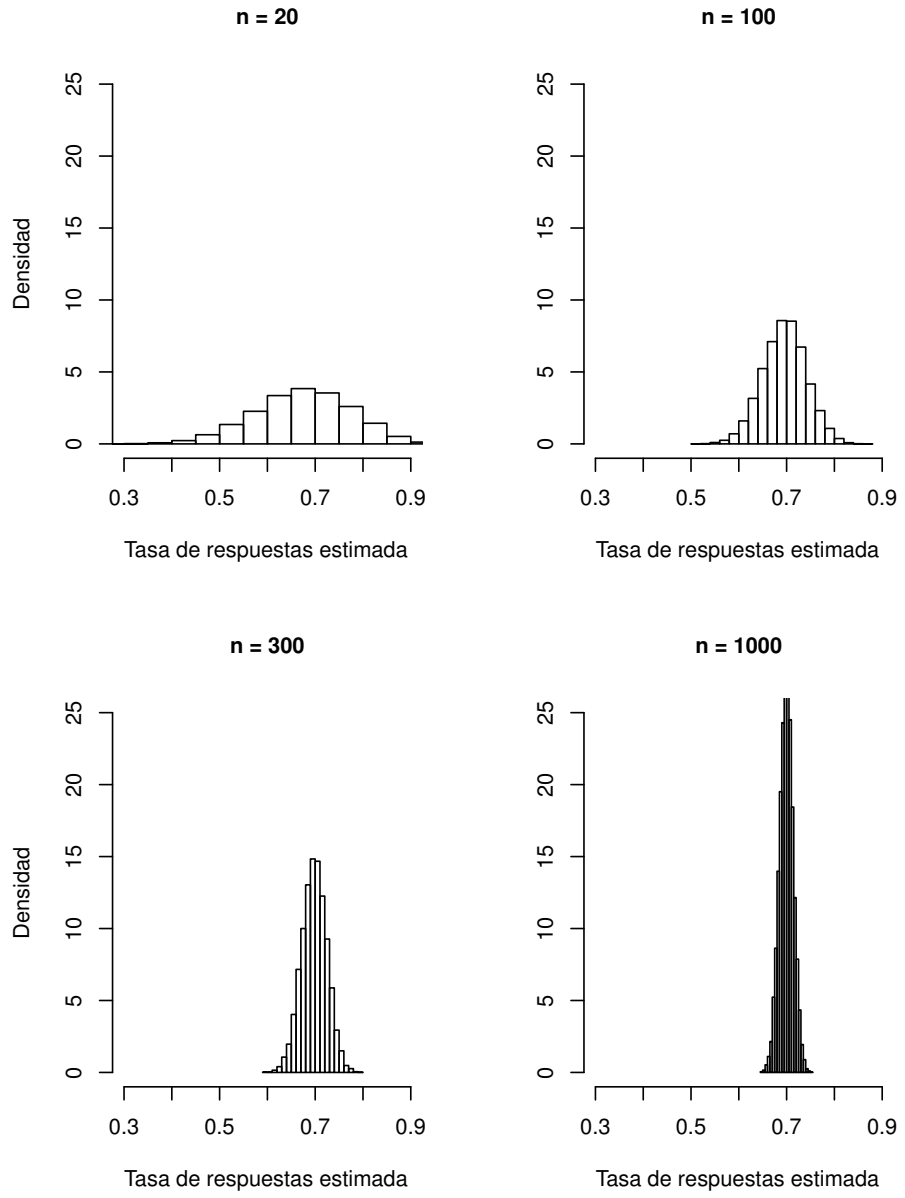


Figura 2.1: Estimador de la tasa de respuestas según tamaño muestral

2.1.1. Definición de estimador puntual

Sea θ un parámetro desconocido que representa un determinado aspecto de una población. Para su estimación puede utilizarse toda la información disponible del mismo. Una parte importante la aportan los datos obtenidos sobre una selección de elementos extraídos (normalmente al azar) de la referida población. De esta forma, si representa-

mos el conjunto de datos por $\mathcal{X}$, el estimador en general será una función de éstos que representaremos por $\hat{\theta}_n = \hat{\theta}_n(\mathcal{X})$.

La naturaleza aleatoria de los datos conlleva la naturaleza aleatoria del estimador. Tal como ya se ha señalado, su fiabilidad depende de la forma de su distribución de probabilidad. Dos propiedades deseables de cualquier estimador son: (i) que su valor esperado coincida o sea próximo al parámetro a estimar (que sea *exacto*); (ii) que tenga poca desviación estándar (que sea *preciso*). Estas ideas se concretan en los conceptos de *estimación centrada* y *error estándar*.

2.1.2. Estimación centrada

Se dice que un estimador $\hat{\theta}_n$ es *centrado* para un parámetro θ si $E\left[\hat{\theta}_n\right] = \theta$. Esto significa que todas las posibles estimaciones que se hagan de θ mediante $\hat{\theta}_n$ tienen como centro de gravedad al verdadero valor de θ.

Mediante una observación de la figura 2.1, el lector puede hacerse una idea de que el centro de gravedad de la distribución de probabilidad de $\hat{\pi}_n$ es $\pi = 0{,}70$.

2.1.3. Error estándar

Se llama *error estándar* de un estimador a su desviacion estándar; esto es: $\mathrm{sd}\left(\hat{\theta}_n\right)$. El error estándar por tanto mide la precisión del estimador en el sentido de que cuando es una cantidad próxima a cero, todas las posibles estimaciones son similares.

Ejemplo 2.1. Sea π la tasa de elementos de una población que tienen un cierto carácter. Para su estimación se seleccionan aleatoriamente n elementos de la población de estudio y se considera el estimador $\hat{\pi}_n =$ *proporción de elementos de la muestra que poseen el carácter del estudio*. Este estimador tiene las siguientes propiedades:

- El estimador es centrado; esto es: $E[\hat{\pi}_n] = \pi$
- $\mathrm{sd}(\hat{\pi}_n) = \sqrt{\pi(1-\pi)/n}$

Ejemplo 2.2. Supóngase que en una cierta población un marcador numérico X es tal que $E[X] = \mu$ y $\mathrm{var}(X) = \sigma^2$. En orden a estimar μ se observa el marcador en una muestra aleatoria se elementos de la población. Para las observaciones obtenidas $X_1, \ldots, X_n$ consideramos el estimador *media muestral* obtenido como: $\hat{\mu}_n = (1/n)\sum_{i=1}^{n} X_i$. El estimador satisface las siguientes propiedades:

- $E\left[\hat{\mu}_n\right] = \mu$
- $\mathrm{sd}\left(\hat{\mu}_n\right) = \sigma/\sqrt{n}$

Ejemplo 2.3. Sean π_E y π_C las tasas de elementos que poseen un cierto carácter en las poblaciones E y C respectivamente. Para estimar el riesgo relativo $\rho = \pi_E/\pi_C$ se seleccionan muestras aleatorias de las poblaciones E y C de tamaños n_E y n_C respectivamente y se obtiene el estimador $\hat{\rho} = \hat{\pi}_E/\hat{\pi}_C$, donde $\hat{\pi}_E$ y $\hat{\pi}_C$ son las estimaciones de π_E y π_C obtenidas en la forma indicada en el ejemplo 2.1. Las propiedades de este estimador, analizadas en la escala logarítmica, son las siguientes:

- $E\left[\log\hat{\rho}\right] \approx \log\rho$
- $\mathrm{sd}\left(\log\hat{\rho}\right) \approx \sqrt{\left(1-\pi_E\right)/\left(n_E\pi_E\right) + \left(1-\pi_C\right)/\left(n_C\pi_C\right)}$

2.1.4. Consistencia de un estimador y el problema de la determinación del tamaño muestral

Nótese que los estimadores descritos en la sección anterior son centrados. Además, el error estándar de $\hat{\pi}_n$ y $\hat{\mu}_n$ se aproximan a cero según aumenta el valor del tamaño muestral n. Para el caso del riesgo relativo, el error estándar del estimador se aproxima a cero si aumentan ambos tamaños muestrales; esto es: n_E, $n_C \to \infty$. En tales casos, la *distribución del estimador tiende a concentrarse sobre el verdadero valor del parámetro.* Cuando esto ocurre se dice que el estimador es *consistente.*

La primera consecuencia de la consistencia es que permite estimar el parámetro con toda la fiabilidad que se quiera siempre y cuando se aumente suficientemente el tamaño de la muestra. Para analizar esta cuestión, consideramos la siguiente ecuación de precisión:

$$\Pr\left(\theta - B \leq \hat{\theta}_n \leq \theta + B\right) = 0{,}95$$

La ecuación depende del tamaño muestral n y de la cantidad B llamada *cota de error.* Ésta expresa que, con un 95 % de probabilidad, la diferencia entre el verdadero valor del parámetro desconocido θ y cualquier estimación $\hat{\theta}_n$ es menor que la cota de error B prefijada. La solución de la ecuación conduce a obtener el tamaño muestral n que satisface tal condición

Analizamos ahora el caso en el que se desea estimar en una cierta población la tasa π de elementos que poseen un cierto carácter. Para tal fin se selecciona una muestra de n elementos y se estima π mediante la proporción muestral $\hat{\pi}_n$ de elementos que poseen el carácter de estudio. La ecuación que expresa la precisión es:

$$P\left(\pi - B \leq \hat{\pi}_n \leq \pi + B\right) = 0{,}95$$

La solución de la ecuación tiene la forma:

$$n = \frac{3{,}84}{B^2}\pi\left(1 - \pi\right)$$

Nótese que esta expresión del tamaño muestral depende del valor de π, el cual obviamente es desconocido. Para resolver este problema puede procederse de dos formas alternativas:

- Si el investigador *espera* que el valor de π sea una cantidad determinada, puede sustituirla en la expresión de n y obtener de esta forma un valor aproximado para el tamño muestral.

- Es fácil comprobar que cualquiera que sea el verdadero valor de π, se satisface siempre la desigualdad: $\pi\left(1 - \pi\right) \leq 1/4$. De esta forma, haciendo la oportuna sustitución se obtiene la siguiente cota para el valor de n:

$$n \leq \frac{3{,}84}{4B^2}$$

2.2. Intervalos de confianza

La estimación por intervalo de confianza de un parámetro θ consiste en determinar dos cantidades L y U tales que, con un grado de confianza determinado podamos afirmar que $L \leq \theta \leq U$. Los extremos L y U normalmente se basan en datos que contienen información acerca del verdadero valor de θ.

Más concretamente, supóngase que para estimar un parámetro θ se dispone de un conjunto de datos $\mathfrak{X}$ (obtenidos aleatoriamente de una población). Decimos que el intervalo de extremos $L = L\left(\mathfrak{X}\right)$ y $U = U\left(\mathfrak{X}\right)$ (sus valores quedan determinados por los datos $\mathfrak{X}$) es un intervalo de confianza al nivel $1 - \alpha$ para θ si:

$$P\left(L \leq \theta \leq U\right) = 1 - \alpha$$

Para estimadores $\hat{\theta}_n$ con distribución de probabilidad normal (o aproximadamente normal), los intervalos de la forma:

$$\left[\hat{\theta}_n - z_{1-\alpha/2} \cdot \mathrm{sd}\left(\hat{\theta}_n\right) \,;\, \hat{\theta}_n + z_{1-\alpha/2} \cdot \mathrm{sd}\left(\hat{\theta}_n\right)\right]$$

tienen probabilidad $1-\alpha$ de cubrir al verdadero valor del parámetro, y por tanto, son intervalos de confianza al nivel $1-\alpha$ para el parámetro θ.

Ejemplo 2.4. En un cierto estudio de diseño transversal realizado en Telde (Gran Canaria) cuyo tamaño muestral era de $n = 1030$ personas, se encontró que 128 padecían diabetes mellitus (DM2). De esta forma, la prevalencia estimada de DM2 fue $\hat{\pi}_n = 0{,}1243$ (12.43 %).

La expresión del error estándar es:

$$\mathrm{sd}\left(\hat{\pi}_n\right) = \sqrt{\frac{\pi\left(1-\pi\right)}{n}} \approx 0{,}01028$$

De esta forma, teniendo en cuenta que $\hat{\pi}_n$ tiene aproximadamente distribución normal y que el cuantil 0.975 de la distribución normal estándar es $z_{0{,}975} = 1{,}96$, un intervalo de confianza al 95 % es:

$$[0{,}1243 - 1{,}96 \times 0{,}01028\,;\, 0{,}1243 + 1{,}96 \times 0{,}01028] = [0{,}1041\,;\, 0{,}1444]$$

Hay por tanto una *confianza* del 95 % de que el verdadero valor de la prevalencia de DM2 esté comprendido entre el 10.41 % y 14.44 %.

En el ejemplo 1.3 se evaluó la asociación entre obesidad e hipertensión arterial (OMS) a través del riesgo relativo. Un intervalo de confianza al nivel $1-\alpha$ para el parámetro $\log \varrho$ es:

$$\left[\log\hat{\rho} - z_{1-\alpha/2} \cdot \mathrm{sd}\left(\log\hat{\rho}\right) \,;\, \log\hat{\rho} + z_{1-\alpha/2} \cdot \mathrm{sd}\left(\log\hat{\rho}\right)\right]$$

El valor del error estándar $\mathrm{sd}\left(\log\hat{\rho}\right)$ puede aproximarse por:

$$\mathrm{sd}\left(\log\hat{\rho}\right) \approx \sqrt{\frac{1-\hat{\pi}_E}{n_E\hat{\pi}_E} + \frac{1-\hat{\pi}_C}{n_C\hat{\pi}_C}}$$

Nótese que se han sustituido las tasas desconocidas π_E y π_C por sus estimaciones respectivas $\hat{\pi}_E$ y $\hat{\pi}_C$. De esta forma, el intervalo de confianza al 95 % ($\alpha = 0{,}05$) para $\log\rho$ es: [0,700 ; 1,052]. Ello supone que el intervalo de confianza para ρ es [2,015 ; 2,864].

Esto supone que con una confianza del 95 %, el riesgo relativo al menos es de 2.015, lo que indica que la probabilidad de hipertensión arterial en los sujetos obesos al menos duplica la de los controles.

2.3. Revisión del concepto de contraste de hipótesis

En esta sección se trata el problema de decidir sobre la verosimilitud de una hipótesis acerca de un parámetro de interés. Tal decisión se basará principalmente en un conjunto de datos que contengan información sobre el referido parámetro. Un *test de hipótesis* es una regla de decisión que *acepta la hipótesis planteada* si se evidencia que la hipótesis contraria es inconsistente con los datos. La fiabilidad de un test de hipótesis se evalúa a través de dos medidas llamadas *significación* y *potencia*.

2.3.1. Aproximación al problema de contraste de hipótesis

Supóngase que el tratamiento estándar para una enfermedad tiene una tasa de respuestas favorables del 50 % y que se piensa que podría elevarse mediante un nuevo tratamiento experimental. Si θ representa la tasa de respuestas correspondiente al tratamiento experimental, el investigador tendrá que decidir si la hipótesis que expresa su superioridad frente al estándar ($H_1 : \theta > 0{,}5$) es más verosimil que la hipótesis contraria o *hipótesis nula* ($H_0 : \theta = 0{,}5$).

Una vez formuladas las hipótesis a contrastar, el decisor puede desarrollar un simple ensayo clínico aplicando el tratamiento experimental a una muestra de n pacientes (pongamos $n = 40$) y observando las correspondientes respuestas. Los datos obtenidos pueden resumirse en la variable aleatoria $T_n =$ *número de respuestas favorables en una muestra de tamaño* n, la cual tiene distribución $b\,(n = 40; \theta)$. Nótese en que forma los datos contienen información acerca del verdadero valor del parámetro.

Parece natural que en este problema la regla de decisión o *test de hipótesis* tenga la siguiente forma: *rechazar la hipótesis nula* H_0 si $T_n > C$. La cuestion esencial es decidir cuál es el valor adecuado para C. Para ello hay que definir los tipos de error que pueden darse en un problema de contraste de hipótesis. En tal sentido, se dice que se comete un error de tipo *alpha* cuando se rechaza la hipótesis nula siendo ésta cierta y uno de tipo *beta* cuando se acepta una hipótesis nula que es falsa. Los escenarios que pueden presentarse en un problema de contraste de hipótesis se resumen en el siguiente cuadro.

		Decisión	
		Aceptar H_0	Rechazar H_0
H_0	Cierta	Correcta	Error *alpha*
	Falsa	Error *beta*	Correcta

Tabla 2.1: Escenarios posibles en un problema de contraste de hipótesis

En el problema que nos ocupa, el error de tipo *alpha* se cuantifica en la siguiente forma:

$$\alpha = \Pr\left(T_n > C \mid \theta = 0{,}5\right)$$

esto es, la probabilidad de rechazar la hipótesis nula supuesto que esta es cierta ($\theta = 0{,}5$).

En este punto cabe señalar que en un problema de contraste de hipótesis el error *alpha* es inadmisible. En el contexto del ensayo clínico supondría admitir que el tratamiento experimental es superior al estándar cuando en realidad no lo es. Si el tratamiento estándar fuese un placebo, se estaría dando por válido un tratamiento que carece de efectividad (sería tan efectivo como un placebo). En la práctica, ello supone que el máximo valor admisible para el error *alpha* es del 5 %. Esta determinación conduce a que el valor del umbral de rechazo C se determine como solución de la ecuación anterior. En el caso que nos ocupa (n= 40) la solución del problema es $C = 25$.

Surge ahora el problema de evaluar el error beta. Para tal fin consideramos la siguiente función del parámetro θ:

$$\Pr\left(T_{40} > 25 \mid \theta\right)$$

esto es: la probabilidad de rechazar la hipótesis nula H_0 en función del valor de θ.

Esta función recibe el nombre de *función de potencia del test*. Obsérvese que:

$$\Pr\left(T_{40} > 25 \mid \theta = 0{,}6\right) \approx 0{,}31{,}7$$

Esto significa que si el tratamiento experimental eleva la tasa de respuestas del 50 % al 60 %, la probabilidad de que se rechace H_0, esto es, de que el test *detecte* la superioridad del tratamiento experimental es sólo del 31.7 %, lo que significa que para esta alternativa ($\theta = 0{,}6$) el error *beta* del test es del 68.3 %. Tal solución es obviamente

inadmisible para el promotor del nuevo tratamiento. El test sería admisible si el verdadero valor de θ fuera por ejemplo del 70 % dado que $P\left(T_{40} > 25 \mid \theta = 0{,}7\right) = 0{,}8074$ lo que supondría que para esta alternativa el error *beta* sería del 18.26 %. Obsérvese que para $\theta = 0{,}5$ el valor de la función de potencia es el error *alpha* del test.

¿Cómo puede mejorarse la potencia del test?. Obviamente no puede ser a costa de elevar el error *alpha*. La solución pasa por aumentar el tamaño muestral. La tabla muestra para cada valor de n los correspondientes umbrales C de los test de hipótesis y las potencias alcanzadas según los valores del parámetro θ.

La tabla 3.1 y la figura 3.4 muestran para la familia de test de hipótesis *rechazar* H_0 *si* $T_n > C$ las funciones de potencia con errores *alpha* próximos al 5 % para los tamaños muestrales 40 y 100.

		θ				
n	C	0.5	0.6	0.7	0.8	0.9
40	25	0.0403	0.3174	0.8074	0.9921	1
100	58	0.0443	0.6225	0.9928	1	1

Tabla 2.2: Funciones de potencia en función del tamaño n del estudio

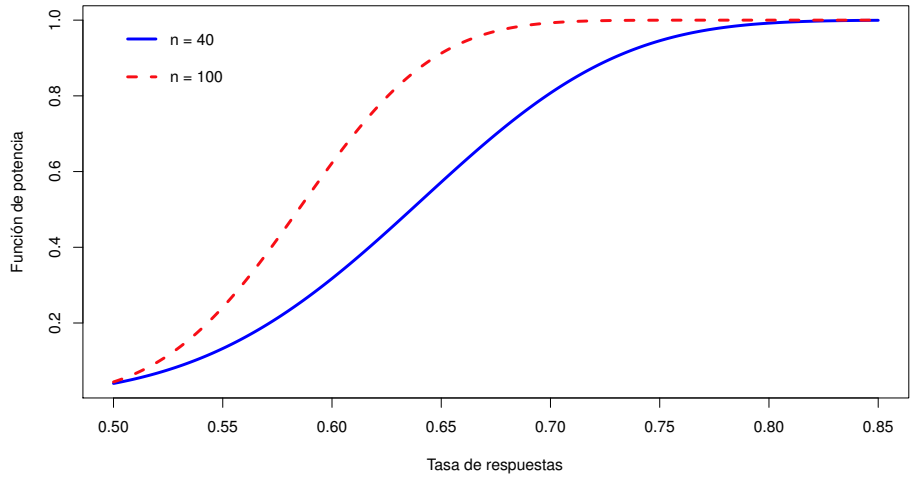


Figura 2.2: Funciones de potencia según tamaño n del estudio

2.3.2. Elementos de un problema de contraste de hipótesis

En un escenario en el que se quiere decidir si el verdadero valor de un parámetro θ pertenece o no a un conjunto Θ_0 se define el problema de contraste de hipótesis por los siguientes elementos:

1. *Hipótesis a contrastar:*

   a) Hipótesis nula $H_0 : \theta \in \Theta_0$

   b) Hipótesis alternativa $H_1 : \theta \notin \Theta_0$

2. Un conjunto de *datos* $\mathfrak{X}$ conteniendo información acerca del verdadero valor del parámetro θ.

3. Una regla de decisión o *test de hipótesis* (T, W) donde $T = T(\mathfrak{X})$ es una función de los datos llamada *test estadístico* y W un conjunto llamado *región crítica*. La regla consiste en *rechazar* H_0 cuando $T \in W$.

Los datos en general son de naturaleza aleatoria, lo que significa que el test estadístico T es una variable aleatoria cuya ley de probabilidad depende del valor del parámetro θ.

2.3.3. Significación y potencia

Para el contraste de hipótesis $H_0 : \theta \in \Theta_0$ frente a $H_1 : \theta \notin \Theta_0$, la función de potencia del test (T, W) se define por:

$$P(T \in W \mid \theta)$$

Esta función expresa la probabilidad de *rechazar* la hipótesis nula H_0 como función del valor del parámetro θ. Las propiedades deseables para esta función son obvias; a saber: debe tomar valores *pequeños* cuando $\theta \in \Theta_0$ y *grandes* en caso contrario. Formalmente, el error de tipo *alpha* se cuantifica como la máxima probabilidad del test de cometer un error de este tipo; esto es:

$$\alpha = \sup_{\theta \in \Theta_0} P(T \in W \mid \theta)$$

El error *alpha* también recibe el nombre de *significación* del test. En la práctica el test se determina en orden a que su significación sea un valor prefijado (habitualmente el 5 %).

El criterio clásico de bondad de un test de hipótesis es el siguiente: entre todos los test con la misma significación, el mejor es el que tiene mayor potencia, para todas los valores del parámetro que corresponden a la hipótesis alternativa; esto es: $\theta \notin \Theta_0$.

En el contraste de hipótesis presentado en 6.1, $\Theta_0 = \{0{,}5\}$. La tabla 6.1 y la figura 6.2 muestran las potencias de tres test de hipótesis alternativos para este contraste. Nótese que los tres tienen aproximadamente la misma significación, pero el basado en la muestra de tamaño $n = 100$ da la máxima potencia para todas las alternativas ($\theta > 0{,}5$).

2.3.4. Nivel mínimo de significación ó *p*-valor.

En las secciones anteriores se ha enfatizado en la importancia de controlar en un contraste de hipótesis el error *alpha* o *significación* del test (como máximo el 5 %). Supóngase ahora que se ha realizado un contraste mediante un test con significación del 5 % y que la decisión que procede es el rechazo de la hipótesis nula H_0. Al investigador debería surgirle la siguiente inquietud: ¿se estará cometiendo un error *alpha*?. Podría también plantearse esta cuestión: ¿con los datos actuales podría mantenerse el rechazo de H_0 con un test de menor significación?. En este escenario surge de manera natural la definición de p-valor como la mínima significación que permite rechazar H_0 con los datos observados.

Para ilustrar esta idea volvamos al contraste de hipótesis descrito en 2.3.1 correspondiente al tamaño muestral $n = 40$ y supóngase que el número de respuestas favorables fue $T_{40} = 30$. Para el test con significación $\alpha = 0{,}05$ la decisión que procede es rechazar H_0 dado que $T_{40} > 25$. ¿Podría mantenerse el rechazo de H_0 utilizando un test de hipótesis de menor significación?. Para responder a esta pregunta nos fijamos en la tabla 2.3 en la que se muestran las significaciones del test correspondientes a varios umbrales de rechazo.

C	25	26	27	28	29	30
α	0.0403	0.0192	0.0083	0.0032	0.0011	0.0003

Tabla 2.3:

Obsérvese que si el umbral C se eleva de 25 a 26, la significación del test baja de 0.0403 a 0.0192 y con el resultado $T_{40} = 30$ puede seguirse rechazando H_0. Si se eleva C a 27, la significación del test baja a 0.0083 y puede aún rechazarse H_0 con el resultado

$T_{40} = 30$. Pero si se eleva C a 30 ya no es posible rechazar H_0. Por tanto, la mínima significación que permite rechazar H_0 con el resultado $T_{40} = 30$ es $\alpha = 0{,}0011$ ($C = 29$). Esta cantidad es por tanto el p-valor.

2.4. Contrastes básicos

2.4.1. Test de la ji-cuadrado para la comparación de tasas

Sean π_E y π_C las probabilidades de un cierto evento F en las poblaciones E y C respectivamente. Para su comparación se considera el contraste de hipótesis $H_0 : \pi_E = \pi_C$ frente a la alternativa $H_1 : \pi_E \neq \pi_C$. Este tipo de contraste es frecuente en ensayos clínicos cuando se compara la tasa de respuestas de un tratamiento experimental frente a un control. Los datos necesarios para el contraste conviene resumirlos en la siguiente tabla:

	E	C	
F	$n_{E,1}\,(e_{E,1})$	$n_{C,1}\,(e_{C,1})$	n_1
F^C	$n_{E,0}\,(e_{E,0})$	$n_{C,0}\,(e_{C,0})$	n_0
	n_E	n_C	n

Tabla 2.4:

Aquí, n_E y n_C representan los tamaños muestrales correspondientes a las poblaciones E y C respectivamente ($n_E + n_C = n$), mientras que $n_{E,1}$ y $n_{C,1}$ son las frecuencias de ocurrencia de F en los referidos grupos ($n_{E,1} + n_{C,1} = n_1$) . Para las no ocurrencias de F los 1 se sustituyen por 0. Nótese que la frecuencia $n_{E,1}$ es una variable aleatoria con distribución de probabilidad $b\,(n_E, \pi_E)$.

Supóngase ahora que H_0 es cierta y que por tanto $\pi_E = \pi_C = \pi$. En tal caso, $E\,[n_{E,1}] = n_E\pi$. Bajo H_0, un estimador natural para π es $\hat{\pi} = n_1/n$. De esta forma, una estimación del valor esperado bajo H_0 para $E\,[n_{E,1}]$ es (sustituyendo π por su estimador) $e_{E,1} = n_E n_1/n$.

Para la realización del contraste consideramos en primer lugar el test estadístico J^2 definido por

$$J^2 = \sum_{i=E,C} \sum_{j=1,0} \frac{(n_{i,j} - e_{i,j})^2}{e_{i,j}}$$

Bajo la hipótesis nula H_0, $J^2 \cong \chi^2(1)$. De esta forma, el test de hipótesis *rechazar* H_0 *si:*

$$J^2 > \chi^2_{1-\alpha}(1)$$

tiene significación α. La cantidad $\chi^2_{1-\alpha}(1)$ es el cuantil $1-\alpha$ de la distribución $\chi^2(1)$.

Ejemplo 2.5. El gen $p53$, también llamado *el guardián del genoma*, se encuentra en el cromosoma 17 y es de importancia crucial en el control del ciclo celular, la apoptosis y reparación del ADN. Un p53 defectuoso podría permitir que las células anormales proliferen siendo el cáncer su peor consecuencia. El gen está formado por los alelos W y M, siendo éste último el desfavorable. Representamos por π_E y π_C las probabilidades de encontrar el genotipo WW en sujetos con cáncer de pulmón y sujetos controles respectivamente. En orden a contrastar las hipótesis $H_0 : \pi_E = \pi_C$ frente a $H_1 : \pi_E \neq \pi_C$ se realizó un estudio de caso control en el que se incluyeron 516 pacientes y 542 controles. Los resultados para el intrón 3 se resumen en la siguiente tabla de contingencia (se consignan simultáneamente los valores esperados bajo H_0).

	Cáncer (E)	Control (C)	Total
WW	$n_{E,1} = 377$	$n_{C,1} = 440$	817
	$e_{E,1} = 398{,}46$	$e_{C,1} = 418{,}54$	
WM ó MM	$n_{E,0} = 139$	$n_{C,0} = 102$	241
	$e_{E,0} = 117{,}54$	$e_{C,0} = 123{,}46$	
Total	516	542	1058

Tabla 2.5:

El valor del test estadístico J^2 es:

$$J^2 = \sum_{i=E,C} \sum_{j=1,0} \frac{(n_{i,j} - e_{i,j})^2}{e_{i,j}} = 9{,}9056$$

La tabla 2.6 muestra diversos niveles de significación α y la decisión que procede de acuerdo con el test J^2

α	$\chi^2_{1-\alpha}(1)$	Decisión sobre H_0
0.05	3.84	Rechazar H_0
0.01	6.63	Rechazar H_0
0.005	7.88	Rechazar H_0
0.0016478	9.9056	Mínimo α que permite rechazar H_0
0.001	10.83	No es posible rechazar H_0

Tabla 2.6: La mínima significación que permite rechazar H_0 es el p-valor (0.0016478)

2.4.2. *t*-test para la comparación de esperanzas

Sean las distribuciones $N(\mu_E, \sigma)$ y $N(\mu_C, \sigma)$ correspondientes a un marcador numérico evaluado en las poblaciones E y C respectivamente. Consideremos el contraste de hipótesis $H_0 : \mu_E = \mu_C$ frente a $H_1 : \mu_E \neq \mu_C$. Para su realización se seleccionan muestras aleatorias $X_{E,1}, \ldots, X_{E,n_E}$ y $X_{C,1}, \ldots, X_{C,n_C}$ de cada una de las ditribuciones. La siguiente tabla muestra para cada uno de los grupos E y C los estimadores de la esperanza y varianza.

$\hat{\mu}_E = (1/n_E)\sum_{j=1}^{n_E} X_{E,j}$	$S_E^2 = (1/(n_E-1))\sum_{j=1}^{n_E}(X_{E,j}-\hat{\mu}_E)^2$
$\hat{\mu}_C = (1/n_C)\sum_{j=1}^{n_C} X_{C,j}$	$S_C^2 = (1/(n_C-1))\sum_{j=1}^{n_C}(X_{C,j}-\hat{\mu}_C)^2$

Tabla 2.7: Estimación de medias y varianzas

La expresión del estimador combinado de la varianza es:

$$S_p^2 = \frac{1}{n_E + n_C - 2}\left\{(n_E - 1)\, S_E^2 + (n_C - 1)\, S_C^2\right\}$$

Definimos finalmente el test estadístico por:

$$T = \frac{\hat{\mu}_E - \hat{\mu}_C}{S_p\sqrt{1/n_E + 1/n_C}}$$

Utilizando los resultados del ejercicio 5.4, puede comprobarse que bajo la hipótesis nula H_0, el test estadístico T sigue una distribución de probabilidad $t(n_E + n_C - 2)$. El test de hipótesis: *rechazar* H_0 *si* $\mid T \mid > t_{1-\alpha/2}(n_E + n_C - 2)$ tiene significación α.

El contraste anterior recibe el nombre de contraste de la t de Student *bilateral* o a *dos*

colas. A menudo el investigador está seguro de que la desigualdad entre las esperanzas no puede darse en un determinado sentido; por ejemplo, ha de ser siempre $\mu_E \geq \mu_C$. En este caso, el contraste de interés sería de tipo *unilateral* o a *una cola* en el cual las hipótesis a contrastar son de la forma $H_0 : \mu_E = \mu_C$ frente a $H_1 : \mu_E > \mu_C$. Para este contraste se utiliza el mismo test estadístico T, pero el test de hipótesis con significación α es de la forma: rechazar H_0 si $T > t_{1-\alpha}\,(n_E + n_C - 2)$.

Capítulo 3

Modelos lineales

En este capítulo de analizarán modelos lineales cuya finalidad es la predicción de una variable aleatoria numérica a partir de una función lineal de un conjunto de variables predictivas.

3.1. Regresión lineal simple

En esta sección se considerará un conjunto de datos de la forma:

$$\{(X_i, Y_i) : i = 1, \ldots, n\}$$

Supondremos que cada Y_i es el resultado de observar una variable aleatoria cuya distribución de probabilidad depende del correspondiente valor X_i.

Ejemplo 3.1. En la figura 3.1 se muestran los niveles de contaminación por gérmenes aerobios-mesófilos (ufc/gr) en las agallas de un conjunto de doradas según el tiempo transcurrido después del sacrificio. Nótese que el nivel de contaminación es una variable aleatoria cuya distribución de probabilidad cambia según los días transcurridos desde el sacrificio.

3.1.1. Definición del modelo

El conjunto de datos $\{(X_i, Y_i) : i = 1, \ldots, n\}$, obedece al modelo de *regresión lineal simple* si:

- $Y_1, \ldots, Y_n$ son variables aleatorias independientes

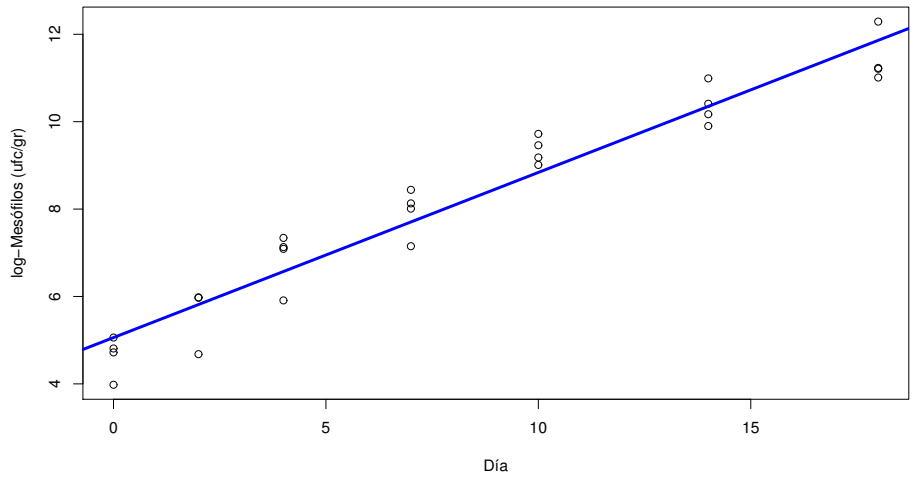


Figura 3.1: Evolución de la contaminación en la agalla de doradas por mesófilos

- Condicionalmente a X_i, la variable aleatoria Y_i tiene distribución de probabilidad $N(\alpha + \beta X_i, \sigma)$ para $i = 1, \ldots, n$

Llamaremos normalmente a los valores X_i *predictores* y a los Y_i, *respuestas*. La esperanza de Y_i para el predictor X_i la representamos por $\mu(X_i) = E[Y_i \mid X_i] = \alpha + \beta X_i$. Nótese que el parámetro β representa el incremento esperado de la respuesta por cada unidad de incremento de X_i. En efecto:

$$\mu(X_i + 1) - \mu(X_i) = \alpha + \beta(X_i + 1) - \alpha - \beta X_i = \beta$$

3.1.2. Estimación de los parámetros

Consideramos para el parámetro β el estimador:

$$\hat{\beta}_n = \frac{1}{nS_X^2} \sum_{i=1}^{n} \left(X_i - \bar{X}\right) Y_i$$

siendo $\bar{X} = (1/n) \sum_{i=1}^{n} X_i$ y $S_X^2 = (1/n) \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$.
El estimador tiene las propiedades:

- $E\left[\hat{\beta}_n\right] = \beta$
- $\mathrm{sd}\left(\hat{\beta}_n\right) = \sigma / \left(S_X \sqrt{n}\right)$

y para el parámetro α:

$$\hat{\alpha_n} = \bar{Y} - \hat{\beta}_n \bar{X}$$

siendo ahora $\bar{Y} = (1/n) \sum_{i=1}^{n} Y_i$. Las propiedades de $\hat{\alpha}_n$ son:

- $E[\hat{\alpha}_n] = \alpha$
- $\mathrm{sd}(\hat{\alpha}_n) = \sigma \sqrt{1/n + \overline{X}^2 / (nS_X^2)}$

Sea $\hat{\mu}_i = \hat{\alpha}_n + \hat{\beta}_n X_i$ la predicción estimada de la respuesta para X_i. Puede comprobarse que $\hat{\mu}_i$ es centrado para $\mu(X_i)$. Además:

$$\mathrm{sd}(\hat{\mu}_i) = \frac{\sigma}{\sqrt{n}} \sqrt{1 + \left(X_i - \bar{X}\right)^2 / S_X^2}$$

lo que demuestra su consistencia. Nótese que la varianza del estimador decrece a medida que la predicción se realiza en un valor X_i próximo a $\bar{X}$.

3.1.3. Contraste de la regresión

El modelo de regresión lineal simple predice bien si la nube de puntos está muy próxima a la recta de esperanzas. La bondad de ajuste la cuantificaremos utilizando el llamado coeficiente de determinación R^2. Para su definición consideramos en primer lugar la siguiente identidad:

$$\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{\mu}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \hat{\mu}_i\right)^2$$

Esta expresión recibe el nombre de *análisis de la varianza de la regresión.* Sus elementos reciben las siguientes denominaciones:

- $\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 =$ *Variabilidad total*
- $\sum_{i=1}^{n} \left(\hat{\mu}_i - \bar{Y}\right)^2 =$ *Variabilidad atribuible a la regresión*
- $\sum_{i=1}^{n} \left(Y_i - \hat{\mu}_i\right)^2 =$ *Variabilidad residual*

Para que los valores esperados de la respuesta $\mu(x) = \alpha + \beta x$ varíen con el predictor x se requiere que $\beta \neq 0$. En tal caso se dice que existe regresión lineal. El contraste de la regresión lineal se formula comparando la hipótesis nula $H_0 : \beta = 0$ frente a $H_1 : \beta \neq 0$. Para la realización del contraste consideramos el siguiente test estadístico:

$$F = \frac{\sum_{i=1}^{n} \left(\hat{\mu}_i - \bar{Y}\right)^2}{\left(1/\left(n-2\right)\right) \sum_{i=1}^{n} \left(Y_i - \hat{\mu}_i\right)^2}$$

Supuesto que sea cierta la hipótesis nula H_0, el test estadístico F tiene distribución de probabilidad $\mathfrak{F}\left(1, n-2\right)$. Para un valor $0 < \alpha < 1$, definimos el siguiente test de hipótesis: *rechazar* H_0 *si* $F > f_{1-\alpha}\left(1, n-2\right)$. Es inmediato probar que el test tiene significación α.

3.1.4. Bondad de ajuste: Coeficiente de determinación R^2

El coeficiente de determinación se define por:

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{\mu}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}$$

La identidad del análisis de la varianza de la regresión garantiza que $0 \leq R^2 \leq 1$. El ajuste perfecto se produce cuando todos los puntos de la nube se encuentran sobre la recta de regresión. Ello supone que $Y_i = \hat{\mu}_i$ y de ahí la variabilidad residual sería nula implicando esto que la atribuible a la regresión coincida con la total. En ese caso, $R^2 = 1$. En el otro extremos nos encontramos que no existe variabilidad atribuible a la regresión, coincidiendo de esta forma la residual con la total lo que implica que $R^2 = 0$. Concluimos que los buenos ajustes se producen cuando el coeficiente R^2 tiene un valor próximo a la unidad.

```
Call:
lm(formula = mesofilo ~ dia, data = sa)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1352 -0.5020  0.1548  0.4740  0.8819

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.05943    0.19346   26.15   <2e-16 ***
dia          0.37787    0.01916   19.72   <2e-16 ***
---
Signif. codes:  0
```

```
Analysis of Variance Table

Response: mesofilo
          Df  Sum Sq Mean Sq F value    Pr(>F)
dia        1 141.621 141.621  388.86 < 2.2e-16 ***
Residuals 25   9.105   0.364
---
Signif. codes:  0
```

3.2. Análisis de la covarianza

Las asociaciones identificadas en los estudios observacionales entre dos variables X e Y a menudo se explican más por la presencia de factores de confusión F que por una relación de causalidad. En tal contexto interesa medir la asociación entre X e Y para un mismo valor de F. Esto es lo que se entiende por *ajustar por la variable de confusión* C. En este capítulo nos ocupamos de analizar el problema de comparar medias entre dos grupos ajustadas por un factor de confusión utilizando los llamados modelos de análisis de la covarianza.

Los datos que se resumen en la siguiente tabla corresponden a un estudio transversal realizado en el municipio de Telde en el que se incluyeron 1,020 personas con más de 30 años.

	HTA (IDF)		
	Si ($n = 406$)	No ($n = 614$)	P
Edad, años	$54{,}5 \pm 11{,}8$	$43{,}8 \pm 10{,}0$	$< 0{,}001$
LDL (mg/dL)	$137{,}3 \pm 31{,}9$	$131{,}8 \pm 32{,}0$	0,007

Tabla 3.1: Edad y nivel de la LDL en los grupos determinados por la presencia o no de hipertensión arterial (International Diabetes Federation). Los valores indican medias ± desviación estándar.

Obsérvese que el valor medio de la LDL en el grupo de sujetos hipertensos es significativamente mayor que en el de los controles ($p = 0{,}007$). Ello no significa que tenga que existir una relación de causalidad entre la hipertensión y la elevación de la LDL. Es bien sabido que los niveles de este marcador tienden a incrementarse al paso de los años. Dado que los hipertensos tienen una edad media estimada superior en casi 11

años a la que tienen los controles, ésta podría explicar total o parcialmente la diferencia en los niveles de LDL entre hipertensos y normotensos. Para investigar esta cuestión utilizaremos el siguiente modelo de análisis de la covarianza:

$$E\left[LDL \mid HTA, EDAD\right] = \theta + \alpha \cdot HTA + \beta \cdot EDAD$$

donde HTA es una variable binaria con valores 1 y 0 los cuales indican la presencia o no de hipertensión arterial. Nótese que el parámetro α se corresponde con la diferencia entre los valores esperados de la LDL de hipertensos y normotensos de una *misma edad*. El parámetro β representa el incremento esperado de la LDL por año cumplido, incremento que el modelo considera común para ambos grupos de estudio.

Este modelo tiene como finalidad, medir la asociación entre los niveles de LDL e hipertensión entre sujetos en una misma edad; esto es: *asociación ajustada por edad*.

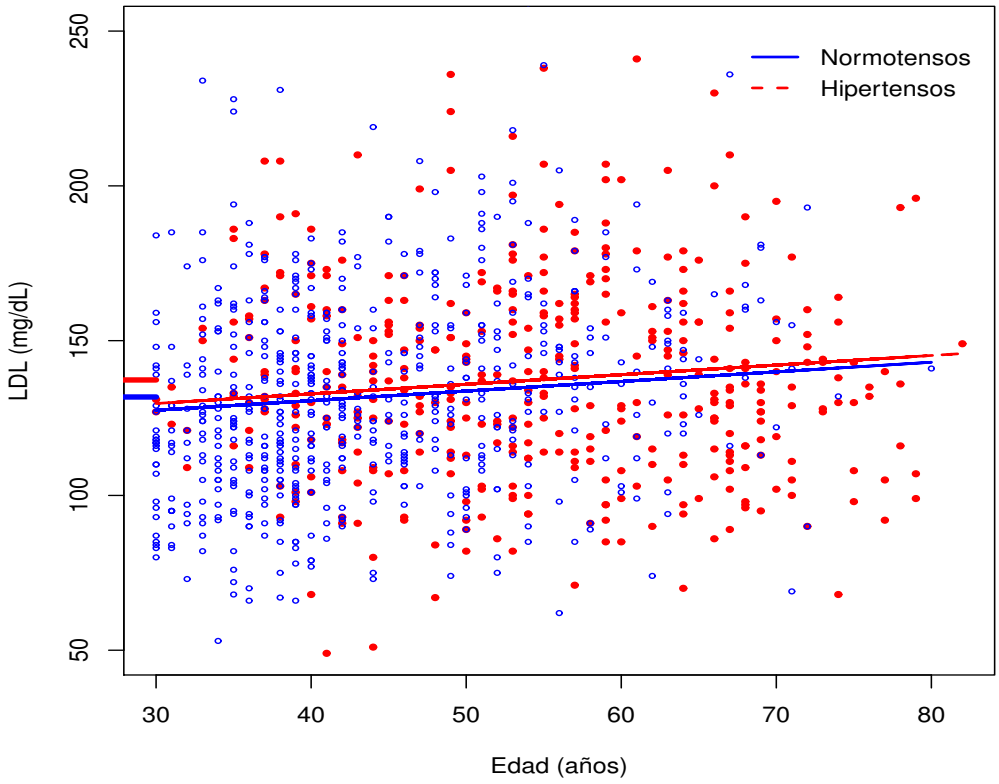


Figura 3.2: Niveles de LDL según edad y estatus de hipertensión arterial (IDF)

En la figura 3.2 muestra los niveles de LDL frente a la edad según la presencia o no de hipertensión arterial. Las marcas que se muestran en el eje de ordenadas corresponden a las medias observadas. Las rectas representan los valores esperados según edad en cada uno de los grupos. La distancia entre las rectas corresponden a la diferencia entre los valores esperados para una misma edad. Nótese que esta diferencia es menor que la que hay entre las medias observadas.

La siguiente tabla muestra la estimación de los parámetros del modelo.

Parámetro	Estimador	Error estándar	P
θ	118.20	4.263	< 0.001
α	2.191	2.265	0.334
β	0.310	0.093	< 0.001

Tabla 3.2:

En orden a resumir los valores de la LDL en magnitudes comparables, obtenemos del modelo de análisis de la covarianza los valores esperados para hipertensos y normotensos tomando como valor de la edad la media de toda la muestra. Esta última puede obtenerse de la tabla 4.1 en la siguiente forma:

$$Edad\,media = \frac{54{,}5 \times 406 + 43{,}8 \times 614}{406 + 614} = 48{,}06$$

De esta forma obtenemos:

- Hipertensos: $E\left[LDL \mid HTA = 1; EDAD = 48{,}06\right] = 135{,}304$
- Normotensos: $E\left[LDL \mid HTA = 0; EDAD = 48{,}06\right] = 133{,}113$

Nótese que la diferencia entre estas medias coinciden exactamente con la estimación del parámetro α (2.191) el cual carece de significación estadística (p=0.334).

Se incluye a continuación la salida de la estimación del modelo correspondiente al paquete R.

```
Call:
lm(formula = LDL ~ HTA_IDF + EDAD)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-84.114 -21.682  -2.172  20.037 134.836

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 118.20443    4.26308  27.727  < 2e-16 ***
HTA_IDF       2.19090    2.26542   0.967 0.333721
EDAD          0.31022    0.09289   3.340 0.000869 ***
---
Signif. codes:  0
```

Capítulo 4

Regresión logística

En este capítulo se consideran modelos de regresión logística cuya finalidad es prdecir una variable aleatoria binaria Y cuyos valores codificaremos por 1 y 0 ($Y \in \{0,1\}$).

4.1. Formulación del modelo

El conjunto de datos de la forma:

$$\{(X_{i,1},\ldots,,X_{i,p};Y_i) \,:\, i=1,\ldots,n\}$$

obedece al modelo de regresión logística o modelo logit si $Y_1,\ldots,Y_n$ son variables aleatorias binarias independientes ($Y_i \in \{0,1\}$) tales que:

$$\text{logit}\ \Pr\left(Y_i=1 \mid X_{i,1},\ldots,X_{i,p}\right)=\beta_0+\beta_1X_{i,1}+\cdots+\beta_pX_{i,p}$$

siendo $\text{logit}\,(x)=\log\left(x/\left(1-x\right)\right) \,:\, 0<x<1$.

La expresión anterior puede expreserse como:

$$\frac{\Pr\left(Y_i=1 \mid X_{i,1},\ldots,X_{i,p}\right)}{1-\Pr\left(Y_i=1 \mid X_{i,1},\ldots,X_{i,p}\right)}=\exp\left(\beta_0+\beta_1X_{i,1}+\cdots+\beta_pX_{i,p}\right)$$

de donde se deduce finalmente:

$$\Pr\left(Y_i=1 \mid X_{i,1},\ldots,X_{i,p}\right)=\frac{\exp\left(\beta_0+\beta_1X_{i,1}+\cdots+\beta_pX_{i,p}\right)}{1+\exp\left(\beta_0+\beta_1X_{i,1}+\cdots+\beta_pX_{i,p}\right)}$$

lo cual garantiza que la probabilidad condicional toma valores en $[0,1]$.

4.2. Interpretación de los coeficientes del modelo logit

Asumimos ahora el modelo logístico para un conjunto de datos de la forma:

$$\{(X_i, Y_i) \,:\, i = 1, \ldots, n\}$$

donde para cada valor $X_i \in \{0, 1\}$, Y_i es una variable aleatoria binaria tal que:

$$\text{logit}\ \Pr\left(Y_i = 1 \mid X_i\right) = \beta_0 + \beta_1 X_i$$

La odd-ratio que mide la asociación entre ambas variables puede obtenerse como:

$$OR = \frac{\Pr\left(Y_i = 1 \mid X_i = 1\right) \Pr\left(Y_i = 0 \mid X_i = 0\right)}{\Pr\left(Y_i = 1 \mid X_i = 0\right) \Pr\left(Y_i = 0 \mid X_i = 1\right)}$$

Veamos ahora que a partir del modelo logístico puede obtenerse la odd-ratio mediante la relación $OR = \exp\left(\beta_1\right)$

En efecto:

$$\text{logit}\ \Pr\left(Y_i = 1 \mid X_i = 1\right) - \text{logit}\ \Pr\left(Y_i = 1 \mid X_i = 0\right) = \beta_1$$

Haciendo operaciones en el primer miembro de la igualdad queda:

$$\log\left\{\frac{\Pr\left(Y_i = 1 \mid X_i = 1\right) / \left(1 - \Pr\left(Y_i = 1 \mid X_i = 1\right)\right)}{\Pr\left(Y_i = 1 \mid X_i = 0\right) / \left(1 - \Pr\left(Y_i = 1 \mid X_i = 0\right)\right)}\right\} = \beta_1$$

Nótese que el primer miembro es $\log OR$. Esta observación completa la demostración.

4.3. Odd-ratio ajustada

Para los datos descritos en el ejemplo 1.1. se había considerado el siguiente modelo:

$$\text{logit}\ \Pr\left(OBESIDAD \mid HTA,\, Edad\right) = \beta_0 + \beta_1 HTA + \beta_2 EDAD$$

Nótese que para cualquier valor de la edad que se considere, puede escribirse:

$$\text{logit}\ \Pr\left(OBESIDAD \mid HTA = 1,\, Edad\right) - \text{logit}\ \Pr\left(OBESIDAD \mid HTA = 0,\, Edad\right) = \beta_1$$

Esto significa que (haciendo operaciones similares a las de la sección anterior) que $\exp(\beta_1)$ corresponde a la odd-ratio que mide la asociación entre obesidad e hipertensión arterial, *a una misma edad*. Por tanto, si ésta odd-ratio fuese estadísticamente significativa, no podría atribuirse tal significación a un efecto de confusión por la edad. La odd-ratio obtenida de esta forma recibe el nombre de *odd-ratio ajustaa por la edad*. Se muestra a continuación la salida de la estimación del modelo mediante el paquete R.

```
Call:
glm(formula = OBESO ~ HTA_OMS + EDAD, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3760  -0.7503  -0.6835   1.1046   1.8216

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.870825   0.301639  -6.202 5.57e-10 ***
HTA_OMS      1.227068   0.156412   7.845 4.33e-15 ***
EDAD         0.014095   0.006328   2.227   0.0259 *
---
Signif. codes:  0
```

Esta salida se puede resumirse en la siguiente tabla:

	Estimación ($\hat{\beta}$)	SD $\left(\hat{\beta}\right)$	p-valor	Odd-Ratio (IC-95 %)
HTA	1.221	0.158	$< .001$	3.39 (2.49 ; 4.62)
$Edad$	0.014	0.006	< 0.028	1.01 (1.00 ; 1.03)

Tabla 4.1: