Regresión linear y correlación

Ejemplo 1

Un estudiante de introducción a la biología desea determinar la relación entre la temperatura y la frecuencia cardíaca en la rana leopardo común, Rana pipiens. Manipula la temperatura en 2 incrementos que van de 2 a 18C y registra la frecuencia cardíaca (latidos por minuto) en cada intervalo. Sus datos se presentan a continuación. Completa un análisis de regresión.

Recording number	*X* Temperature (°Celsius)	*Y* Heart rate (bpm)
1	2	5
2	4	11
3	6	11
4	8	14
5	10	22
6	12	23
7	14	32
8	16	29
9	18	32

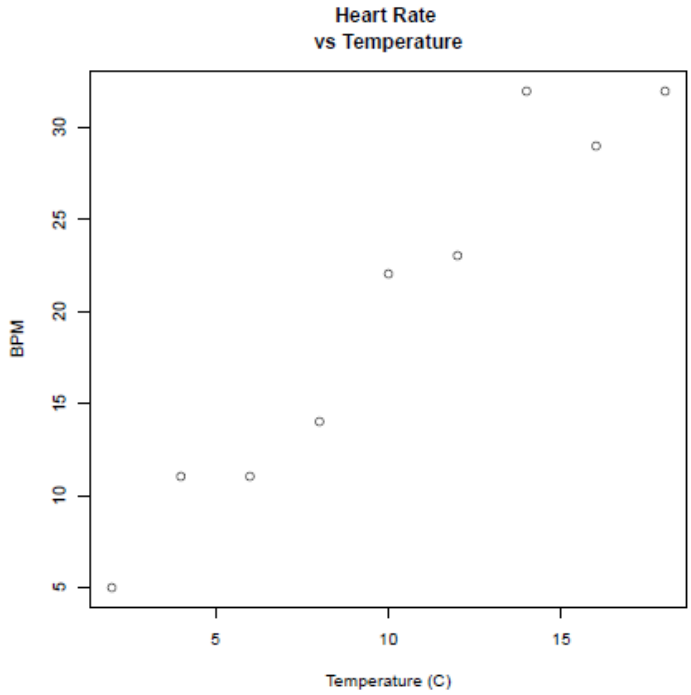
```
> data.Ex10.1 <- read.table("http://waveland.com/Glover-Mitchell/Example10-1.txt",
+ header = TRUE)
> tail(data.Ex10.1, n = 3)

##   Temperature HrtRt
## 7          14    32
## 8          16    29
## 9          18    32
```

Observe la forma de los datos: Cada fila está formada por una temperatura y la correspondiente frecuencia cardíaca. Para determinar si puede haber una relación lineal entre estas dos variables comience por hacer un gráfico de dispersión de los datos

```
> plot(data.Ex10.1$Temperature, data.Ex10.1$HrtRt, main = "Heart Rate
vs Temperature", xlab = "Temperature (C)", ylab = "BPM")
```

Como se ha señalado anteriormente, el siguiente comando también habría trazado los datos.

```
> plot(data.Ex10.1, main = "Heart Rate vs Temperature",
+ xlab = "Temperature (C)", ylab = "BPM")
```



A continuación, realice los cálculos preliminares para el análisis de regresión utilizando la función lm( ). La variable de respuesta es HrtRt y la variable independiente es Temperatura.

lm(y x, datos = dataFrame) Detalles: y es la medida o los datos de respuesta, mientras que x es la variable independiente, normalmente bajo el control del experimentador, y data = dataFrame especifica el marco de datos en el que residen x e y. Observe el orden de las variables de entrada, y se introduce primero y x después. Invertirlas cambia el resultado.

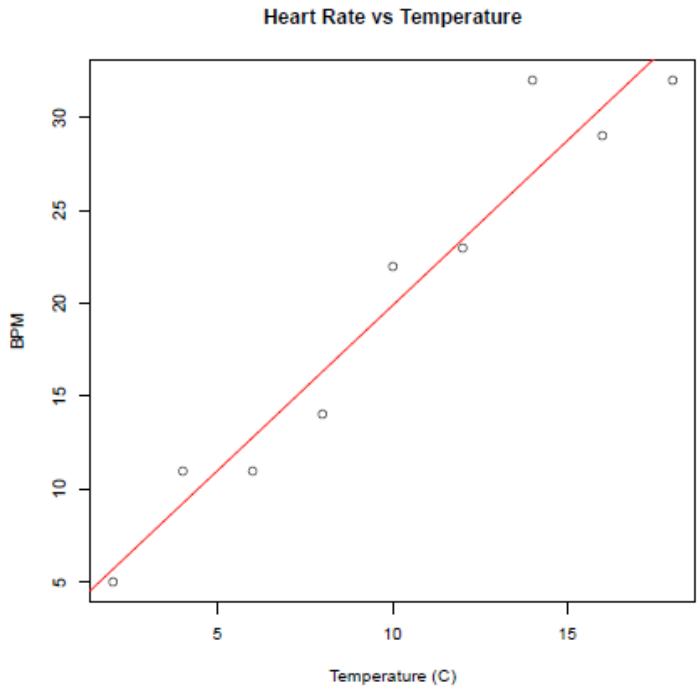
```
> lm.Ex10.1 <- lm(HrtRt ~ Temperature, data = data.Ex10.1)
> lm.Ex10.1       # print the results

##
## Call:
## lm(formula = HrtRt ~ Temperature, data = data.Ex10.1)
##
## Coefficients:
## (Intercept)  Temperature
##        2.14         1.78
```

Los resultados indican que el intercepto de la línea de regresión de mínimos cuadrados es a = 2,14 y la pendiente es b = 1,78. Por lo tanto, la ecuación de la línea de regresión es

$$Y = 2.14 + 1.78X$$

Para trazar la línea de regresión utilice una función adicional con el comando plot( ).

abline(a, b, col = "color") Detalles: a es el intercepto de la línea y b es la pendiente, de ahí el nombre abline( ). El argumento opcional col = "color" hace que la línea se dibuje en el color seleccionado, siendo el predeterminado "negro". El comando debe seguir inmediatamente a un comando plot( ) y la línea se incluirá en el gráfico.



En el contexto de esta pregunta de regresión, a y b pueden especificarse utilizando lm.Ex10.1.

```
> plot(data.Ex10.1, main = "Heart Rate vs Temperature",
+ xlab = "Temperature (C)", ylab = "BPM")
> abline(lm.Ex10.1, col = "red")  # lm.Ex10.1 consists of intercept a and slope b
```

Una vez calculada la ecuación de regresión, compruebe si explica o no una parte significativa de la variabilidad de las Y. Las hipótesis son $H_0: \beta = 0$ y $H_a: \beta \neq 0$

```
> aov.Ex10.1 <- aov(HrtRt ~ Temperature, data = data.Ex10.1)
> summary(aov.Ex10.1)

##             Df Sum Sq Mean Sq F value   Pr(>F)
## Temperature  1    756     756     109 0.000016 ***
## Residuals    7     49       7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El valor P es ínfimo, por lo que se rechaza H_0 : Una parte significativa de la variabilidad de la frecuencia cardíaca se explica por la regresión lineal sobre la temperatura.

Ejemplo 2

Una malacóloga interesada en la morfología de los quitones antillanos, Chiton olivaceus, midió la longitud (anterior-posterior) y la anchura de las ocho placas superpuestas que

componen el caparazón de 10 de estos animales. Sus datos (en cm) se presentan a continuación

Animal	Length	Width
1	10.7	5.8
2	11.0	6.0
3	9.5	5.0
4	11.1	6.0
5	10.3	5.3
6	10.7	5.8
7	9.9	5.2
8	10.6	5.7
9	10.0	5.3
10	12.0	6.3

Analice estos datos como un problema de correlación.

```
> data.Ex10.2 <- read.table("http://waveland.com/Glover-Mitchell/Example10-2.txt",
+ header = TRUE)
> data.Ex10.2

##    Length Width
## 1    10.7   5.8
## 2    11.0   6.0
## 3     9.5   5.0
## 4    11.1   6.0
## 5    10.3   5.3
## 6    10.7   5.8
## 7     9.9   5.2
## 8    10.6   5.7
## 9    10.0   5.3
## 10   12.0   6.3
```

Para determinar si puede haber una relación lineal entre estas dos variables, comience haciendo un gráfico de dispersión de los datos.

```
> plot(data.Ex10.2$Length, data.Ex10.2$Width, main = "Scatterplot Width
versus Length", xlab = "Length (cm)", ylab = "Width (cm)")
```

Para realizar el análisis de correlación utilice

cor.test(x, y, método = "testMethod", nivel conf.= 0.95)
Detalles: x e y son vectores de datos emparejados de la misma longitud; method = "testMethod" especifica el método de correlación a utilizar. El método por defecto es "pearson" y las otras opciones son los métodos no paramétricos "kendall" y "spearman". Como es habitual, conf.level especifica el nivel de confianza con el que se va a realizar la prueba, siendo el valor por defecto 0,95, que equivale a a = 0,05. Hay otras opciones que se detallarán cuando se discutan los métodos no paramétricos.

En este ejemplo los datos están contenidos en data.Ex10.2$Length y data.Ex10.2$Width. Utilice el método = "pearson" y el conf.level = 0,95. Para determinar el coeficiente de correlación y probar las hipótesis $H_0: \rho = 0$ y $H_a: \rho \neq 0$ con α = 0,05

```
> cor.test(data.Ex10.2$Length, data.Ex10.2$Width, method = "pearson",
+ conf.level = 0.95)

##
##  Pearson's product-moment correlation
##
## data:  data.Ex10.2$Length and data.Ex10.2$Width
## t = 11.136, df = 8, p-value = 3.781e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.871327 0.992922
## sample estimates:
##      cor
## 0.969226
```

El coeficiente de correlación (estimación de la muestra) es r = 0,9692. El valor P, muy pequeño, de 0,0000038, significa que debe H_0 rechazarse: Existe una fuerte correlación lineal entre la longitud y la anchura de los caparazones de los quitones. El intervalo de confianza para ρ es [0,871, 0,993]. El intervalo no contiene 0, lo que confirma que debe rechazarse la hipótesis nula.

Ejemplo 3

Se han realizado varios estudios sobre el efecto de la edad relativa en varios tipos de logros (por ejemplo, académicos, deportivos y sociales). En resumen, los efectos de la edad relativa pueden producirse siempre que haya requisitos de edad mínima para participar en una actividad. Por ejemplo, para entrar en el jardín de infancia en Ginebra, Nueva York, donde residen los autores, un niño debe tener 5 años el 1 de diciembre del año escolar en curso. En consecuencia, el 1 de diciembre, los niños de la misma clase de jardín de infancia pueden tener una edad mínima de 5 años y 0 días (los nacidos el 1 de diciembre) y 5 años y 364 días (los nacidos el 2 de diciembre). Del mismo modo, para participar en equipos deportivos (fútbol, liga infantil, etc.) los niños deben alcanzar una edad mínima en una fecha determinada, lo que también da lugar a una diferencia de edad de 1 año en las cohortes iniciales de participantes. ¿Están estas diferencias de edad asociadas a diferencias de rendimiento, es decir, hay efectos relativos de la edad?

Un estudio de DeMeis y Stearns examinó los efectos de la edad relativa en el rendimiento académico y social en Ginebra. Los datos de la siguiente tabla proceden de una parte de su estudio, que examinó el número de alumnos de los grados K a 4 evaluados para el programa de alumnos superdotados y con talento del distrito.

La primera y la segunda columna indican el mes posterior a la fecha de corte en el que nació el alumno. La tercera columna indica el número de alumnos de cada categoría de corte en los grados K a 4 evaluados para el Programa de Alumnos Superdotados y con Talento del distrito.

Birth month	Month after cut-off	Students evaluated
December	1	53
January	2	47
February	3	32
March	4	42
April	5	35
May	6	32
June	7	37
July	8	38
August	9	27
September	10	24
October	11	29
November	12	27

La tabla muestra que, en general, los estudiantes de mayor edad (aquellos con fechas de nacimiento en los primeros meses después de la fecha de corte) tienden a estar sobrerrepresentados y los estudiantes más jóvenes infrarrepresentados en los evaluados para el Programa de Estudiantes Superdotados y con Talento del distrito. Por ejemplo, el mes de diciembre (el primer mes después de la fecha límite) fue el que tuvo más remisiones. ¿Existe una correlación entre estas dos medidas? Determine el coeficiente de correlación t de Kendall y determine si es significativamente diferente de 0 al nivel $\alpha = 0{,}05$. (Basado en los datos reportados en: DeMeis, J. y E. Stearns. 1992. Relationship of school entry age to academic and social performance. The Journal of Educational Research, 86: 20-27.)

```
> data.Ex10.5 <- read.table("http://waveland.com/Glover-Mitchell/Example10-5.txt",
+ header = TRUE)
> data.Ex10.5

##        Month AfterCutOff Students
## 1   December           1       53
```

```
## 2    January           2       47
## 3   February           3       32
## 4      March           4       42
## 5      April           5       35
## 6        May           6       32
## 7       June           7       37
## 8       July           8       38
## 9     August           9       27
## 10 September          10       24
## 11   October          11       29
## 12  November          12       27
```

Para realizar un análisis de correlación utilizando la medida de correlación de Kendall, utilice de nuevo cor.test( ) pero con el método = "kendall". En este ejemplo los datos están contenidos en data.Ex10.5$AfterCutOff y data.Ex10.5$Students. Para determinar el coeficiente de correlación y probar las hipótesis $H_o: \tau = 0$ y $H_a: \tau \neq 0$ con $\alpha = 0{,}05$.

```
> cor.test(data.Ex10.5$AfterCutOff, data.Ex10.5$Students, method = "kendall",
+ conf.level = 0.95)

## Warning in cor.test.default(data.Ex10.5$AfterCutOff, data.Ex10.5$Students, :
Cannot compute exact p-value with ties

##
##  Kendall's rank correlation tau
##
## data:  data.Ex10.5$AfterCutOff and data.Ex10.5$Students
## z = -2.7559, p-value = 0.005853
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## -0.615457
```

El coeficiente de correlación (estimación de la muestra) es τ = -0,6155. Parece que existe una correlación negativa moderadamente fuerte entre el mes posterior a la fecha de corte y las remisiones para la evaluación de superdotados. El valor P de 0,0059 significa que debe rechazarse H_0.

¿Por qué aparece la advertencia en la salida? Vuelva a mirar los datos originales y fíjese en los empates. Los pares de meses febrero y mayo y agosto y noviembre tenían el mismo número de alumnos. La presencia de empates hace imposible calcular un valor P exacto. Si sabe de antemano que habrá empates en los datos, esta advertencia puede evitarse utilizando el argumento adicional: exact = FALSE, de modo que el comando completo podría haber sido

```
> cor.test(data.Ex10.5$AfterCutOff, data.Ex10.5$Students, method = "kendall",
+ exact = FALSE, conf.level = 0.95)
```

Ejemplo 4

Un estudio de Musch y Hay examinó los efectos de la edad relativa en el fútbol en varios países del hemisferio norte y del sur. Para cada país, se investigó una muestra formada por todos los jugadores de la liga de fútbol profesional más alta. A continuación se presentan los datos de Alemania. En Alemania se aplica la fecha de corte del 1 de agosto. Dado que las fechas de corte de participación varían según el país, los jugadores extranjeros fueron excluidos de su análisis. Para cada país, se calculó la distribución de los cumpleaños de los jugadores profesionales por meses. Estas distribuciones de cumpleaños se compararon con la de la población general de ese país.

La primera columna es el mes posterior a la fecha de corte en el que nació el jugador. La segunda columna es el número de futbolistas profesionales nacidos en los respectivos meses del año de la competición. La tercera columna es el número de jugadores de fútbol que se esperaría sobre la base de las estadísticas oficiales de nacimiento, suponiendo que la distribución de los nacimientos de los profesionales del fútbol es la misma que la de la población general, y que no existe ningún efecto de la edad relativa. La cuarta columna es la diferencia entre este número esperado y el observado de jugadores.

Month	Actual players	Expected players	Difference
1	37	28.27	8.73
2	33	27.38	5.62
3	40	26.26	13.74
4	25	27.60	−2.60
5	29	29.16	−0.16
6	33	30.05	2.95
7	28	31.38	−3.38
8	25	31.83	−6.83
9	25	31.16	−6.16
10	23	30.71	−7.71
11	30	30.93	−0.93
12	27	30.27	−3.27

La tabla muestra que, en general, los jugadores de mayor edad en las cohortes de 1 año tienden a estar sobrerrepresentados y los más jóvenes infrarrepresentados en el número total de jugadores profesionales de fútbol en Alemania. Calcule e interprete r_s para estos datos, donde X es el mes e Y es la diferencia entre el número real y el esperado de jugadores profesionales. (Based on data reported by: Musch, J. and R. Hay. 1999. The relative age effect in soccer: Crosscultural evidence for a systematic discrimination against children born late in the competition year. Sociology of Sport Journal, 16: 54–64.)

```
> data.Ex10.8 <- read.table("http://waveland.com/Glover-Mitchell/Example10-8.txt",
+ header = TRUE)
> head(data.Ex10.8, n = 3)

##   Month Actual Expected
## 1     1     37    28.27
## 2     2     33    27.38
## 3     3     40    26.26
```

Para poder realizar la prueba, primero hay que calcular la diferencia entre el número real de jugadores y el número esperado de jugadores para cada mes y poner el resultado en una nueva columna de datos.Ej10.8.

```
> data.Ex10.8["Difference"] <- data.Ex10.8$Actual - data.Ex10.8$Expected
> data.Ex10.8

##    Month Actual Expected Difference
## 1      1     37    28.27       8.73
## 2      2     33    27.38       5.62
## 3      3     40    26.26      13.74
## 4      4     25    27.60      -2.60
## 5      5     29    29.16      -0.16
## 6      6     33    30.05       2.95
## 7      7     28    31.38      -3.38
## 8      8     25    31.83      -6.83
## 9      9     25    31.16      -6.16
## 10    10     23    30.71      -7.71
## 11    11     30    30.93      -0.93
## 12    12     27    30.27      -3.27
```

Para realizar el análisis de correlación mediante el método de Spearman, utilice cor.test( ) y establezca method = "spearman". En este ejemplo, los datos están contenidos en data.Ex10.8$Month y data.Ex10.8$Difference. Para determinar el coeficiente de correlación y probar las hipótesis $H_0: \rho_s = 0$ y $H_a: \rho_s \neq 0$ con $\alpha = 0{,}05$.

```
> cor.test(data.Ex10.8$Month, data.Ex10.8$Difference, method = "spearman",
+ conf.level = 0.95)

##
##  Spearman's rank correlation rho
##
## data:  data.Ex10.8$Month and data.Ex10.8$Difference
## S = 494, p-value = 0.01
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.727273
```

El coeficiente de correlación (estimación muestral) es r_s = -0,7273. El valor P de la prueba es 0,01, por lo que se rechaza H_0 y se acepta H_a. La correlación negativa de rango de Spearman (r_s = -0,7273) es significativa e indica que hay un exceso de jugadores "goliath" (los nacidos a principios del año de la competición) y una falta de jugadores nacidos a finales del año de la competición entre los futbolistas profesionales de Alemania.

Pruebas de bondad de ajuste para datos categóricos

Ejemplo 5

La grave sequía de 1987 en Estados Unidos afectó a la tasa de crecimiento de los árboles establecidos. Se cree que la mayoría de los árboles de las zonas afectadas tienen un anillo de crecimiento de 1987 que es menos de la mitad del tamaño de los demás anillos de crecimiento del árbol. Se recoge una muestra de 20 árboles y 15 presentan esta característica. ¿Apoyan estos datos la afirmación?

Utilice la prueba binomial con α = 0,05, es decir, con el conf.level= 0,95. Sea p la proporción de árboles con un anillo de crecimiento de 1987 que tiene menos de la mitad de su tamaño habitual. La hipótesis alternativa es que "la mayoría de los árboles" tienen esta propiedad, es decir, $H_a: p > 0.5$. La hipótesis nula es $H_0: p \leq 0.5$. El número de aciertos es x = 15, el número de ensayos es n = 20, y la probabilidad de éxito hipotética es la p = 0,5 por defecto. Se trata de una prueba de cola derecha, por lo que alternativa = "mayor". No hay datos que leer para el problema.

binom.test(x, n, p = proportion, alternative = "two. Sided", conf.level = 0.95)
Detalles: x es el número de aciertos; n especifica el número de ensayos; p = proporción especifica la probabilidad hipotética de éxito, con p = 0,5 por defecto; alternativa indica la hipótesis alternativa donde el valor por defecto es "two. Sided" y las otras opciones son "greater" y "less"; y conf.level es el nivel de confianza para el intervalo de confianza devuelto. El valor por defecto es 0,95, que equivale a α = 0,05.

```
> binom.test(x = 15, n = 20, alternative = "greater")

##
##  Exact binomial test
##
## data:  15 and 20
## number of successes = 15, number of trials = 20, p-value = 0.02069
```

```
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.544418 1.000000
## sample estimates:
## probability of success
##                   0.75
```

Como el valor P es inferior a α = 0,05, se rechaza H_0. Hay pruebas de que la mayoría de los árboles tienen anillos de crecimiento para 1987 de menos de la mitad de su tamaño habitual.

Un intervalo de confianza de cola derecha del 95 por ciento para la proporción p de árboles con un anillo de crecimiento de 1987 que es menor que la mitad de su tamaño habitual es [0,5444, 1,0000] y no contiene 0,5. Estamos seguros al 95 por ciento de que la proporción real es al menos 0,5444.

Ejemplo 6

Los pardalotes son aves pequeñas (8-12 cm) que se alimentan sobre todo en el dosel exterior, en lo alto de los eucaliptos. Sin embargo, anidan en agujeros excavados en bancos de tierra. ("Pardalote" viene de la palabra griega que significa "manchado") Hay dos razas diferentes de pardalote estriado, Pardalotus striatus, en el sureste de Queensland. Supongamos que los registros históricos indican que la raza A comprendía el 70% de la población. Un pequeño censo en el monte Coot-tha localiza 18 pardalotes: 11 de la raza A y 7 de la raza B. ¿Indican estas cifras alguna diferencia con el patrón histórico?

```
> binom.test(x = 11, n = 18, p = 0.7)

##
##  Exact binomial test
##
## data:  11 and 18
## number of successes = 11, number of trials = 18, p-value = 0.4429
## alternative hypothesis: true probability of success is not equal to 0.7
## 95 percent confidence interval:
##  0.357451 0.827014
## sample estimates:
## probability of success
##               0.611111
```

El valor P es de 0,4429, que es mucho mayor que α = 0,05. No hay pruebas suficientes para apoyar la afirmación de un cambio en las proporciones de población de las dos razas. El intervalo de confianza del 95 por ciento para la proporción de pardalotes de la raza A es [0,357, 0,827] y por tanto contiene la proporción histórica de 0,7.

Ejemplo 7

En 2011, según la Oficina del Censo de Estados Unidos, el 30,4% de los adultos que residían en el país habían recibido al menos una licenciatura. Una empresa de alta tecnología está estudiando la posibilidad de trasladar su sede a una ciudad en la que se cree que hay una proporción de licenciados universitarios superior a la media. En una muestra aleatoria de n = 480 individuos de esta ciudad, 169 afirman tener una licenciatura. ¿Existen pruebas de que los ciudadanos de esta ciudad están más de que los ciudadanos de esta ciudad tienen un mayor nivel de estudios que el conjunto de la población?

Con una prueba binomial de cola derecha con α = 0,05 es apropiada con las hipótesis nula y alternativa hipótesis $H_0: p \leq 0.304$ y$H_a: p > 0.304$, respectivamente

```
> binom.test(x = 169, n = 480, p = 0.304, alternative = "greater")

##
##  Exact binomial test
```

```
##
## data:  169 and 480
## number of successes = 169, number of trials = 480, p-value =
## 0.01331
## alternative hypothesis: true probability of success is greater than 0.304
## 95 percent confidence interval:
##  0.315934 1.000000
## sample estimates:
## probability of success
##               0.352083
```

El valor P de la prueba es 0,013. Como este valor es inferior a α = 0,05, se rechaza H_0. De forma equivalente, un intervalo de confianza del 95 por ciento [0,316, 1,000] no contiene la proporción hipotetizada p = 0,304 de titulados universitarios. Hay pruebas de que los ciudadanos de esta ciudad tienen más educados que la población estadounidense en su conjunto.

Ejemplo 8

Supongamos que las hipótesis son las mismas que en el ejemplo 6, pero que el tamaño de la muestra es n = 180 aves: 110 de la raza A y 70 de la raza B. ¿Indicarían estas cifras alguna diferencia con respecto al patrón histórico?

Una prueba binomial de dos caras con a = 0,05 es adecuada y las hipótesis nula y alternativa siguen siendo las mismas que en el ejemplo 6 El estadístico de prueba o número de aciertos es x = 110, el número de ensayos es n = 180.

```
> binom.test(x = 110, n = 180, p = 0.7)

##
##  Exact binomial test
##
## data:  110 and 180
## number of successes = 110, number of trials = 180, p-value =
## 0.01149
## alternative hypothesis: true probability of success is not equal to 0.7
## 95 percent confidence interval:
##  0.535762 0.682736
## sample estimates:
## probability of success
##               0.611111
```

El valor P es 0,0115, que es menor que α = 0,05. El intervalo de confianza del 95 por ciento para la proporción de pardalotes de raza A es [0,536, 0,683] y no contiene la proporción histórica de 0,7. Rechace H_0 esta vez. Hay pruebas de que ha habido un cambio en el patrón histórico. El beneficio de un mayor tamaño de la muestra puede verse en el intervalo de confianza mucho más estrecho para la proporción p de pardalotes de raza A. Más trabajo vale la pena.

Ejemplo 9

Un stent es un pequeño tubo de malla metálica, a menudo insertado en una arteria, que actúa como un andamio para proporcionar apoyo para mantener la arteria abierta. Los stents son un tratamiento habitual para mantener abiertas las arterias totalmente ocluidas en pacientes coronarios. Pueden implantarse incluso varios días después de un infarto, bajo el supuesto de que cualquier procedimiento que aumente el flujo sanguíneo supondrá una mejora de la tasa de supervivencia.

Se realizó un estudio para determinar si el momento en que se implantan los stents en las arterias de los pacientes modifica la eficacia del tratamiento. En el estudio se asignaron aleatoriamente a dos grupos 2166 pacientes estables que habían sufrido infartos en los 3 a 28 días anteriores y tenían una oclusión total de una arteria relacionada con este ataque. El grupo 1 estaba formado por 1.082 pacientes que recibieron stents y una terapia médica óptima, mientras que el grupo 2 estaba formado por 1.084 pacientes que sólo recibieron terapia médica sin stents. Los resultados fueron que 171 de los pacientes del Grupo 1 y 179 del Grupo 2 habían muerto al cabo de cuatro años.

Los investigadores esperaban encontrar una reducción de la tasa de mortalidad en los pacientes que recibieron stents en las condiciones descritas. ¿Hubo alguna prueba de ello?

prop.test(x, n, alternative = "two.sided", conf.level = 0.95)
Detalles: x es un vector bidimensional que especifica los aciertos de cada grupo; n es un vector bidimensional que especifica el número de ensayos de cada grupo; alternative indica la hipótesis alternativa donde el valor por defecto es "two. Sided" y las otras opciones son "greater" y "less"; y conf.level es el nivel de confianza para el intervalo de confianza devuelto. El valor predeterminado es 0,95, que equivale a equivalente a α = 0,05.

Sin embargo, prop.test( ) utiliza un método diferente (chi-cuadrado).Así que hemos proporcionado la función z.prop.test( ) que utiliza este último método. Las dos funciones sólo difieren ligeramente en el valor P calculado.

Es adecuada una prueba de proporciones de dos muestras de cola izquierda al nivel α = 0,05. Las hipótesis son $H_0: p1 \geq p2$ y $H_a: p1 < p2$. Cree los vectores que contienen el número de éxitos y el número de ensayos.

```
> x <- c(171, 179)               # successes (deaths) for Group 1 and  Group 2
> n <- c(1082, 1084)             # trials for Group 1 and  Group 2
> source("http://waveland.com/Glover-Mitchell/z.prop.txt") # Download source file

## Downloaded: z.prop.test( ).

> z.prop.test(x, n, alternative = "less")

##
##  2-sample Z-test for equality of proportions
##
## data:  x out of n
## z-statistic = -0.4481, p-value = 0.327
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000  0.0189272
## sample estimates:
##   prop 1   prop 2
## 0.158041 0.165129
```

El valor P resultante es 0,327 y es mucho mayor que a = 0,05. No se rechaza H0; no hay pruebas de una reducción significativa de la tasa de mortalidad con la implantación de stents en pacientes entre 3 y 28 días después del infarto. Para completar, aquí está el análisis usando prop. Test Los resultados son bastante similares.

```
> prop.test(x, n, alternative = "less")

##
##  2-sample test for equality of proportions with continuity
##  correction
##
```

```
## data:  x out of n
## X-squared = 0.1519, df = 1, p-value = 0.3484
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000  0.0198505
## sample estimates:
##   prop 1   prop 2
## 0.158041 0.165129
```

Ejemplo 10

El primer ejemplo que consideramos es un modelo extrínseco. En un modelo extrínseco, todos los parámetros de la población necesarios para el análisis se suponen antes de recoger los datos. No es necesario estimar ningún parámetro (por ejemplo, la media o la varianza) a partir de los datos.

chisq.test(x = observado, p = esperado)
Detalles: x = observado es el vector que contiene los recuentos observados de cada categoría; p = esperado es el vector que especifica las proporciones esperadas para cada categoría. (Se pueden utilizar los valores esperados en lugar de las proporciones.) Los recuentos y las correspondientes proporciones esperadas (valores) deben aparecer en el en el mismo orden. Hay argumentos adicionales que pueden utilizarse con esta función que se detallarán más adelante.

Los cuatro relojes, Mirabilis jalapa, son plantas originarias de América tropical. Su nombre se debe a que sus flores tienden a abrirse al final de la tarde. Las plantas individuales de las cuatro horas pueden tener flores rojas, blancas o rosas. Se cree que el color de las flores en esta especie está controlado por un único locus genético con dos alelos que expresan una dominancia incompleta, de modo que los heterocigotos tienen flores rosas, mientras que los homocigotos para un alelo tienen flores blancas y los homocigotos para el otro alelo tienen flores rojas. Según los principios genéticos mendelianos, la autopolinización de las plantas de flor rosa debería producir una progenie con flores rojas, rosas y blancas en una proporción de 1:2:1. Un horticultor autopoliniza varias plantas de flor rosa y produce 240 progenies con 55 de flor roja, 132 de flor rosa y 53 de flor blanca. ¿Son estos datos razonablemente consistentes con el modelo mendeliano de un locus de un solo gen con dominancia incompleta?

Las hipótesis son

- H_0: Los datos son consistentes con un modelo mendeliano (las flores rojas, rosas y blancas se dan en la proporción 1:2:1).
- H_a: Los datos son inconsistentes con un modelo mendeliano.

Los tres colores son las categorías. Para calcular las frecuencias esperadas, no es necesario estimar ningún parámetro, ya que las proporciones mendelianas (25% de rojo, 50% de rosa y 25% de blanco) se establecieron antes de la recogida de datos, por lo que se trata de una prueba extrínseca. Para realizar la prueba, primero se crea el vector observado que contiene los recuentos observados para los diferentes colores y luego se crea el vector esperado. Observe a continuación que se han utilizado fracciones en lugar de proporciones decimales. En general, el uso de fracciones es más preciso que el uso de decimales que pueden necesitar ser redondeados, aunque en este caso el uso de esperado <- c(0,25, 0,5, 0,25) habría producido el mismo resultado.

```
> observed <- c(55, 132, 53)
> expected <- c(1/4, 2/4, 1/4)
> chisq.test(x = observed, p = expected)

##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 2.4333, df = 2, p-value = 0.2962
```

El valor P resultante es 0,2962 y es mayor que α = 0,05. Se acepta H_0. Hay apoyo al modelo genético mendeliano.

Ejemplo 11

El siguiente ejemplo que consideramos es un modelo intrínseco. El modelo intrínseco requiere una estimación de algún parámetro de la población a partir de los datos recogidos.

La distribución de Poisson es útil para describir acontecimientos aleatorios poco frecuentes, como las tormentas graves. En el periodo de 98 años comprendido entre 1900 y 1997, hubo 159 huracanes que tocaron tierra en Estados Unidos. ¿Sigue el número de huracanes que tocan tierra/año (véase la tabla siguiente) una distribución de Poisson? (Based on data reported in: Bove, M. et al. 1998. Effect of El Niño on U.S. landfalling hurricanes, revisited. Bulletin of the American Meteorological Society, 79: 2477–2482.)

Huracanes/año	0	1	2	3	4	5	6
Frecuencia	18	34	24	16	3	1	2

Las hipótesis son

- H_0: El número anual de huracanes que llegan a tierra en Estados Unidos sigue una distribución de Poisson.
- H_a: El número anual de huracanes que tocan tierra en EE.UU. es inconsistente con una distribución de Poisson.

Realizar una prueba de chi-cuadrado intrínseca requiere varios cálculos previos. Para utilizar la fórmula de la densidad de Poisson necesitamos especificar el número medio de huracanes/año. En este ejemplo, este parámetro debe estimarse a partir de los datos, y esto es lo que hace que se trate de un modelo intrínseco. A partir de la información dada, la estimación de la media μ es

```
> mu.est <- round(159/98, digits = 3)
> mu.est

## [1] 1.622
```

Utilizar la función de densidad de probabilidad de Poisson dpois(x, mu) para determinar la probabilidad de x huracanes por año. Por ejemplo, para determinar la probabilidad de 2 huracanes al año, utilice

```
> dpois(2, mu.est)

## [1] 0.259804
```

Para calcular las probabilidades de 0 a 5 huracanes/año, utilice el vector c(0:5) como x en dpois( ).

```
> dpois(c(0:5), mu.est)

## [1] 0.1975033 0.3203503 0.2598041 0.1404674 0.0569595 0.0184777
```

La última categoría "observada" en los datos dados es de 6 huracanes/año. Sin embargo, teóricamente existe una pequeña probabilidad de que se observen más de 6 huracanes en un año, por lo que utilizamos "6 o más huracanes/año" como última categoría.

$$P\left(x \geq \frac{6huracanes}{años}\right) = 1 - p\left(x < 6\frac{huracanes}{año}\right) = 1 - P\left(x \leq 5\frac{huracanes}{año}\right)$$

```
> ppois(5, mu.est, lower.tail = FALSE)    # P(x >= 6)

## [1] 0.00643757
```

Ponga todas estas probabilidades en una sola variable llamada expected.proportionsmediante la operación c( )

```
> expected.proportions <- c(dpois(c(0:5), mu.est), ppois(5, mu.est, lower.tail = FALSE))
> expected.proportions

## [1] 0.19750330 0.32035035 0.25980413 0.14046743 0.05695954 0.01847768
## [7] 0.00643757
```

La prueba χ^2 requiere que todos los valores esperados sean al menos 5. Para determinar los valores esperados, multiplique las Proporciones de x huracanes/año por el número de años, 98.

```
> expected.values <- 98*expected.proportions
> expected.values

## [1] 19.355323 31.394334 25.460805 13.765809  5.582035  1.810812  0.630882
```

La frecuencia esperada de observar 0 huracanes/año es de 19,35, 1 huracán/año es de 31,39, y así sucesivamente, hasta 6 o más huracanes/año, que es de 0,63. Las frecuencias esperadas para 5 y para 6 o más huracanes/año son ambas inferiores a 5, por lo que no satisfacen los supuestos de la prueba. Para sortear esta dificultad, se deben combinar las categorías adyacentes hasta que el número esperado sea al menos 5. En este caso, se requiere colapsar las tres últimas categorías en una sola categoría " ≥ 4", o, en términos de R, ppois(3, mu.est, llower. Tail= FALSE).

```
> # the probabilities of 0 to 3 hurricanes and 4 or more hurricanes combined
> expected <- c(dpois(c(0:3), mu.est), ppois(3, mu.est, lower.tail = FALSE))
> expected

## [1] 0.1975033 0.3203503 0.2598041 0.1404674 0.0818748
```

Ahora estamos preparados para realizar una prueba de chi-cuadrado intrínseca. R no tiene una función incorporada para hacer esto, pero hemos creado el intrinsic.chisq.test( ) para llenar esta laguna

intrinsic.chisq.test(x = observado, p = esperado, est.params)
Detalles: Al igual que con chisq.test( ), x = observado es un vector que contiene los recuentos observados de cada categoría y p = esperado es un vector que especifica las proporciones esperadas para cada categoría. Los recuentos y las correspondientes proporciones esperadas deben aparecer en el mismo orden. Por último, est.params es el número de parámetros que se estimaron a partir de los datos para determinar las proporciones esperadas. Debe ser un número entero no negativo y el valor por defecto es est.params = 1. Se utiliza para determinar los grados de libertad en el análisis.

En este ejemplo se estimó un parámetro, μ, a partir de los datos, por lo que est.params = 1. A partir de los datos originales, el número observado de años sin huracanes fue 18, con 1 huracán fue 34, con 2 huracanes fue 24, con 3 huracanes fue 16, y con 4 o más huracanes fue de 6. Poner estos valores en una variable llamada observado.

```
> observed <- c(18, 34, 24, 16, 6)
```

Las probabilidades correspondientes se han almacenado en expected. Así que después de descargar la función requerida todo el análisis se realiza con un solo comando.

```
> source("http://waveland.com/Glover-Mitchell/intrinsic.chisq.txt")

## Downloaded: intrinsic.chisq.test( ).

> intrinsic.chisq.test(x = observed, p = expected, est.params = 1)

##
##  Chi-squared test for given probabilities
##
## data:  x = observed and p = expected
## X-squared = 1.268, df = 3, p-value = 0.7367
```

El valor P es de 0,7367, que es mucho mayor que α = 0,05. No podemos rechazar la afirmación de que el número anual de huracanes que tocan tierra en Estados Unidos está descrito por un proceso de Poisson con $\mu = 159/98 \approx 1.622$. Acepte H_0.

Ejemplo 12

En un estudio sobre el comportamiento del cangrejo ermitaño en Point Lookout, en la isla de North Stradbroke, se recogió una muestra aleatoria de tres tipos de conchas de gasterópodos. Cada concha se calificó como ocupada por un cangrejo ermitaño o vacía. No se tomaron muestras de conchas con gasterópodos vivos. ¿Prefieren los cangrejos ermitaños un determinado tipo de caparazón? ¿O los cangrejos ermitaños ocupan los caparazones en la misma proporción que los vacíos? En otras palabras, ¿es la especie de concha de concha es independiente de si está ocupada?

Species	Occupied	Empty	Total
Austrocochlea	47	42	89
Bembicium	10	41	51
Cirithiidae	125	49	174
Total	182	132	314

Las hipótesis son

- H_0: El estado (ocupado o no) es independiente de la especie de concha.
- H_a: El estado no es independiente de la especie de concha.

El aspecto más complicado del análisis es la introducción de los datos. Los datos deben introducirse en una tabla (o matriz) antes de poder analizarlos. Introduzca los datos, excluyendo los totales, fila por fila (o columna por columna, véanse los comentarios más abajo). A continuación, vincule las filas con el comando rbind( ) (o vincule las columnas con cbind( )) para formar la tabla de datos. Por último, añada los nombres de las columnas (o de las filas).

```
> Austrocochlea <- c(47, 42)      # or Occupied <- c(47, 10, 42)
> Bembicium <- c(10, 41)          # and Empty <- c(42, 41, 49)
> Cirithiidae <- c(125, 49)
> Example11.10.table <- rbind(Austrocochlea, Bembicium, Cirithiidae)
> # or Example11.10.table <- cbind(Occupied, Empty)
> colnames(Example11.10.table) <- c("Occupied", "Empty")
> # rownames(Example11.10.table) <- c("Austrocochlea", "Bembicium", "Cirithiidae")
> Example11.10.table

##              Occupied Empty
## Austrocochlea       47    42
## Bembicium           10    41
## Cirithiidae        125    49
```

El análisis se realiza ahora con

```
> chisq.test(Example11.10.table)

##
##  Pearson's Chi-squared test
##
## data:  Example11.10.table
## X-squared = 45.5116, df = 2, p-value = 1.31e-10
```

El diminuto valor P indica que debe rechazarse H0. Hay razones para creer que las especies de caparazón y la ocupación no son independientes. Es decir, los cangrejos ermitaños son "selectivos" en la especie de concha que ocupan.

En este punto es útil mencionar un par de características más de R. Para sumar los totales marginales a la tabla del Ejemplo 11.10, utilice

```
> addmargins(Example11.10.table)

##               Occupied Empty Sum
## Austrocochlea       47    42  89
## Bembicium           10    41  51
## Cirithiidae        125    49 174
## Sum                182   132 314
```

Para ver los valores esperados utilizados en la prueba, utilice

```
> chisq.test(Example11.10.table)$expected

##               Occupied   Empty
## Austrocochlea  51.5860 37.4140
## Bembicium      29.5605 21.4395
## Cirithiidae   100.8535 73.1465
```

Referencia

[1] K. Mitchell y T. Glover, *An Introduction to biostatistics using R*.