Introducción

El objetivo de los métodos estadísticos es utilizar la evidencia empírica para mejorar la comprensión de una población que puede incluir un grupo de sujetos o individuos y, a partir de la información recopilada y analizada, se pueden tomar decisiones informadas. El objetivo de muchos estudios científicos es comprender los cambios en características específicas (variables) en una población de interés. Por ejemplo, podríamos estar interesados en el rango de temperatura corporal normal de las personas sanas, el tamaño de los tumores en pacientes con cáncer de mama o la tasa de crecimiento de los nogales [1].

El análisis estadístico comienza con un problema científico, generalmente presentado en forma de prueba de hipótesis o problema de predicción. Los métodos estadísticos se utilizan para probar hipótesis basadas en datos empíricos. A través de estos métodos, podemos decidir si una hipótesis debe ser rechazada. Estas decisiones, a su vez, nos ayudan a tomar decisiones más informadas sobre las cuestiones científicas que inspiran nuestra investigación. Los problemas científicos a veces se presentan como problemas de predicción. La predicción es el proceso de utilizar un conjunto de variables predictoras para adivinar el valor de una variable de respuesta. Por ejemplo, podríamos querer usar la circunferencia abdominal para predecir el porcentaje de grasa corporal o el tamaño del tumor para predecir el tiempo de supervivencia de los pacientes con cáncer. Recopilamos datos de muestras de población para comprender el panorama general de la misma población [1].

El proceso de usar datos para sacar conclusiones sobre una población completa, mientras se reconoce el grado de incertidumbre en nuestros resultados, se llama inferencia estadística. Nos permiten tomar decisiones sobre las cuestiones científicas que motivan nuestra investigación y análisis de datos. Por lo general, usamos programas de computadora para preparar, explorar y analizar datos. Los programas estadísticos más utilizados son MINITAB, MATLAB, R, SAS, SPSS y STATA [1].

R

R es tres cosas: un proyecto, un lenguaje y un entorno de software. Como proyecto, R forma parte del Proyecto de software libre de GNU (www.gnu.org), cuyo objetivo es compartir software de forma gratuita, sin restricciones de licencia. Por lo tanto, usar R no le cuesta nada al usuario [2].

El proyecto R es una actividad académica y la mayoría de los colaboradores son estadísticos. El proyecto R fue iniciado en 1995 por un grupo de estadísticos de la Universidad de Auckland y ha seguido creciendo. Dado que la estadística es una ciencia interdisciplinaria, el uso de R atrae a investigadores académicos de todos los campos de la estadística aplicada. R tiene muchos usuarios de nicho, que incluyen: estadísticas ambientales, econometría, aplicaciones médicas y de salud pública, y bioinformática [2].

Como lenguaje, R es un dialecto del lenguaje S, un lenguaje de programación estadística orientado a objetos desarrollado a finales de los años 80 por los laboratorios Bell de AT&T.

Ejemplos

El fundamento de la estadística bayesiana es una regla de probabilidad bastante sencilla conocida regla de Bayes (también llamada teorema de Bayes, ley de Bayes). La regla de Bayes es una identidad contable que obedece a los axiomas de la probabilidad [1] .

esta simple regla que es la fuente de las ricas aplicaciones y la controversia sobre los elementos subjetivos en la estadística bayesiana

La regla de Bayes comienza relacionando la probabilidad conjunta y la condicional

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Sin embargo, la probabilidad de A y B también puede reescribirse como:

$$P(A \cap B) = P(A|B) * P(A)$$

$$P(A \cap B) = P(B|A) * P(B)$$

Se conoce como regla de Bayes. La regla de Bayes se denomina a veces regla de la probabilidad inversa. Esto se debe a que muestra cómo una probabilidad condicional P(B|A) puede convertirse, o invertirse, en una probabilidad condicional P(A|B).

La tabla 9-1 muestra la probabilidad conjunta de dos eventos, el evento A que es una proteína unida a la membrana y el evento B que tiene una alta proporción de residuos hidrofóbicos (aminoácidos). Las dos columnas con datos representan las distribuciones marginales de A, siendo una proteína unida a la membrana, y el complemento de A (escrito como ~A o A^c), no siendo una proteína unida a la membrana. Las dos filas representan las distribuciones marginales de B, que tiene un alto contenido hidrofóbico y el complemento de B^c (~B o B^c) [1]. Cada celda representa la probabilidad conjunta de dos eventos

		Tipo de proteína	
		Membrana unida (A)	Membrana no unida (A)
Contenido hidrofóbico	Alto (B)	0.3	0.2
	Bajo (~B)	0.1	0.4

Supongamos que queremos calcular la probabilidad de que una proteína tenga un alto contenido hidrofóbico dado que es una proteína unida a una membrana. Para ello podemos aplicar la regla de Bayes de esta forma:

$$P(B|A) = \frac{P(A|B)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

P (A∩B) es la probabilidad conjunta de A y B se encuentra simplemente en la celda de la tabla de probabilidad conjunta y es 0,3. P(B) se puede calcular por la ley de la probabilidad

total, calculando la P(B|A) P(A)+P(B|~A) P(~A) o, ya que tenemos la tabla por la suma de la fila del suceso B, que es 0,5.

$$P(B|a) = \frac{P(A \cap B)}{P(B)} = \frac{0.3}{0.5} = 0.6$$

Y llegamos a la conclusión de que, dado que la proteína está unida a la membrana, hay un 60% de posibilidades de que la proteína tenga un alto contenido de residuos hidrofóbicos.

Ejemplo 2

Una cepa particular de ratones consanguíneos tiene una forma de distrofia muscular que tiene una base genética clara. En esta cepa, la probabilidad de aparición de la distrofia muscular en cualquier ratón nacido de padres específicos es de 1/4. Si se crían 20 crías de estos padres, encuentre las siguientes probabilidades [3].

a) Menos de 5 tendrán distrofia muscular.
b) Cinco tendrán distrofia muscular.
c) Menos de 8 y más de 2 tendrán distrofia muscular.

Nota

La distribución de variables discretas más importante en biología es la distribución binomial. La función de densidad de probabilidad de una variable aleatoria binomial se caracteriza por los dos parámetros n, el número de ensayos o tamaño de la muestra, y p, la probabilidad de éxito en un ensayo. R ha incorporado funciones de densidad y de densidad acumulada para las variables aleatorias binomiales.

pbinom(x, n, p = probability)
Detalles: x es el número de éxitos; n especifica el número de ensayos; p = probabilidad especifica la probabilidad de éxito, con p = 0,5 por defecto

El número de pruebas es n = 20 y la probabilidad de éxito es de p = ¼.

a) Para determinar la probabilidad de que menos de 5 tengan distrofia muscular es de

```
> pbinom(4, 20, p = 1/4)  # pbinom is the CDF

## [1] 0.414842
```

b) Para determinar que cinco tendrán distrofia muscular

```
> dbinom(5, 20, p = 1/4)  # dbinom is the pdf

## [1] 0.202331
```

c) Determinar la probabilidad de que menos de 8 y más de 2 tengan distrofia muscular

```
> pbinom(7, 20, p = 1/4) - pbinom(2, 20, p = 1/4)

## [1] 0.806928
```

Ejemplo 3

Una ictióloga que estudia el Cottus ricei (cabeza de cuchara) capturas ejemplares en una gran red de cerco que arrastra por el lago. Sabe, por su experiencia de muchos años, que por término medio capturará 2 peces por cada recorrido de arrastre. Encuentra las probabilidades de capturar [3]

1. No hay peces en un tramo concreto.
2. Menos de 6 peces en una zona determinada;
3. Entre 3 y 6 peces en una zona determinada.

Nota

La distribución de Poisson

La variable aleatoria discreta que describe el número de ocurrencias de un suceso en un intervalo continuo de tiempo o espacio, a veces llamado sucesos raros y aleatorios, surge de lo que se conoce como procesos de Poisson. El número esperado de sucesos durante un periodo cualquiera es el mismo que durante cualquier otro periodo de la misma duración y se denota por μ. Para poder utilizar una distribución de Poisson, el valor de μ debe conocerse o determinarse a partir de una muestra. R ha incorporado funciones de densidad y de densidad acumulativa acumulativas, dpois() y ppois( ), respectivamente.

dpois(x, mu)
ppois(x, mu, lower.tail = TRUE)

x es el número de aciertos y mu especifica el número esperado (media) de aciertos en un periodo. La función de densidad acumulativa tiene un argumento adicional llamado lower.tail. Por defecto, se establece en TRUE para que se utilice la cola izquierda de la función de densidad. Utilizando ppois(n, mu.est, lower.tail = FALSE), la función devuelve el valor 1 - ppois(n, mu.est) = 1-F(n), cual es igual a P(X>n) o la cola derecha de la función.

mu = 2 peces/tramo

1. Con no peces

```
> dpois(0, 2)     # pdf: P(X= 0) = f(0)

## [1] 0.135335
```

2. Con menos de 6 peces P(X<6) = F(5)

```
> ppois(5, 2)     # CDF: F(5)

## [1] 0.983436
```

3. Para entre 3 y 6 peces, P (3 < X < 6) = F (5)- F (3)

```
> ppois(5, 2) - ppois(3, 2)

## [1] 0.126313
```

Ejemplo 4

Suponga que la presión arterial diastólica X en mujeres hipertensas se centra en unos 100 mmHg y tiene una desviación estándar de 16 mmHg y se distribuye normalmente. Encuentre P (X < 90), P (X > 124), P (96 < X < 104). Posterior, encuentre x de modo que P (X ≤ x) = 0,95 [3].

Nota:
La distribución normal
Las distribuciones normales dependen de dos parámetros, la media μ y la desviación estándar σ. En R se utilizan las funciones dnorm () y pnorm () para la pdf (Función de densidad de probabilidad) y la CDF (función de distribución acumulativa) de varias distribuciones normales.

dnorm(x, mean = mu, sd = sigma)
pnorm(x, mean = mu, sd = sigma, lower.tail = TRUE)

x es un valor de la variable aleatoria normal X; mean especifica la media de la distribución; sd especifica la desviación estándar. Los valores por defecto son media = 0 y sd = 1 que especifican la distribución normal estándar. El argumento lower.tail es por defecto TRUE. Cuando se establece en FALSE, pnorm(x, media = mu, sd = sigma, lower.tail = FALSE) devuelve 1 - pnorm(x, media = mu, sd = sigma) o la cola superior de la distribución[3].

Esta vez la media = 100 y la sd = 16.

```
> pnorm(90, mean = 100, sd = 16)                                  # P(X < 90)

## [1] 0.265986

> pnorm(124, mean = 100, sd = 16, lower.tail = FALSE )            # P(X > 124)

## [1] 0.0668072

> pnorm(104, mean = 100, sd = 16) - pnorm(96, mean = 100, sd = 16) # P(96 < X < 104)

## [1] 0.197413
```

Por último, encuentre x de manera que P (X x) = 0,95. Bueno, R también tiene una función para eso.

qnorm(p, mean = mu, sd = sigma)

p es la probabilidad acumulada particular de la variable aleatoria normal X que estamos tratando de alcanzar; mean especifica la media de la distribución; sd especifica la desviación estándar. Los valores por defecto son media = 0 y sd = 1 que especifican la distribución normal.

```
> qnorm(0.95, mean = 100, sd = 16)

## [1] 126.318
```

Ejemplo 5

La concentración media de colesterol en sangre de una gran población de hombres adultos (50-60 años) es de 200 mg/dl con una desviación estándar de 20 mg/dl. Suponga que las mediciones de colesterol en sangre se distribuyen normalmente. ¿Cuál es la probabilidad de que un individuo seleccionado al azar de este grupo de edad tenga un nivel de colesterol en sangre inferior a 250 mg/dl? [3].

Utilizamos la función pnorm

```
> pnorm(250, mean = 200, sd = 20)

## [1] 0.99379
```

¿Cuál es la probabilidad de que un individuo seleccionado al azar de este grupo de edad tenga un nivel de colesterol superior a 225 mg/dl?

Aplicando pnorm( ) con el argumento opcional lower.tail = FALSE para determinar la cola superior por encima de 225 de la distribución.

```
> pnorm(225, mean = 200, sd = 20, lower.tail = FALSE)

## [1] 0.10565
```

Ejemplo 6

Supongamos que se sabe que una especie particular de plantas del sotobosque tiene una varianza en las alturas de 16 cm^2 (σ^2 = 16 cm^2). Si esta especie se muestrea con las alturas de 25 plantas con un promedio de 15 cm, encuentre el intervalo de confianza del 95% para la media de la población.

Aquí n = 25, $\bar{X}$ = 15 cm, σ^2 = 16 cm^2, y σ = 4 cm. Introduce todo esto en R y calcula el error estándar.

```
> n <- 25                  # sample size
> mx <- 15                 # sample mean
> sd <- 4                  # population sd
> se <- sd/sqrt(n)         # standard error
> se

## [1] 0.8
```

Observe que para determinar un intervalo de confianza se necesitan cuatro números diferentes:

1. Una estimación puntual, aquí la media muestra $\bar{X}$
2. Una medida de variabilidad, aquí el error estándar de la media $\frac{\sigma}{\sqrt{n}}$
3. Un nivel de confianza deseado 1 - α, en este caso 1- α = 0.95, por lo tanto, α = 0.05
4. La distribución de muestreo de la estimación puntual, aquí la distribución normal estándar, que proporciona el factor de confianza b con el que se ajusta la variabilidad para el nivel de confianza deseado, en este caso $F(b) = 1 - \frac{\alpha}{2} = 0.975$.

Utilizando R, los puntos finales del intervalo de confianza tienen la forma

Punto estimado ± (Factor de confianza) (error estándar)

Calculando el factor b mediante la función qnorm( )

```
> qnorm(0.975)                          # with the defaults: mean = 0, sd = 1

## [1] 1.95996

> L1 <- mx - qnorm(0.975)*se            # lower endpt
> L2 <- mx + qnorm(0.975)*se            # upper endpt
> c(L1, L2)                             # print both endpoints at once

## [1] 13.432 16.568
```

Estamos seguros al 95% de que los valores 13,43 y 16,57 capturan la media paramétrica, m. Recalculemos esto una vez más, pensando en qnorm(0,975)*se cómo el término de error completo que sumaremos o restaremos de la media. Esto nos dará un código altamente reutilizable código altamente reutilizable y nos permitirá centrarnos en la idea clave: determinar el error

```
> error <- qnorm(0.975)*se
> L1 <- mx - error        # lower endpt
> L2 <- mx + error        # upper endpt
> c(L1, L2)       # print both endpoints at once

## [1] 13.432 16.568
```

¿Y si queremos estar más seguros (digamos, el 99%) de que hemos incluido m en nuestro intervalo?

Ahora α = 1 - 0,99 = 0,01 por lo que 1- $\frac{\alpha}{2}$ = 0,995 en el término de error. Este es el único cambio necesario.

```
> error <- qnorm(0.995)*se
> L1 <- mx - error
> L2 <- mx + error
> c(L1, L2)

## [1] 12.9393 17.0607
```

Ejemplo 7

Una ecologista forestal, que estudia la regeneración de las comunidades de la selva tropical en los huecos causados por la caída de grandes árboles durante las tormentas, leyó que las plántulas de Dendrocnide excelsa crecerán 1,5 m/año con luz solar directa en dichos huecos. En los huecos de su parcela de estudio identificó 9 ejemplares de esta especie y los midió en 2005 y de nuevo un año después. A continuación, se muestran los cambios de altura de los 9 ejemplares. ¿Apoyan sus datos la afirmación publicada de que las plántulas de esta especie crecerán una media de 1,5 m al año en luz solar directa?

1.9 2.5 1.6 2.0 1.5 2.7 1.9 1.0 2.0

El ecologista busca desviaciones de 1,5 m en cualquier dirección, por lo que se trata de una prueba de dos colas:

$$H_0: \mu_d = 1.5 \frac{m}{año}$$

$$H_a: \mu_d \neq 1.5 \frac{m}{año}$$

Datos en http://waveland.com/Glover-Mitchell/Example06-1.txt.

```
> data.Ex06.1 <- read.table("http://waveland.com/Glover-Mitchell/Example06-1.txt",
+ header = TRUE)
> data.Ex06.1

##   Difference
## 1        1.9
## 2        2.5
## 3        1.6
## 4        2.0
## 5        1.5
## 6        2.7
## 7        1.9
## 8        1.0
## 9        2.0
```

Para realizar una prueba t sobre estos datos, utilice *t.test( )* con mu = 1,5, la hipótesis por defecto de dos lados y el nivel *conf.level* = 0.95 (o α = 0.05).

```
> t.test(data.Ex06.1$Difference, mu = 1.5)

##
##  One Sample t-test
##
## data:  data.Ex06.1$Difference
## t = 2.3534, df = 8, p-value = 0.04643
## alternative hypothesis: true mean is not equal to 1.5
## 95 percent confidence interval:
##  1.50805 2.29195
## sample estimates:
## mean of x
##       1.9
```

La función t.test( ) proporciona dos formas de responder si m difiere significativamente de 1,5 m/año. El valor P es 0,046, que es menor que $\alpha = 0,05$. Así que hay pruebas para rechazar la hipótesis nula a favor de la alternativa de que la media es diferente de 1,5 m/año. De forma equivalente, t.test( ) proporciona un intervalo de confianza para la media m. Si el valor hipotético (aquí esperamos $\mu = 1,5$) cae dentro del intervalo de confianza, entonces se retiene la hipótesis nula. En caso contrario, se rechaza la hipótesis nula. En este ejemplo, el intervalo de confianza del 95 por ciento [1,508, 2,292] no contiene el valor esperado de m. Hay pruebas para rechazar la hipótesis nula a favor de la alternativa de que la media es diferente de 1,5 m/año.

Ejemplo 8

Para comprobar la eficacia de una nuevo Spray para el control de los ácaros de la roya en los huertos, un investigador desea comparar el rendimiento medio de las arboledas tratadas con la media mostrada en las arboledas no tratadas en años anteriores. Se eligió una muestra aleatoria de 16 huertos de un acre y se roció según un programa recomendado. El rendimiento medio de esta muestra de 16 arboledas fue de 814 cajas de fruta comercializable con una desviación estándar de 40 cajas. Los rendimientos de las arboledas de un acre en la misma zona, sin la pulverización para el control del ácaro de la roya, han sido de una media de 800 cajas en los últimos 10 años. ¿Presentan los datos pruebas que indiquen que el rendimiento medio es suficientemente mayor en las arboledas rociadas que en las no rociadas?

Nota:

t.test2(mx, sx, nx, mu = 0, alternative = "two.sided", conf.level = 0.95)

Detalles: mx, sx y nx son la media, la desviación estándar y el tamaño de la muestra para la muestra x. El argumento opcional mu es el valor esperado de la media. El valor por defecto es mu = 0. El argumento opcional alternative es la hipótesis alternativa. El valor por defecto es "dos caras", siendo las otras opciones posibles "menor" o "mayor". Por último, el argumento conf.level da el nivel de confianza que se utilizará en la prueba t; el valor por defecto es 0,95, que equivale a a = 0,05. Hay argumentos adicionales que se analizan en el siguiente capítulo.
Puede descargar esta función utilizando el comando source( ). La función sólo necesita descargarse una vez por sesión

> source (https://waveland.com/Glover-Mitchell/t.test2.txt)

Anticipar un cambio concreto, un aumento del rendimiento, significa que hay que realizar una prueba de cola derecha (alternative = "greater"):

$$H_0: \mu \leq 800\ cajas$$

$$H_a: \mu > 800\ cajas$$

Sin los datos originales, utilice t.test2( ) se utiliza media muestral mx = 814, una desviación estándar muestral sx = 40, un tamaño de muestra nx = 16 y una media hipotética mu = 800.

```
> t.test2(mx = 814, sx = 40, nx = 16, mu = 800, alternative = "greater")

##
##  One Sample t-test
##
## data:  mx = 814, sx = 40, and nx = 16
## t = 1.4, df = 15, p-value = 0.09093
## alternative hypothesis: true mean is greater than 800
## 95 percent confidence interval:
##  796.469     Inf
## sample estimates:
## mean of x
##       814
```

El valor P de 0,09093 es mayor que α = 0,05 lo que indica que no se puede rechazar la hipótesis nula. De forma equivalente, el intervalo de confianza del 95 por ciento para μ es [796.469, ∞). Este intervalo contiene la media hipotetizada de 800 cajas, por lo que no hay pruebas suficientes de que el rendimiento medio sea significativamente mayor en las arboledas rociadas que en las no rociadas.

Ejemplo 9

Entre los pocos linajes de reptiles que surgieron de la extinción del Mesozoico se encuentran los actuales. Uno de ellos, el tuátara, Sphenodon punctatum, de Nueva Zelanda, es el único superviviente de un grupo que, por lo demás, desapareció hace 100 millones de años. A continuación se indican las masas (en g) de muestras aleatorias de tuatara macho adulto procedentes de dos islotes del Estrecho de Cook. ¿Es la variabilidad de la masa de los machos adultos es diferente en los dos islotes?

Locación A		Locación B
510	790	650
773	440	600
836	435	600
505	815	575
765	460	452

780	690	320
235		660

Puede formularse como el siguiente par de hipótesis

$$H_0: \sigma_A^2 = \sigma_B^2$$

$$H_a: \sigma_A^2 \neq \sigma_B^2$$

Datos disponibles: http://waveland.com/Glover-Mitchell/Example07-1.txt

```
> data.Ex07.1 <- read.table("http://waveland.com/Glover-Mitchell/Example07-1.txt",
+ header = TRUE)
> data.Ex07.1

##      A   B
## 1  510 650
## 2  773 600
## 3  836 600
## 4  505 575
```

```
## 5  765 452
## 6  780 320
## 7  235 660
## 8  790  NA
## 9  440  NA
## 10 435  NA
## 11 815  NA
## 12 460  NA
## 13 690  NA
```

El término NA, "no disponible", indica que no hay datos en ese momento en la tabla. (Las entradas en blanco no se permiten en los marcos de datos; cada fila debe tener el mismo número de entradas). Para realizar una prueba F con estos datos, utilice var.test( ) con los valores predeterminados ratio = 1 alternative = two.sided, y conf.level = 0.95 o α = 0.05.

```
> var.test(data.Ex07.1$A, data.Ex07.1$B)

##
##  F test to compare two variances
##
## data:  data.Ex07.1$A and data.Ex07.1$B
## F = 2.5173, num df = 12, denom df = 6, p-value = 0.2661
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.469106 9.385360
## sample estimates:
## ratio of variances
##            2.51734
```

La función var.test( ) proporciona dos formas de responder si las varianzas de las poblaciones son significativamente diferentes. El valor P es 0,226, que es mayor que $\alpha = 0,05$, por lo que no hay pruebas suficientes para rechazar la hipótesis nula de que las varianzas son iguales. De forma equivalente, var.test( ) proporciona un intervalo de confianza para el valor de F. Como es habitual, si el valor hipotetizado (aquí esperamos que el cociente F sea 1) cae dentro del intervalo de confianza, entonces se retiene la hipótesis nula. En caso contrario, se rechaza la hipótesis nula. En este ejemplo, el intervalo de confianza del 95 por ciento [0,469, 9,385] contiene el valor esperado de F de 1. Así que la hipótesis nula no se rechaza. Continúe asumiendo que las varianzas son iguales.

Referencias

[1] B. Shahbaba, *Biostatistics with R*. New York, NY: Springer New York, 2012. doi: 10.1007/978-1-4614-1302-8.

[2] K. Seefeld, M. Ed, y E. Linder, “Statistics Using R with Biological Examples”, p. 325, 2007.

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