universidad de león

TEMA 8

Inferencia estadística

Métodos Numéricos y Estadísticos
Grado en Ingeniería Mecánica

Curso 2021-2022

Eva María Mazcuñán Navarro

Eva María Mazcuñán Navarro
Departamento de Matemáticas
Universidad de León
E-mail: emmazn@unileon.es

Contenidos

Introducción

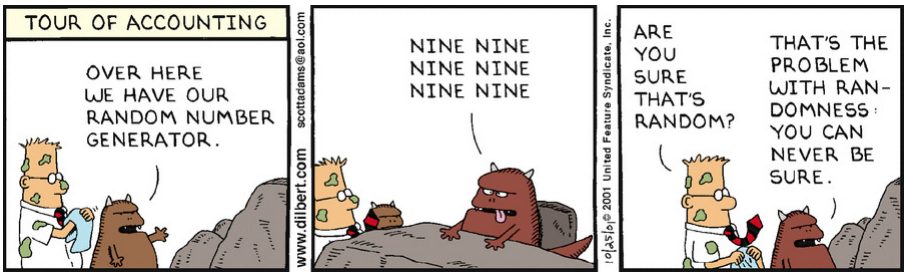


Figura 1: Viñeta de la tira cómica *Dilbert*, de Scott Adams.

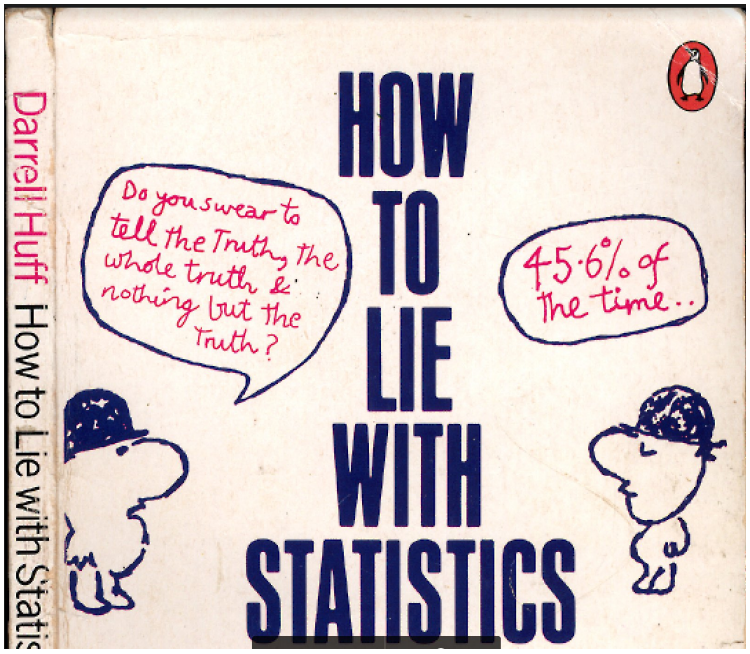


Figura 2: Portada del libro *How to Lie with Statistics*, de Darrell Huff.

Hasta ahora hemos realizado un estudio teórico de las variables aleatorias. Y hemos aprendido a resolver problemas como el siguiente:

> Sabemos que, para un determinado jugador de baloncesto, la probabilidad con que acierta un lanzamiento de tiro libre es 0.75.
>
> Calcula la probabilidad de que, en una tanda de 50 lanzamientos, acierte como máximo 30 de ellos.

Para resolver el problema anterior, consideramos la variable aleatoria

$$X = \text{número de lanzamientos acertados.}$$

Identificamos que X sigue una distribución binomial de parámetros $n = 50$ y $p = 0.75$, abreviadamente

$$X \sim Binom(50, 0.75),$$

y calculamos la probabilidad pedida como

$$P(X \leq 30) = F(30) = \texttt{pbinom(30,50,0.75)} = 0.0139.$$

Ahora, en una situación real, lo más frecuente es que no conozcamos el valor $p = 0.75$ para la probabilidad de acierto del jugador en el lanzamiento de tiros libres. Porque este valor p es un parámetro teórico, que interpretamos como la proporción de aciertos del jugador cuando el número de lanzamientos tiende a infinito. Para obtener el valor de p deberíamos observar los infinitos potenciales lanzamientos que podría efectuar el lanzador, situación imposible en la práctica. En consecuencia, el tipo de problema que se plantea en la práctica es más parecido al siguiente:

Problema 1

Nos planteamos investigar el valor del parámetro

$p =$ probabilidad de que el jugador acierte un tiro libre.

Para ello, observamos el resultado de 50 lanzamientos efectuados por el jugador en cuestión, anotando en cada caso si acierta o falla. Se obtienen 28 aciertos y 22 fallos.

El objetivo es realizar inferencias sobre el parámetro p basándonos en los resultados de los lanzamientos observados. Queremos en concreto responder a las siguientes cuestiones:

- El jugador afirma que su probabilidad de acertar un tiro libre es 0.75 ¿resultan los datos compatibles con esta afirmación?
- A la vista de los datos ¿cuál sería un rango de valores plausibles para la probabilidad p?

En este tema estudiaremos los procedimientos de inferencia estadística que permitirán responder a las preguntas anteriores. Responderemos a la primera pregunta utilizando un procedimiento de inferencia estadística denominado **contraste de hipótesis** (sección 2) y abordaremos la segunda cuestión construyendo un **intervalo de confianza** (sección 3).

En general, un procedimiento de inferencia estadística es un método o técnica que permite inferir, a partir de la información empírica proporcionada por una muestra (en nuestro ejemplo los aciertos observados en 50 lanzamientos), el comportamiento de una determinada población (en nuestro ejemplo la probabilidad de acierto p). Las conclusiones de un procedimiento de inferencia estadística nunca serán afirmaciones categóricas, sino que estarán sujetas a un riesgo de error, que se cuantifica en términos de probabilidades (ver figuras 1 y 2).

Presentaremos el comando de R `binom.test()` para realizar el contraste de hipótesis y obtener el intervalo de confianza para la proporción p que responderán a las preguntas planteadas antes. Basándonos en nuestro conocimiento de las variables binomiales estudiadas en el tema anterior, seremos capaces de reproducir los cálculos que realiza este comando. De esta forma, podremos entender los fundamentos generales de los contrastes de hipótesis y los intervalos de confianza.

En la última práctica de ordenador del curso, estudiaremos más comandos de R para realizar contrastes de hipótesis y calcular intervalos de confianza para otros parámetros, como la media de una variable continua. Gracias a lo aprendido en este tema, sabremos interpretar las salidas de dichos comandos, aún sin conocer el detalle de los cálculos realizados para obtenerlas.

1. Estadística descriptiva

Todo estudio estadístico comienza con una exploración inicial de los datos obtenidos, que habitualmente incluye el cálculo determinados valores de interés que resumen o describen los datos, y la visualización de los datos mediante los gráficos adecuados en cada caso.

Para nuestro ejemplo calcularemos la llamada proporción muestral y realizaremos un diagrama de barras.

Comenzamos cargando el paquete `tidyverse` para tener acceso a todas las funciones de R que usaremos a lo largo del tema.

```
library("tidyverse")
```

1.1. Estimación puntual

Supongamos, como se indicó en el problema planteado en la introducción, que en los 50 lanzamientos observados, el jugador acertó 28 de ellos (y falló los 22

restantes).

Recordemos que nuestro objetivo es estimar el valor teórico de la probabilidad o proporción de aciertos p del jugador. Una forma natural de hacerlo es calcular la proporción de aciertos en nuestra muestra de 50 lanzamientos:

$$\hat{p} = \frac{28}{50} = 0.56.$$

Este valor se denomina **proporción muestral** y se denota $\hat{p}$, porque nos da una idea aproximada de cuál puede ser el verdadero valor del valor teórico p, que se llama **proporción poblacional**.

Decimos que la proporción muestral $\hat{p} = 0.56$ es una **estimación puntual** de la proporción poblacional p. En general, una estimación puntual de un parámetro es un valor numérico calculado a partir de la muestra que nos orienta sobre el verdadero valor del parámetro.

Hay que tener claro que la proporción muestral $\hat{p} = 0.56$ nos proporciona una estimación puntual de la proporción poblacional p, pero que el verdadero valor de p sigue siendo desconocido. El verdadero valor de p es la probabilidad de que el jugador acierte un tiro libre, que identificamos con la proporción de aciertos en el conjunto ideal de los infinitos lanzamientos que podría efectuar el jugador. Por su parte, el valor $\hat{p} = 0.56$ es tan sólo la proporción de aciertos en los 50 lanzamientos que hemos observado. Podría ser que los 50 lanzamientos observados sean una "mala racha", y que el verdadero valor de la proporción de aciertos p sea superior a $\hat{p} = 0.56$.

1.2. Gráficos

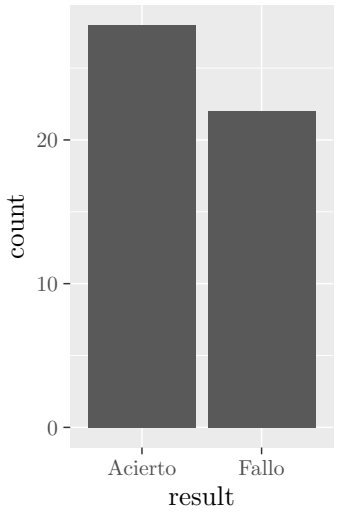
Ahora vamos a representar gráficamente los datos mediante un diagrama de barras.

En primer lugar creamos una tabla con el recuento de aciertos y fallos:

```
basket <- tibble(
  result = c("Acierto", "Fallo"),
  count = c(28, 22)
)
```

Y usamos la tabla creada para crear un diagrama de barras con la función `geom_col()`:

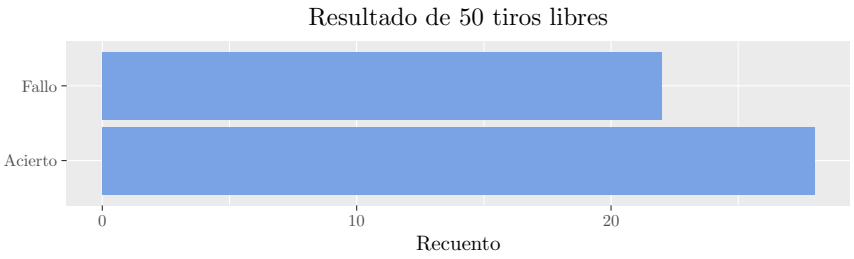
```
ggplot(
  data = basket,
  mapping = aes(
    x = result,
    y = count
  )
) +
  geom_col()
```



Mejoramos un poco el aspecto inicial de nuestro gráfico de barras añadiendo rótulos, cambiando el color de las barras, y girándolo con `coord_flip()`:

```
ggplot(
  data = basket,
  mapping = aes(
    x = result,
    y = count
  )
) +
  geom_col(
    fill = "#7AA3E5"
  ) +
  coord_flip() +
  labs(
    title = "Resultado de 50 tiros libres",
    x = "",
    y = "Recuento"
```

```
) +
theme(plot.title = element_text(hjust = 0.5)) # centrar
título
```



2. Contrastes de hipótesis

Hasta ahora hemos descrito los datos numérica y gráficamente. Concretamente, hemos calculado la proporción de aciertos en la muestra, obteniendo el valor $\widehat{p} = 0.56$ para la proporción muestral, como estimación de la proporción poblacional p, y hemos realizado un diagrama de barras de los recuentos de aciertos y fallos.

Nos planteamos ahora responder a la primera cuestión planteada en el problema de la introducción:

> Supongamos que el jugador afirma que su probabilidad de acertar un tiro libre es 0.75. El objetivo es contrastrar si los datos observados resultan compatibles con la afirmación del jugador.

Con la exploración inicial que hemos hecho de los datos, podemos concretar el objetivo planteando lo que en estadística se denomina un **contraste de hipótesis**. Dividiremos el procedimiento en tres pasos: planteamiento de las hipótesis, cálculo del p-valor y conclusión.

2.1. Planteamiento de las hipótesis

Según el jugador la proporción verdadera es $p = 0.75$. Y la estimación de p según nuestra muestra es $\widehat{p} = 0.56$, inferior al valor de p según el jugador.

Como resaltamos antes, no podemos concluir con seguridad, basándonos en los datos obtenidos en la muestra de 50 lanzamientos, que el jugador mienta. Puede que el jugador diga la verdad, y que, por azar, hayamos observado una racha peor de lo habitual.

Pero nuestros datos levantan la sospecha de que el jugador exagera, y nos sugieren que $p < 0.75$. Esta hipótesis sugerida por los datos se denomina **hipótesis alternativa**, y se denota H_1.

Por otra parte, a la igualdad $p = 0.75$, se le llama **hipótesis nula** y se denota H_0.

Resumimos el contraste de hipótesis que planteamos escribiendo

$$\begin{cases} H_0: & p & = & 0.75 \\ H_1: & p & < & 0.75. \end{cases}$$

2.2. Cálculo del p-valor

Vamos a resolver el conflicto, entre la hipótesis nula H_0, y la hipótesis alternativa H_1 que sugieren los datos, contestando a la pregunta siguiente:

> Si la realidad fuera que $p = 0.75$ (la hipótesis nula es cierta, el jugador dice la verdad) ¿cuál es la probabilidad de observar una racha tan mala como la observada o peor?

Respondiendo a esta pregunta, cuantificaremos la sospecha que han levantado nuestros datos y que provocó la formulación de la hipótesis alternativa.

Si la respuesta es que la probabilidad es muy baja, tomaremos la decisión de acusar de mentiroso al jugador y nos decantaremos por la hipótesis alternativa de que $p < 0.75$. Si por el contrario la probabilidad no es muy baja, lo dejaremos estar y no acusaremos al jugador de mentiroso, pensaremos que los datos observados pueden deberse al azar y no a que el jugador mienta.

Adoptaremos una actitud conservadora, y fijaremos el umbral de lo que consideramos una probabilidad muy baja en 0.05. De esta manera, si los datos observados tienen una probabilidad superior a 0.05, no acusaremos de mentiroso al jugador. Esta probabilidad se llama **nivel de significación** del contraste y se denota α. Lo más frecuente en el ámbito científico es trabajar con este nivel de significación $\alpha = 0.05$, pero también pueden convenirse otros valores como $\alpha = 0.01$.

El procedimiento estadístico con el que vamos a resolver el contraste de hipótesis que hemos planteado y a calcular la probabilidad que buscamos se llama **test binomial exacto**. El comando de R que lo lleva a cabo es `binom.test()`. Concretamente para resolver nuestro contraste tenemos que ejecutar:

```r
binom.test(
  x = 28,
  n = 50,
  p = 0.75,
  alternative = "less"
)
```

```
##
##  Exact binomial test
##
## data:  28 and 50
## number of successes = 28, number of trials = 50, p-value =
0.002618
## alternative hypothesis: true probability of success is less
than 0.75
## 95 percent confidence interval:
##  0.0000000 0.6801741
## sample estimates:
## probability of success
##                   0.56
```

Los argumentos `x = 28` y `n = 50` indican que hemos observado 28 aciertos en 50 lanzamientos. Con el argumento `p = 0.75` indicamos que queremos comparar nuestros datos con el valor de referencia 0.75 de la hipótesis nula. Y el argumento `alternative = "less"` indica que nuestra hipótesis alternativa es que p es menor que el valor umbral 0.75.

La probabilidad que buscamos la encontramos en el fragmento de la salida

`p-value = 0.002618`.

Esta probabilidad 0.002618 asociada a nuestro contraste de hipótesis se denomina **p-valor**. La letra p en p-valor es inicial de probabilidad, no tiene que ver con que estemos realizando un contraste de hipótesis sobre el parámetro p (en un contraste de hipótesis para una media también hablaríamos del p-valor de ese contraste).

En el apartado siguiente, utilizaremos el p-valor para decidir entre las hipótesis contrastadas, esto es, entre la hipótesis nula $p = 0.75$ y la hipótesis alternativa sugerida por los datos $p < 0.75$. Para entender el criterio que seguiremos, vamos a analizar cómo se calcula el p-valor que hemos obtenido con `binom.test()`.

El p-valor 0.002618 que hemos obtenido con la función `binom.test()` cuantifica

nuestra sospecha de que $p < 0.75$, en el sentido que se explica a continuación: Si fuera verdad que $p = 0.75$, la probabilidad de observar en 50 lanzamientos una racha tan mala o peor como la de nuestra muestra sería de 0.002618.

Dicho con otras palabras: si repetimos un gran número de veces el experimento (observar otros 50 lanzamientos y contabilizar los aciertos), en aproximadamente 26 de cada 10000 repeticiones observaríamos un resultado tan desfavorable como acertar tan sólo 28 de los 50 lanzamientos.

Con las explicaciones anteriores, que describen el significado del p-valor, y la teoría que conocemos de variables aleatorias, tenemos herramientas suficientes para saber cómo se ha calculado el p-valor 0.002618: se obtiene como

Reproduce el cálculo del p-valor 0.002618 como la probabilidad de un suceso asociado a la variable aleatoria

$$X = \text{Número de lanzamientos acertados entre los 50 observados.}$$

Sigue los siguientes pasos:

- Piensa cuál es el suceso asociado a X cuya probabilidad coincide con p-valor, según el significado del p-valor que se ha explicado antes.
- Calcula la probabilidad del suceso en cuestión utilizando los comandos de R que hemos estudiado en el Tema 6. Verifica que coincide con el p-valor para confirmar que has acertado.
- Basándote en el significado del p-valor, escribe un breve párrafo explicando por qué coincide con la probabilidad del suceso propuesto.

2.3. Conclusión

Continuará

3. Intervalos de confianza

Continuará ...

Problema *tipo examen*

Continuará ...

Bibliografía

Para ampliar conocimientos de este tema puedes consultar los Capítulos 6, 7, 8 y 9 de [1] y los Capítulos 6, 7, 8 y 9 de [2].

[1] Janet Susan Milton y Jesse C. Arnold. *Introduction to probability and statistics: principles and applications for engineering and the computing sciences*. 2003. ISBN: 007246836X.

[2] Douglas C. Montgomery y George C. Runger. *Applied Statistics and Probability for Engineers, 5th Edition*. Wiley, 2010, pág. 784. ISBN: 1118050177.