# Practica 2

Gonzalo Cruz Gómez y Daniel Fernández Magariños

2024-11-26

## Introducción

### Ejercicio 1

Demostrar que la log-verosimilitud de p basada en n observaciones es:

$$x \sim \mathrm{Bin}(n, p)$$

$$L(x) = \binom{n}{r} p^r (1-p)^{n-r}$$

$$\ell(p|r) = \log \binom{n}{r} + r\log(p) + (n-r)\log(1-p)$$

Como $\binom{n}{r}$ no depende de p, podemos quitarlo:

$$r \log\left(\frac{p}{1-p}\right) + n\log(1-p)$$

### Ejercicio 2

$$\bar{I}(p) = \mathbb{E}\left[\left(\frac{\partial^2}{\partial p^2} \log L(p)\right)\right]$$

$$\frac{\partial \ell}{\partial p} = \frac{r}{p} + \frac{r-n}{1-p}, \quad \frac{\partial^2}{\partial p^2} = -\frac{r}{p^2} - \frac{n-r}{(1-p)^2}.$$

$$\mathbb{E}\left[-\frac{\partial^2}{\partial p^2}\right] = \mathbb{E}\left[\frac{r}{p^2}\right] + \mathbb{E}\left[\frac{n-r}{(1-p)^2}\right] =$$

$$= \frac{n}{p} - \frac{np}{(1-p)^2} + \frac{n}{(1-p)^2} = \frac{n(1-p)^2 - np^2 + np}{p(1-p)^2} =$$

$$\frac{n(1-2p+p^2) - np^2 + np}{p(1-p)^2} = \frac{n-np}{p(1-p)^2} = \frac{n(1-p)}{p(1-p)^2} = \frac{n}{p(1-p)}$$

$$\bar{I}(p) = \frac{n}{p(1-p)}$$

**Ejercicio 3**

**Parametrización 1:**

$$\theta = \log \frac{p(1-p_0)}{p_0(1-p)};$$

$$e^{\theta} = \frac{p(1-p_0)}{p_0(1-p)};$$

$$p_0(1-p)e^{\theta} = p(1-p_0);$$

$$p_0e^{\theta} - p_0pe^{\theta} = p - pp_0;$$

$$p_0e^{\theta} = p(1-p_0+p_0e^{\theta});$$

$$p = \frac{p_0e^{\theta}}{1-p_0+p_0e^{\theta}}$$

Cambiamos p en la función de log-verosimilitud:

$$l(\theta) = r\log(\frac{p_oe^{\theta}}{1-p_0+p_0e^{\theta}}) - (r-n)\log(1-\frac{p_oe^{\theta}}{1-p_0+p_0e^{\theta}}) = r\left[\log(p_0e^{\theta}) - \log(1-p_0+p_0e^{\theta})\right] -$$

$$-(r-n)\log(\frac{1-p_0}{1-p_0+p_0e^{\theta}}) = r\left[\log(p_0e^{\theta}) - \log(1-p_0+p_0e^{\theta})\right] - (r-n)\left[\log(1-p_0) - \log(1-p_0+p_0e^{\theta})\right]$$

Derivamos la función respecto de theta:

$$\frac{\partial l(\theta)}{\partial\theta} = r\left(\frac{p_0e^{\theta}}{p_0e^{\theta}} - \frac{p_0e^{\theta}}{1-p_0+p_0e^{\theta}}\right) - (r-n)\left(-\frac{p_0e^{\theta}}{1-p_0+p_0e^{\theta}}\right) = r\left(1-\frac{p_0e^{\theta}}{1-p_0+p_0e^{\theta}}\right) + (r-n)\left(\frac{p_0e^{\theta}}{1-p_0+p_0e^{\theta}}\right) =$$

$$= r - n\left(\frac{p_0e^{\theta}}{1-p_0+p_0e^{\theta}}\right)$$

Ahora sustituimos theta por 0 en la derivada ya así obtenemos Z:

$$Z = \frac{\partial l(0)}{\partial\theta} = r - np_0$$

Para calcular V ahora tendremos que hacer la segunda derivada de la función de log-verosimilitud respecto de theta:

$$\frac{\partial^2 l(\theta)}{\partial\theta^2} = -\frac{np_0e^{\theta}(1-p_0+p_0e^{\theta}) - np_0^2e^2\theta}{1-p_0+p_oe^{\theta}}$$

Sustituimos $\theta = 0$:

$$\frac{\partial^2 l(0)}{\partial\theta^2} = -np_0(1-p_0) = (p_0-1)np_0$$

El siguiente paso es calcular la esperanza de la expresión anterior y cambiarle el signo, así obtendremos V:

$$V = -E\left[\frac{\partial^2 l(0)}{\partial\theta^2}\right] = -E\left[(p_0-1)np_0\right];$$

por las propiedades de la esperanza, podemos sacar las constantes que multiplican:

$$V = np_0(1-p_0)$$

**Parametrización 2:**

$$\theta = p - p_0; p = \theta + p_0$$

Sustituimos la expresión de p en la función de log-verosimilitud:

$$l(\theta) = r\left[\log(\theta + p_0) - \log(1 - \theta - p_0)\right] + n\log(1 - \theta - p_0)$$

Derivamos la expresión anterior respecto de theta:

$$\frac{\partial l(\theta)}{\partial \theta} = r\left(\frac{1}{\theta + p_0} + \frac{1}{1 - \theta - p_0}\right) - \frac{n}{1 - \theta - p_0}$$

Sustituimos theta por 0:

$$Z = \frac{\partial l(0)}{\partial \theta} = \frac{r}{p_0} + \frac{r}{1 - p_0} - \frac{n}{1 - p_0} = \frac{r}{p_0} + \frac{r - n}{1 - p_0} = \frac{r(1 - p_0) + (r - n)p_0}{p_0(1 - p_0)} = \frac{r - np_0}{p_0(1 - p_0)}$$

Para calcular ahora la V, derivamos por segunda vez la función de log-verosimilitud respecto de theta:

$$\frac{\partial^2 l(\theta)}{\partial \theta^2} = r\left(\frac{-1}{(\theta + p_0)^2} + \frac{1}{(1 - \theta - p_0)^2}\right) - \frac{n}{(1 - \theta - p_0)^2} = \frac{-r}{(\theta + p_0)^2} + \frac{r - n}{(1 - \theta - p_0)^2}$$

Sustituimos theta respecto de 0:

$$\frac{\partial^2 l(0)}{\partial \theta^2} = \frac{-r}{p_0^2} + \frac{r - n}{(1 - p_0)^2}$$

El último paso es calcular la esperanza y cambiar el signo a la anterior expresión, como son todo constantes menos r, entonces calcular la esperanza de la anterior expresión es sustituir $r = np_0$:

$$V = -E\left[\frac{\partial^2 l(0)}{\partial \theta^2}\right] = \frac{np_0}{p_0^2} - \frac{n(1 - p_0)}{(1 - p_0)^2} = \frac{n}{p_0} - \frac{n}{1 - p_0} = \frac{n}{p_0(1 - p_0)}$$

**Parametrización 3:**

$$\theta = \arcsin(\sqrt{p}) - \arcsin(\sqrt{p_0})$$

$$\theta + \arcsin(\sqrt{p_0}) = \arcsin(\sqrt{p})$$

$$p = \sin^2(\theta + \arcsin(\sqrt{p_0}))$$

Sustituimos p en la fórmula de log-verosimilitud:

$$l(\theta) = r\log\left(\frac{\sin^2(\theta + \arcsin(\sqrt{p_0}))}{1 - \sin^2(\theta + \arcsin(\sqrt{p_0}))}\right) - n\log\left(1 - \sin^2(\theta + \arcsin(\sqrt{p_0}))\right) =$$

$$= r\log\left(\frac{\sin^2(\theta + \arcsin(\sqrt{p_0}))}{\cos^2(\theta + \arcsin(\sqrt{p_0}))}\right) - n\log\left(\cos^2(\theta + \arcsin(\sqrt{p_0}))\right) =$$

$$= r\log\left(\tan(\theta + \arcsin(\sqrt{p_0})\right) - n\log\left(\cos^2(\theta + \arcsin(\sqrt{p_0}))\right) = \; = 2r\log\left(\tan(\theta + \arcsin(\sqrt{p_0}))\right) - 2n\log\left(\cos(\theta + \arcsin(\sqrt{p_0}))\right)$$

Calculamos la derivada parcial respecto de $\theta$ de la función:

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{2r\sec^2\left(\theta + \arcsin(\sqrt{p_0})\right)}{\tan\left(\theta + \arcsin(\sqrt{(p_0)})\right)} - 2n\frac{\sin\left(\theta + \arcsin(\sqrt{p_0})\right)}{\cos\left(\theta + \arcsin(\sqrt{p_0})\right)} =$$

$$= \frac{2r\sec^2\left(\theta + \arcsin(\sqrt{p_0})\right)}{\tan\left(\theta + \arcsin(\sqrt{(p_0)})\right)} - 2n\tan\left(\theta + \arcsin(\sqrt{p_0})\right)$$

Sustituimos $\theta = 0$:

$$\frac{\partial l(0)}{\partial \theta} = \frac{2r \sec^2\left(\arcsin(\sqrt{p_0})\right)}{\tan\left(\arcsin(\sqrt{(p_0)})\right)} - 2n \tan\left(\arcsin(\sqrt{p_0})\right) =$$

$$= 2r \frac{1 + \tan^2\left(\arcsin(\sqrt{p_0})\right)}{\tan\left(\arcsin(\sqrt{(p_0)})\right)} - 2n \tan\left(\arcsin(\sqrt{p_0})\right)$$

Sustituyendo $\tan\left(\arcsin(\sqrt{p_0})\right) = \frac{\sqrt{p_0}}{\sqrt{1-p_0}}$ en la fórmula anterior:

$$Z = 2r \frac{1}{\sqrt{p_0(1-p_0)}} - 2n \frac{\sqrt{p_0}}{\sqrt{1-p_0}} = \frac{2}{\sqrt{1-p_0}} \left( \frac{r}{\sqrt{p_0}} - n\sqrt{p_0} \right)$$

Para calcular V ahora, hacemos la derivada de p respecto de $\theta$:

$$\frac{\partial p}{\partial \theta} = \sin\left(2\theta + 2\arcsin(\sqrt{p_0})\right)$$

$$V = \bar{I}(p) \left( \frac{\partial p}{\partial \theta} \right)^2 = \frac{n}{p(1-p)} \sin^2\left(2\theta + 2\arcsin(\sqrt{p_0})\right)$$

Usamos la propiedad trigonométrica de $\sin^2(2a) = 4\sin^2(a)\cos^2(a)$, y sustituimos $p = \sin^2\left(\theta + \arcsin(\sqrt{p_0})\right)$:

$$V = \frac{n 4 \sin^2\left(\theta + \arcsin(\sqrt{p_0})\right) \cos^2\left(\theta + \arcsin(\sqrt{p_0})\right)}{\sin^2\left(\theta + \arcsin(\sqrt{p_0})\right) \cos^2\left(\theta + \arcsin(\sqrt{p_0})\right)} = 4n$$

**Resumen de resultados:**

- Parametrización 1:
$$Z = r - np_0$$
$$V = np_0(1-p_0)$$
- Parametrización 2:
$$Z = \frac{r - np_0}{p_0(1-p_0)}$$
$$V = \frac{n}{p_0(1-p_0)}$$
- Parametrización 3:
$$Z = \frac{2}{\sqrt{1-p_0}} \left( \frac{r}{\sqrt{p_0}} - n\sqrt{p_0} \right)$$
$$V = 4n$$

## Contraste de hipótesis

### Ejercicio 4

```
#valores de prueba
p0 <- 0.3
n <- 1000
r <- 400
```

```r
calculate <- function(p0, n, r, parametrizacion){
  if(parametrizacion == 1){
    Z <- r - n*p0
    V <- n*p0*(1-p0)
  }
  else if (parametrizacion == 2){
    Z <- (r-n*p0)/(p0*(1-p0))
    V <- n/(p0*(1-p0))
  }
  else if (parametrizacion == 3) {
    Z <- (2 / sqrt(1 - p0)) * (r / sqrt(p0) - n * sqrt(p0))
    V <- 4*n

}

  return (c(Z, V))
}

result1 <- calculate(p0, n, r, parametrizacion = 1)
result2 <- calculate(p0, n, r, parametrizacion = 2)
result3 <- calculate(p0, n, r, parametrizacion = 3)

cat("Resultados para parametrización 1:\n")
```

```
## Resultados para parametrización 1:
```

```r
print (result1)
```

```
## [1] 100 210
```

```r
cat("Resultados para parametrización 2:\n")
```

```
## Resultados para parametrización 2:
```

```r
print (result2)
```

```
## [1]  476.1905 4761.9048
```

```r
cat("Resultados para parametrización 3:\n")
```

```
## Resultados para parametrización 3:
```

```r
print (result3)
```

```
## [1]  436.4358 4000.0000
```

### Ejercicio 5

Para hacer el contraste de hipótesis de que $p = p_0$ donde, $p_0 = 0.3$, es equivalente a hacer el contraste respecto a $\theta$ donde: $H_0 : \theta = 0$ $H_1 : \theta > 0$

```r
# Ponemos los datos necesarios para el contraste
n = 1000 # Tamaño de muestra
p0 = 0.3 # Probabilidad que asumimos en la hipótesis nula
p = 0.4 # Probabilidad para simular los datos y hacer el contraste, tiene que ser mayor que 0.3

set.seed(14389)
```

```r
sim <- rbinom(n, size= 1, prob= p) # Simular la binomial con parámetro p
r = sum(sim) # Calculamos r (número de éxitos), necesario para los cálculos de Z y V
```

- Primera parametrización:

```r
# Calculamos Z y V para la primera parametrización:
par1 = calculate(p0, n, r, 1)
Z_1 = par1[1]
V_1 = par1[2]
```

Como $Z \sim N(\theta V, V)$ entonces la hipótesis nula la distribución de Z es $N(0, V)$:

```r
z_1 = Z_1 / (sqrt(V_1))
p_valor1 = pnorm(z_1, 0, 1, lower.tail = FALSE)
cat(p_valor1)
```

```
## 1.588515e-12
```

Como $p - valor \approx 0$ rechazamos $H_0$ a favor de la $H_1$.

- Segunda parametrización:

```r
# Calculamos Z y V para la segunda parametrización:
par2 = calculate(p0, n, r, 2)
Z_2 = par2[1]
V_2 = par2[2]

z_2 = Z_2 / (sqrt(V_2))
p_valor2 = pnorm(z_2, 0, 1, lower.tail = FALSE)
cat(p_valor2)
```

```
## 1.588515e-12
```

Como $p - valor \approx 0$ rechazamos $H_0$ a favor de la $H_1$.

- Tercera parametrización:

```r
# Calculamos Z y V para la tercera parametrización:
par3 = calculate(p0, n, r, 3)
Z_3 = par3[1]
V_3 = par3[2]

z_3 = Z_3 / (sqrt(V_3))
p_valor3 = pnorm(z_3, 0, 1, lower.tail = FALSE)
cat(p_valor3)
```

```
## 1.588515e-12
```

Como $p - valor \approx 0$ rechazamos $H_0$ a favor de la $H_1$.

## Cálculo del tamaño muestral

**Ejercicio 6**

$$V_1 = n\, p_0\, (1 - p_0)$$

$$V_2 = \frac{n}{p_0(1 - p_0)}$$

$$V_3 = 4n$$

$$V = \frac{(z_\alpha + z_\beta)^2}{\theta_R}$$

Primera parametrización:

$$n = \frac{(z_\alpha + z_\beta)^2}{\theta_R^2 {\cdot} p_0(1-p_0)} = \frac{(z_\alpha + z_\beta)^2}{\log\left(\frac{p(1-p_0)}{p_0(1-p)}\right)^2\, p_0\,(1-p_0)}$$

Segunda parametrización:

$$n = \frac{(z_\alpha + z_\beta)^2 p_0(1-p_0)}{(p-p_0)^2}$$

Tercera parametrización:

$$n = \frac{(z_\alpha + z_\beta)^2}{4\theta^2} = \frac{(z_\alpha + z_\beta)^2}{4(\arcsin\sqrt{p} - \arcsin\sqrt{p}_0)^2}$$

## Potencia y error de tipo I

**Ejercicio 7**

```
# Parámetros enunciado
p0 <- 0.003
p <- 0.006
alpha <- 0.025
beta <- 0.20

# Percentiles z_alpha/2 y z_beta
z_alpha <- qnorm(1 - alpha/2)
z_beta <- qnorm(1 - beta)

#   = log((p(1-p0)) / (p0(1-p)))
n1 <- (z_alpha + z_beta)^2 / (log((p * (1 - p0)) / (p0 * (1 - p)))^2 * p0 * (1 - p0))

#   = p - p0
n2 <- (z_alpha + z_beta)^2 * p0 * (1 - p0) / ((p-p0)^2)

#   = arcsin(sqrt(p)) - arcsin(sqrt(p0))
n3 <- (z_alpha + z_beta)^2 / (4*(asin(sqrt(p))-asin(sqrt(p0)))^2)

# usamos ceiling(n) para redondear al siguiente entero
cat("Tamaño muestral para cada parametrización:\n")
```

```
## Tamaño muestral para cada parametrización:
```

```
cat("  = log((p(1-p0)) / (p0(1-p))): n =", ceiling(n1), "\n")
```

```
##   = log((p(1-p0)) / (p0(1-p))): n = 6558
```

```r
cat("  = p - p0: n =", ceiling(n2), "\n")
```

```
##   = p - p0: n = 3159
```

```r
cat("  = arcsin(sqrt(p)) - arcsin(sqrt(p0)): n =", ceiling(n3), "\n")
```

```
##   = arcsin(sqrt(p)) - arcsin(sqrt(p0)): n = 4597
```

**Ejercicio 8**

Podemos hacer una función que te estime el error de tipo 1 y la potencia de un contraste:

```r
# Los parámetros de la función son:
# n -> número de simulaciones para estimar el error y la potencia
# parametrización -> número de parametrización a usar
# n_sim -> número de simulaciones de la binomial con la probabilidad real (debe de ser
# n1 si la parametrización es 1 y así con todas las parametrizaciones)
error_potencia <- function(n, parametrizacion, n_sim) {
  alpha = 0
  potencia = 0
  for (i in 1:n) {
    p0 = 0.003
    p_H0 = 0.003 # Probabilidad usada bajo la hipótesis nula en la que p = p0
    p_H1 = 0.006 # Probabilidad usada bajo la hipótesis alternativa en la que p != p0

    # Simular los datos con las distintas probabilidades y obtener el número de éxitos
    # bajo la hipótesis nula y bajo la hipótesis alternativa
    sim_H0 <- rbinom(n_sim, 1, p_H0)
    sim_H1 <- rbinom(n_sim, 1, p_H1)
    r_H0 = sum(sim_H0)
    r_H1 = sum(sim_H1)

    # Obtener Z y V
    par_H0 = calculate(p0, n_sim, r_H0, parametrizacion)
    par_H1 = calculate(p0, n_sim, r_H1, parametrizacion)

    # Z y V bajo hipótesis nula
    Z_H0 = par_H0[1]
    V_H0 = par_H0[2]

    # Z y V bajo hipótesis alternativa
    Z_H1 = par_H1[1]
    V_H1 = par_H1[2]

    # Cálculo de p-valor
    p_valor_H0 = pnorm(Z_H0, 0, sqrt(V_H0), lower.tail = FALSE)
    p_valor_H1 = pnorm(Z_H1, 0, sqrt(V_H1), lower.tail = FALSE)

    # Si el p-valor bajo la hipótesis nula es menor al nivel de significancia (0.025),
    #es decir que la rechazamos, aumentamos los fallos de tipo 1
    if (p_valor_H0 < 0.025) {
      alpha = alpha + 1
    }

    # Si el p-valor bajo la hipótesis alternativa es menor al nivel de significancia
```

```r
    #(0.025), es decir que la rechazamos, aumentamos la potencia
    if (p_valor_H1 < 0.025) {
      potencia = potencia + 1
    }
  }
  return (c(alpha/n, potencia/n))
}

# Calculamos los resultados para las distintas parametrizaciones:
resultados1 = error_potencia(50000, 1, n1)
resultados2 = error_potencia(50000, 2, n2)
resultados3 = error_potencia(50000, 3, n3)
```

```r
cat("Valor de alpha para la primera parametrización:", resultados1[1], "\n")
```

```
## Valor de alpha para la primera parametrización: 0.0284
```

```r
cat("Potencia para la primera parametrización: ", resultados1[2], "\n")
```

```
## Potencia para la primera parametrización:  0.96438
```

```r
cat("Valor de alpha para la segunda parametrización:", resultados2[1], "\n")
```

```
## Valor de alpha para la segunda parametrización: 0.032
```

```r
cat("Potencia para la segunda parametrización: ", resultados2[2], "\n")
```

```
## Potencia para la segunda parametrización:  0.78546
```

```r
cat("Valor de alpha para la tercera parametrización:", resultados3[1], "\n")
```

```
## Valor de alpha para la tercera parametrización: 0.02456
```

```r
cat("Potencia para la tercera parametrización: ", resultados3[2], "\n")
```

```
## Potencia para la tercera parametrización:  0.87738
```

### Ejercicio 9

Simulamos el número de éxitos con $p = 0.006$:

```r
p = 0.006
sim <- rbinom(n1, 1, p)
r = sum(sim)
```

- Primer test:

```r
test <- chisq.test(x = c(r, n1 - r), p = c(p0, 1 - p0), correct= FALSE)

print(test)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(r, n1 - r)
## X-squared = 19.048, df = 1, p-value = 1.274e-05
```

Como el p-valor es menor a 0.025, entonces rechazamos la hipótesis nula de que $p = 0.003$.

- Segundo test:

```r
test <- binom.test(round(r), round(n1), p0, alternative = "greater")

print(test)
```

```
##
##  Exact binomial test
##
## data:  round(r) and round(n1)
## number of successes = 39, number of trials = 6557, p-value = 7.577e-05
## alternative hypothesis: true probability of success is greater than 0.003
## 95 percent confidence interval:
##  0.004475584 1.000000000
## sample estimates:
## probability of success
##            0.005947842
```

Como p-valor es menor a 0.025, entonces rechazamos la hipótesis nula de que $p = 0.003$.

Podemos volver a realizar los mismos contrastes pero ahora simulando la binomial con $p = 0.003$, por lo que el resultado del contraste deberá de ser no rechazar la hipótesis nula:

Simular la binomial con $p = 0.003$:

```r
p = 0.003
sim <- rbinom(n1, 1, p)
r <- sum(sim)
```

- Primer test:

```r
test <- chisq.test(x = c(r, n1 - r), p = c(p0, 1 - p0), correct= FALSE)

print(test)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(r, n1 - r)
## X-squared = 0.005499, df = 1, p-value = 0.9409
```

Como el p-valor es mayor a 0.025, no rechazamos la hipótesis nula de que $p = 0.003$

- Segundo test:

```r
test <- binom.test(round(r), round(n1), p0, alternative = "greater")

print(test)
```

```
##
##  Exact binomial test
##
## data:  round(r) and round(n1)
## number of successes = 20, number of trials = 6557, p-value = 0.5004
## alternative hypothesis: true probability of success is greater than 0.003
## 95 percent confidence interval:
##  0.002022337 1.000000000
## sample estimates:
## probability of success
##            0.003050175
```

Como el p-valor es mayor a 0.025, no rechazamos la hipótesis nula de que $p = 0.003$

Cabe destacar que para simular la binomial hemos usado el tamaño muestral n1, porque es el que se usa en el test.

### Ejercicio 10

Utilizamos el Wilcoxon Rank-Sum Test para comprobar si dos muestras independientes provienen de la misma distribución, comparando las proporciones observadas (con un p = 0.006) y las esperadas (con p0 = 0.003)

```r
# Parámetros iniciales
n <- 1000          # Tamaño de la muestra
p0 <- 0.003        # Proporción esperada
p <- 0.006         # Proporción observada

muestra_observada <- rbinom(n = n, size = 1, prob = p)
muestra_esperada <- rbinom(n = n, size = 1, prob = p0)

test <- wilcox.test(muestra_observada, muestra_esperada, alternative = "greater")

print(test)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  muestra_observada and muestra_esperada
## W = 501000, p-value = 0.1586
## alternative hypothesis: true location shift is greater than 0
```