Лабораторна робота №5
з дисципліни "комп'ютерна статистика"
Варіант №4

Горбунова Даніела Денисовича
4 курс бакалаврату
група "комп'ютерна статистика"

26 листопада 2020 р.

1 Вступ.

У даній роботі побудовано критерій відношення вірогідностей для перевірки простих гіпотез про значення ймовірності успіху в біноміальному розподілі. Визначено обсяг вибірки, при якому тест матиме ймовірності похибок першого та другого роду, які не перевищують 0,05.

2 Побудова критерію.

Нехай $X = (\xi_1, \ldots, \xi_n)$ - кратна вибірка. Обчислимо функцію вірогідності за біноміальним розподілом $Bin(m, \theta)$, $\theta \in (0, 1)$:

$$L(X, \theta) = \prod_{j=1}^{n} C_m^{\xi_j} \theta^{\xi_j} (1-\theta)^{m-\xi_j} = \left(\prod_{j=1}^{n} C_m^{\xi_j} \right) \theta^{\sum_{j=1}^{n} \xi_j} (1-\theta)^{nm - \sum_{j=1}^{n} \xi_j}$$

Беремо $p, q \in (0, 1) : p < q$. Обчислимо за ними логарифмічну функцію вірогідності:

$$LR(X) = \frac{L(X, q)}{L(X, p)} = \frac{\cancel{\left(\prod_{j=1}^{n} C_m^{\xi_j} \right)} q^{\sum_{j=1}^{n} \xi_j} (1-q)^{nm - \sum_{j=1}^{n} \xi_j}}{\cancel{\left(\prod_{j=1}^{n} C_m^{\xi_j} \right)} p^{\sum_{j=1}^{n} \xi_j} (1-p)^{nm - \sum_{j=1}^{n} \xi_j}} = \left(\frac{q(1-p)}{p(1-q)} \right)^{\sum_{j=1}^{n} \xi_j} \left(\frac{1-q}{1-p} \right)^{nm} \Rightarrow$$

$$\ln LR(X) = \ln \frac{L(X, q)}{L(X, p)} = \sum_{j=1}^{n} \xi_j \ln \left(\frac{q(1-p)}{p(1-q)} \right) + mn \ln \left(\frac{1-q}{1-p} \right)$$

Зауважимо, що $\sum_{j=1}^{n} \xi_j$ є монотонно зростаючою функцією відносно $\ln LR(X)$, тому критерій відношення вірогідностей запишемо у вигляді:

$$\pi_C(X) = \mathbb{1} \left\{ \sum_{j=1}^{n} \xi_j > C_\alpha \right\}$$

Тобто якщо сума спостережень перевищує заданий поріг рівня α, тоді приймається нульова гіпотеза ($\pi_C(X) = 0$), інакше приймається альтернатива.

Зафікуємо рівень значущості $\alpha \in (0, 1)$. Тоді поріг рівня α визначимо з умови:

$$\alpha = \alpha_\pi = \mathbb{M}_p \pi_C(X) = \mathbb{P}_p \left\{ \sum_{j=1}^{n} \xi_j > C_\alpha \right\}$$

Ймовірність похибки другого роду для цього тесту:

$$\beta_\pi = 1 - \mathbb{M}_q \pi_C(X) = \mathbb{P}_q \left\{ \sum_{j=1}^{n} \xi_j \leqslant C_\alpha \right\}$$

Якщо $\{\xi_j\}_{j=1}^n$ - н.о.р. з біноміальним розподілом $Bin(m, \theta)$, тоді $\sum_{j=1}^n \xi_j \sim Bin(nm, \theta)$. Це нам може спростити життя хоча б в тому сенсі, що розподіл статистики відомий у явному вигляді. Застосуємо наведені раніше міркування для програмної реалізації тесту:

```
# H0 : Binom(0.5, 2), p := 0.5
# H1 : Binom(0.6, 2), q := 0.6
# n = 65
set.seed(0)
log.lr.test <- function(x, m, p, q, alpha = 0.05, to.print = F)
{
  x.sum <- sum(x)
  n <- length(x)
  c.alpha <- qbinom(1 - alpha, n * m, p)
  h.res <- c.alpha >= x.sum
  if(to.print)
  {
    print(paste("sum(X) = ", x.sum,
                ifelse(h.res, "<=", ">"),
                c.alpha, "= c_{alpha}"))
  }
  list(hypothesis = 1 - h.res, statistic = x.sum) # 0 - H0, 1 - H1
}
p <- 0.5
q <- 0.6
m <- 2
n.fixed <- 65
# Приклад застосування
print(log.lr.test(rbinom(65, 2, 0.5), 2, 0.5, 0.6, to.print = T))
# [1] "sum(X) =  68 <= 74 = c_{alpha}"
# $hypothesis
# [1] 0
# $statistic
# [1] 68
```

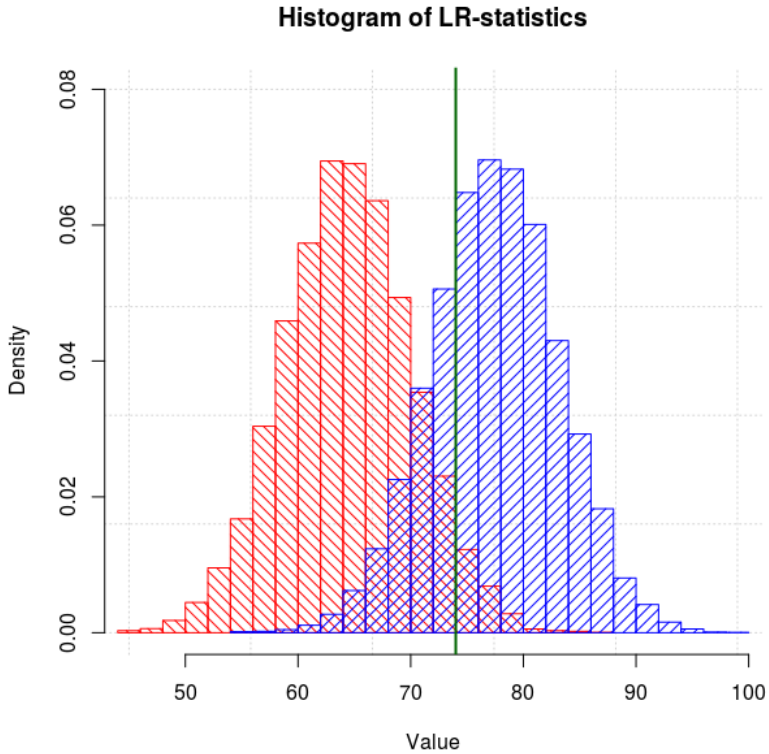

Рис. 1: Гістограма LR-статистик в залежності від обраної гіпотези. Зліва - гістограма за нульовою гіпотезою, справа - гістограма за розподілом альтернативи. Зелена вертикальна лінія - поріг тесту.

За допомогою імітаційного моделювання, після проведення $B = 10^4$ експериментів з використанням тесту відношення вірогідностей, отримані наступні оцінки імовірностей першого та другого роду:

```
# Імітаційне моделювання
B <- 10000
# Генеруємо вибірки за розподілом при виконанні нульової (альтернативної) гіпотези
# У відповідні масиви вносимо результат тесту та значення статистики
counts.0.1 <- numeric(B)
x.stat.0 <- numeric(B)
for(j in 1:B)
{
  x.0 <- rbinom(n.fixed, m, p); r <- log.lr.test(x.0, m, p, q)
  counts.0.1[j] <- r$hypothesis; x.stat.0[j] <- r$statistic
}
counts.1.1 <- numeric(B)
x.stat.1 <- numeric(B)
for(j in 1:B)
{
  x.1 <- rbinom(n.fixed, m, q); r <- log.lr.test(x.1, m, p, q)
  counts.1.1[j] <- r$hypothesis; x.stat.1[j] <- r$statistic
}
```

```r
# Оцінка для ймовірності помилки першого роду
print(
  paste("Estimated probability of I type error:",
        mean(counts.0.1))
  )
# "Estimated probability of I type error: 0.046"

# Оцінка для ймовірності помилки другого роду
print(
  paste("Estimated probability of II type error:",
        1 - mean(counts.1.1))
  )
# "Estimated probability of II type error: 0.2646"

# Гістограма статистик в залежності від параметра ймовірності
min.stat <- min(x.stat.0, x.stat.1)
max.stat <- max(x.stat.0, x.stat.1)

hist(x.stat.0, col = 'red', xlim = c(min.stat, max.stat), ylim = c(0, 0.08),
     freq = F, breaks = 20, main = "Histogram of LR-statistics", angle = -45,
     density = 15, xlab = "Value", panel.first = grid())
hist(x.stat.1, col = 'blue', xlim = c(min.stat, max.stat), ylim = c(0, 0.08),
     freq = F, breaks = 20, add = T, density = 15, angle = 45)

# Емпірична оцінка порогу тесту
c.b <- quantile(x.stat.0, 1 - 0.05)
print(c.b)
# 95%
# 74
abline(v = c.b, lwd = 2, col = 'darkgreen')
```

Практика непогано узгоджується з теорією в сенсі отриманих результатів, близьких до очікуваних.