Statystyka matematyczna - laboratorium nr 1

Bartosz Michalak

10 listopada 2025

Zadanie 1

Celem zadania jest porównanie dwóch estymatorów wariancji dla rozkładu Bernoulliego $\mathcal{B}(1, \vartheta)$:

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Pierwszy z nich jest estymatorem obciążonym, natomiast drugi — estymatorem nieobciążonym. Porównania dokonujemy na podstawie błędów średniokwadratowych (MSE), wyznaczonych metodą Monte Carlo dla różnych wartości n i ϑ.

Opis metody

Generujemy N niezależnych prób $\mathbf{X}^{(k)} = (X_1^{(k)}, \ldots, X_n^{(k)})$ z rozkładu Bernoulliego $\mathcal{B}(1, \vartheta)$. Dla każdej próby obliczamy S_0^2 i S^2, a następnie wyznaczamy błędy średniokwadratowe:

$$\mathrm{MSE}(S_0^2) = \mathbb{E}\big[(S_0^2 - \vartheta(1-\vartheta))^2\big], \qquad \mathrm{MSE}(S^2) = \mathbb{E}\big[(S^2 - \vartheta(1-\vartheta))^2\big].$$

Kod w R

```r
monte_carlo_mse_bernoulli <- function(n, theta, nsim = 20000, seed = 12345) {
  set.seed(seed)
  true_var <- theta * (1 - theta)
  s2_0_vals <- numeric(nsim)
  s2_vals   <- numeric(nsim)
  for (i in seq_len(nsim)) {
    x <- rbinom(n, 1, theta)
    xbar <- mean(x)
    ssq <- sum((x - xbar)^2)
    s2_0_vals[i] <- ssq / n
    s2_vals[i]   <- if (n > 1) ssq / (n - 1) else NA
  }
  list(
```

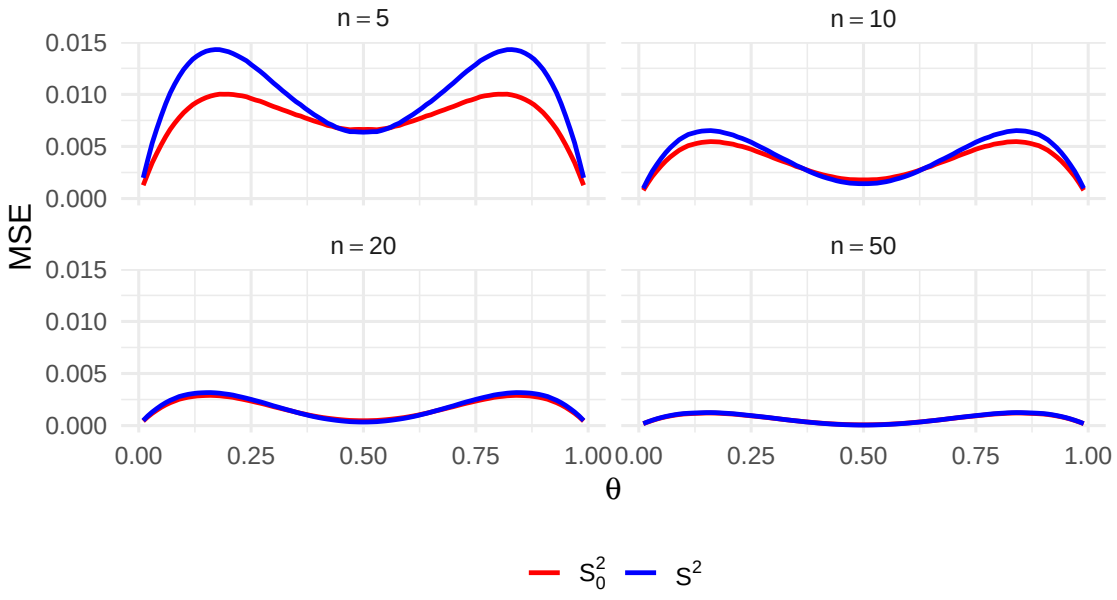
```
    n = n, theta = theta, true_var = true_var,
    mse_s2_0 = mean((s2_0_vals - true_var)^2, na.rm = TRUE),
    mse_s2   = mean((s2_vals - true_var)^2, na.rm = TRUE)
  )
}
```

Wyniki

Eksperymenty przeprowadzono dla różnych wartości $n \in \{5, 10, 20, 50\}$ oraz ϑ w przedziale $[0.01, 0.99]$ z krokiem co 0.01.

Wyniki wskazują, że estymator nieobciążony S_0^2 zazwyczaj osiąga niższe wartości MSE w porównaniu do estymatora maksymalnej wiarygodności S^2, szczególnie dla mniejszych próbek i wartości ϑ bliskich 0 lub 1.

Wraz ze wzrostem rozmiaru próby różnice w MSE między dwoma estymatorami maleją, co sugeruje, że oba estymatory stają się bardziej zbliżone w dużych próbkach.



Rysunek 1: Porównanie błędów średniokwadratowych (MSE) estymatorów $S_0{}^2$ i S^2 dla różnych wartości n i ϑ

Zadanie 2

Rozważamy rozkład dwumianowy $\mathcal{B}(n, \theta)$ oraz rodzinę estymatorów:

$$d_{a,b}(x) = \frac{x + a}{n + a + b},$$

gdzie $a, b > 0$. Funkcją straty jest tzw. *entropic loss* (lub log-loss):

$$L(\theta, a) = \mathbb{E}_\theta \left[\ln \frac{p_\theta(X)}{p_a(X)} \right],$$

gdzie $p_\theta(X)$ to funkcja masy prawdopodobieństwa rozkładu $\mathrm{Bin}(n, \theta)$. Naszym celem jest wyznaczenie wartości funkcji ryzyka:

$$R(\theta, d_{a,b}) = \mathbb{E}_\theta[L(\theta, d_{a,b}(X))].$$

Rozwiązanie

Najpierw rozpiszmy funkcję straty:

$$\begin{aligned} L(\theta, a) &= \mathbb{E}_\theta \left[\ln \frac{p_\theta(X)}{p_{a,b}(X)} \right] \\ &= \mathbb{E}_\theta \left[\ln \frac{\binom{n}{X} \theta^X (1-\theta)^{n-X}}{\binom{n}{X} a^X (1-a)^{n-X}} \right] \\ &= \mathbb{E}_\theta \left[X \ln \frac{\theta}{a} + (n - X) \ln \frac{1-\theta}{1-a} \right]. \end{aligned}$$

Ponieważ $\mathbb{E}_\theta[X] = n\theta$, otrzymujemy

$$L(\theta, a) = n\theta \ln \frac{\theta}{a} + n(1-\theta) \ln \frac{1-\theta}{1-a}.$$

Przejdźmy teraz do funkcji ryzyka:

$$\begin{aligned} R(\theta, d_{a,b}) &= \mathbb{E}_\theta[L(\theta, d_{a,b}(X))] \\ &= n\theta \, \mathbb{E}_\theta \left[\ln \frac{\theta}{d_{a,b}(X)} \right] + n(1-\theta) \, \mathbb{E}_\theta \left[\ln \frac{1-\theta}{1-d_{a,b}(X)} \right]. \end{aligned}$$

Podstawiając postać estymatora $d_{a,b}(X) = \frac{X+a}{n+a+b}$, mamy:

$$\begin{aligned} R(\theta, d_{a,b}) &= n\theta \, \mathbb{E}_\theta \left[\ln \frac{\theta(n+a+b)}{X+a} \right] + n(1-\theta) \, \mathbb{E}_\theta \left[\ln \frac{(1-\theta)(n+a+b)}{n-X+b} \right] \\ &= n\theta \ln\big(\theta(n+a+b)\big) + n(1-\theta) \ln\big((1-\theta)(n+a+b)\big) \\ &\quad - n\theta \, \mathbb{E}_\theta[\ln(X+a)] - n(1-\theta) \, \mathbb{E}_\theta[\ln(n-X+b)]. \end{aligned}$$

Ostatecznie otrzymujemy wzór:

$$\begin{aligned} R(\theta, d_{a,b}) = {} & n\theta \ln\big(\theta(n+a+b)\big) \\ & + n(1-\theta) \ln\big((1-\theta)(n+a+b)\big) \\ & - n\theta \, \mathbb{E}_\theta[\ln(X+a)] \\ & - n(1-\theta) \, \mathbb{E}_\theta[\ln(n-X+b)]. \end{aligned}$$

Wartości oczekiwane $\mathbb{E}_\theta[\ln(X+a)]$ oraz $\mathbb{E}_\theta[\ln(n-X+b)]$ można obliczyć numerycznie, korzystając z rozkładu dwumianowego:

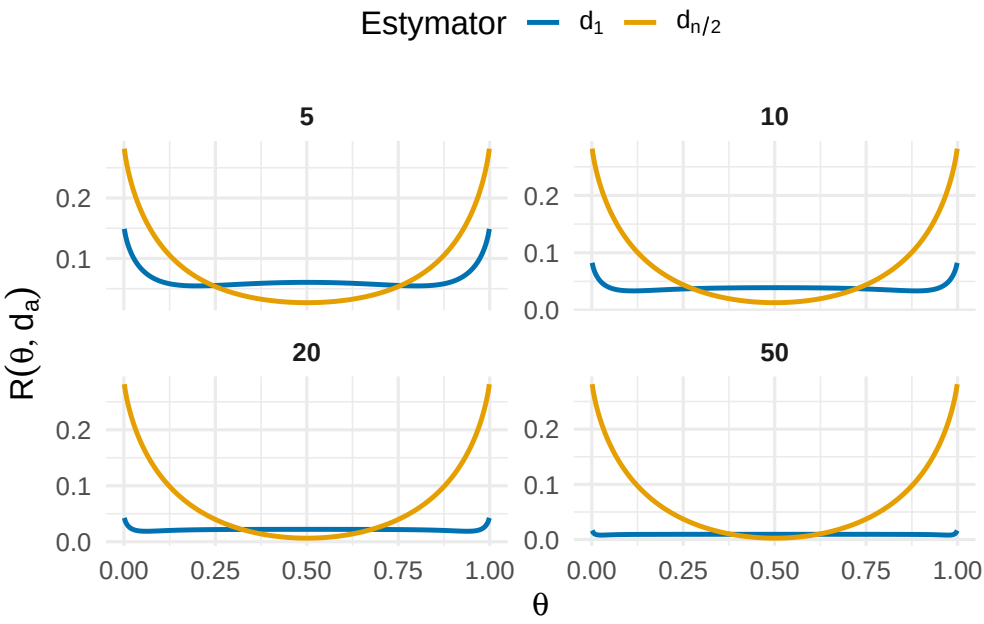
$$\mathbb{E}_\theta[f(X)] = \sum_{k=0}^{n} f(k)\,\mathrm{P}_\theta(X=k) = \sum_{k=0}^{n} f(k)\binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

Kod w R

```r
risk_entropy <- function(n, theta, a, b) {
  x <- 0:n
  pmf <- dbinom(x, size = n, prob = theta)
  d_x <- (x + a) / (n + a + b)
  eps <- .Machine$double.eps
  theta_clamped <- min(max(theta, eps), 1 - eps)
  term1 <- theta_clamped * log(theta_clamped / d_x)
  term2 <- (1 - theta_clamped) * log((1 - theta_clamped) / (1 - d_x))
  loss_x <- term1 + term2
  sum(pmf * loss_x)
}

risk_curve <- function(n, a, b,
                       thetas = seq(0.001, 0.999, length.out = 301)) {
  sapply(thetas, function(th) risk_entropy(n, th, a, b))
}
```

Wyniki



Rysunek 2: Porównanie funkcji ryzyka dla różnych wartości n

Zauważamy, że:

- Estymator $d_{1,1}$ ma stosunkowo płaską funkcję ryzyka — jest bardziej „równomierny" względem ϑ.
- Estymator $d_{n/2,n/2}$ osiąga mniejsze wartości ryzyka w pobliżu $\vartheta = 0.5$, lecz większe przy skrajnych wartościach.
- Wybór a, b wpływa na kompromis między minimalizacją ryzyka globalnie a lokalnie w okolicy konkretnej wartości ϑ.

Zadanie 3

Kontynuując wątek z Zadania 2, zakładamy teraz, że parametr ϑ ma rozkład a priori

$$\pi(\vartheta) = \text{Beta}(\alpha, \beta),$$

czyli gęstość prawdopodobieństwa

$$h(\vartheta) = \frac{\vartheta^{\alpha-1}(1-\vartheta)^{\beta-1}}{B(\alpha,\beta)}, \qquad \vartheta \in [0,1],$$

gdzie $B(\alpha, \beta)$ jest funkcją beta.

Rozważamy rodzinę estymatorów

$$d_{a,b}(x) = \frac{x+a}{n+a+b}, \qquad a, b > 0,$$

dla próby $X \sim \mathcal{B}(n, \vartheta)$.

Celem zadania jest znalezienie optymalnych wartości (a^*, b^*), które minimalizują średnie ryzyko.

Rozwiązanie

Średnie ryzyko można wyznaczyć numerycznie, korzystając z definicji:

$$r(\pi, d_{a,b}) = \int_0^1 R(\vartheta, d_{a,b})\, h(\vartheta)\, d\vartheta,$$

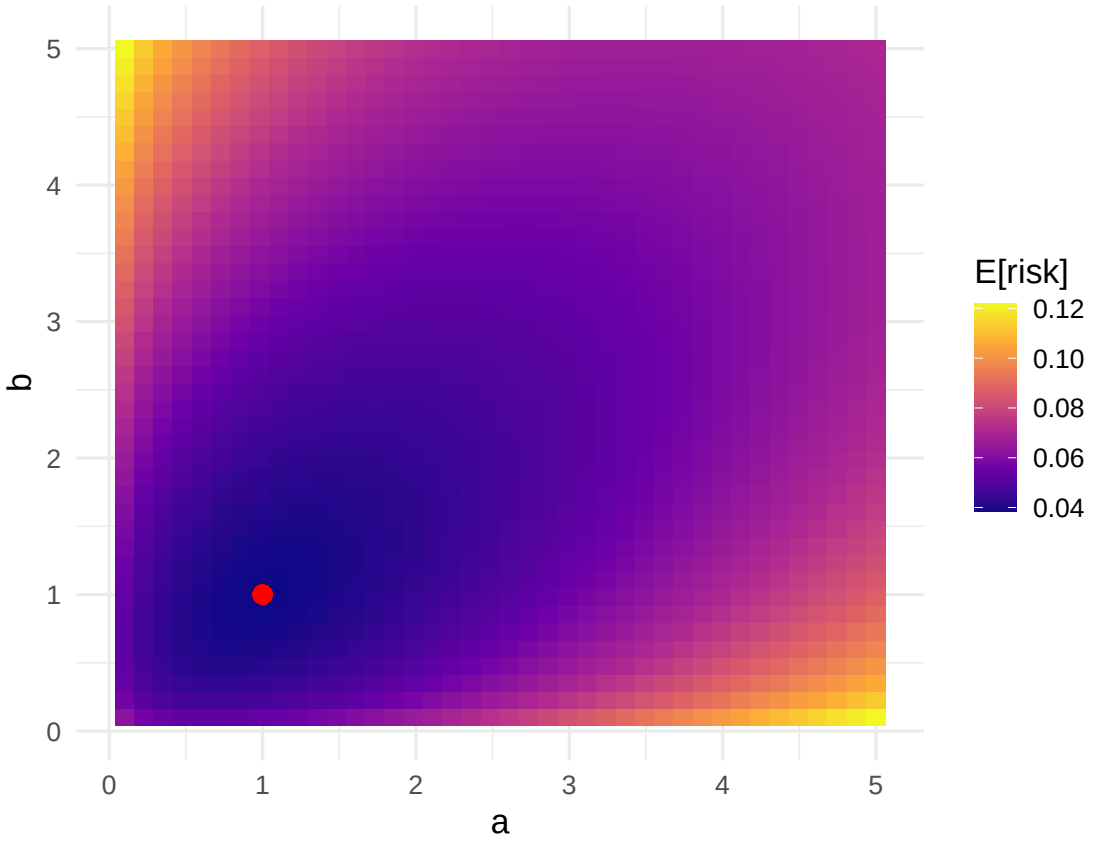
gdzie $R(\vartheta, d_{a,b})$ zostało wyprowadzone w Zadaniu 2.

Kod w R

```
expected_risk_prior <- function(n, a, b,
                                alpha = 1, beta = 1, rel.tol = 1e-8) {
  integrand <- function(theta) {
    sapply(theta, function(th)
      risk_entropy(n, th, a, b) * dbeta(th, shape1 = alpha, shape2 = beta))
  }
  res <- integrate(integrand, lower = 0, upper = 1,
                   rel.tol = rel.tol, subdivisions = 200L)
  res$value
}

optimize_ab_for_prior <- function(n, alpha = 1, beta = 1,
                                  a_start = 1, b_start = 1,
                                  lower = 1e-6, upper = 1000) {
  obj <- function(par) {
    a <- par[1]; b <- par[2]
    if (a <= 0 || b <= 0) return(1e10)
    expected_risk_prior(n, a, b, alpha, beta)
  }
  res <- optim(par = c(a_start, b_start), fn = obj, method = "L-BFGS-B",
               lower = c(lower, lower), upper = c(upper, upper))
  list(par = res$par, value = res$value,
       convergence = res$convergence, message = res$message)
}
```

Wyniki



Rysunek 3: Średnie ryzyko $r(\pi, d_{a,b})$ dla n $= 10$, $\alpha = 1$, $\beta = 1$. Czerwony punkt: optimum (a^*, b^*).

Dla $\alpha = \beta = 1$ rozkład a priori π jest jednostajny na przedziale $[0, 1]$. W tym przypadku średnie ryzyko Bayesowskie osiąga minimum dla estymatora, który symetrycznie uwzględnia wszystkie możliwe wartości ϑ.

Optymalnym estymatorem Bayesowskim dla entropic loss jest

$$d_{1,1}(X) = \frac{X+1}{n+2}.$$

Wyniki numeryczne uzyskane z naszego programu potwierdzają tę obserwację. Dla $n = 10$ optymalnie dobrane wartości parametrów (a^*, b^*) są bardzo bliskie 1 co zgadza się z przewidywaniami teoretycznymi.