Quarto základy - pokračování

Mgr. Zdeňka Geršlová

Příklad 1

Rozdělení výběrového rozptylu a výběrové směrodatné odchylky Pomocí simulační studie ověřte, že pokud náhodná veličina $X \sim N(\mu, \sigma^2)$, potom pro výběrový rozptyl S_n^2 a výběrovou směrodatnou odchylku S_n platí následující vztahy:

1. $S_n^2 \sim \Gamma(\frac{n}{2}, \frac{2\sigma^2}{n})$ exaktně;
2. $S_n^2 \sim N(\sigma^2, \frac{2\sigma^4}{n})$ asymptoticky;
3. $S_n \sim \Gamma_G(\sqrt{\frac{2\sigma^2}{n}}, 2, \frac{n}{2})$ exaktně;
4. $S_n \sim N(\sigma, \frac{\sigma^2}{2n})$ asymptoticky.

Postup

Vygenerujte $M = 1000$ náhodných výběrů z normálního rozdělení $N(\mu, \sigma^2)$ o rozsahu n, kde $\mu = 0$ a $\sigma^2 = 4$. Pro každý náhodný výběr vypočítejte výběrový rozptyl $S_{n_i}^2$, $i = 1, \dots, M$.

Rozptyly vykreslete pomocí histogramu a superponujte je křivkami hustoty exaktního i asymptotického rozdělení statistiky S_n^2.

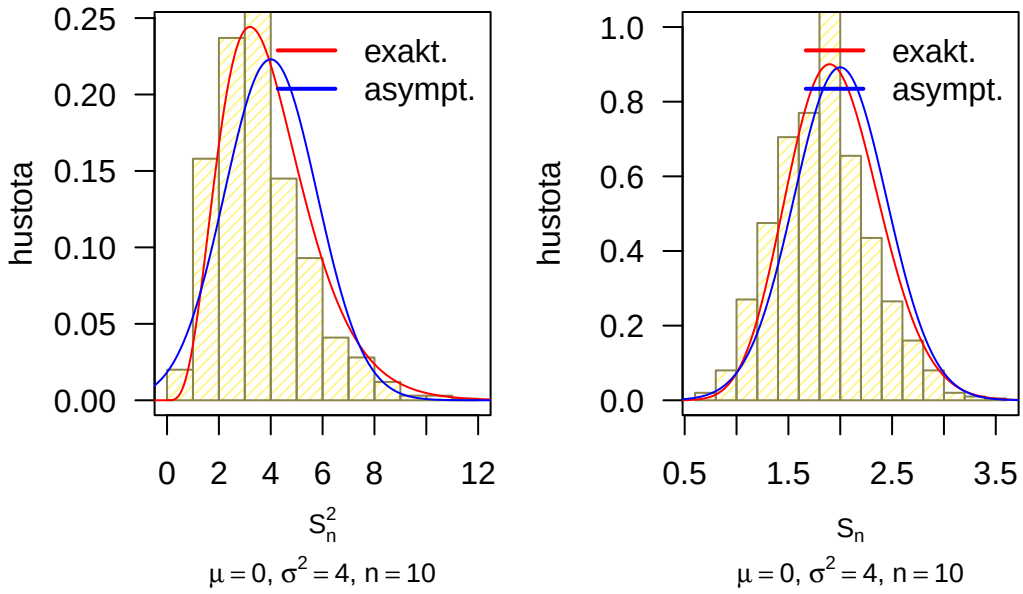
Dále pro každý náhodný výběr vypočítejte výběrovou směrodatnou odchylku S_{n_i}, $i = 1, \dots, M$. Odchylky zaneste do histogramu a superponujte je křivkami hustoty asymptotického a exaktního rozdělení statistiky S_n.

Vytvořte animaci zobrazující konvergenci asymptotického rozdělení obou statistik k exaktnímu rozdělení při zvětšujícím se rozsahu náhodných výběrů n. Hodnoty rozsahu n volte 5, 10, 50, 100, 500, 1000.

Pozn: Opět vytvoříme nejprve funkce v souboru source. Funkci pro výpočet výběrového rozptylu s použitím 1/n a funkci `RozdeleniSn2Sn`*, která bude pro vstupní parametry* `sigma, n, M` *(tj. sm. odchylka zadaného normálního rozdělení, rozsah a počet náhodných výběrů) vytvářet požadované grafy.*

Výsledek staticky

```
source("M8986-source.R")
RozdeleniSn2Sn(2, n = 10)
```



Nápověda

```
VyberovyRozptyl <- function(X) {
  ... # vypocet rozptylu s pouzitim /n
}

RozdeleniSn2Sn <- function(sigma, n, M = 1000) {
  Sn2 <- replicate(M, VyberovyRozptyl(rnorm(n, 0, sigma))) # zopakuje M-krat
  Sn <- # smerodatna odchylka - odmocnenim
  xfit_Sn2 <- seq(from = min(Sn2) - 1, to = max(Sn2) + 1, length = 512) # sekvence rozptyl
  xfit_Sn <- ... # sekvence sm. odchylek

  yfit_e <- dgamma(xfit_Sn2, shape = n / 2, scale = 2 * sigma^2 / n)
  # pozor na spravne zadane shape a scale u gamma rozdeleni
  yfit_a <- ... # asymptoticke rozd. rozptylu
```

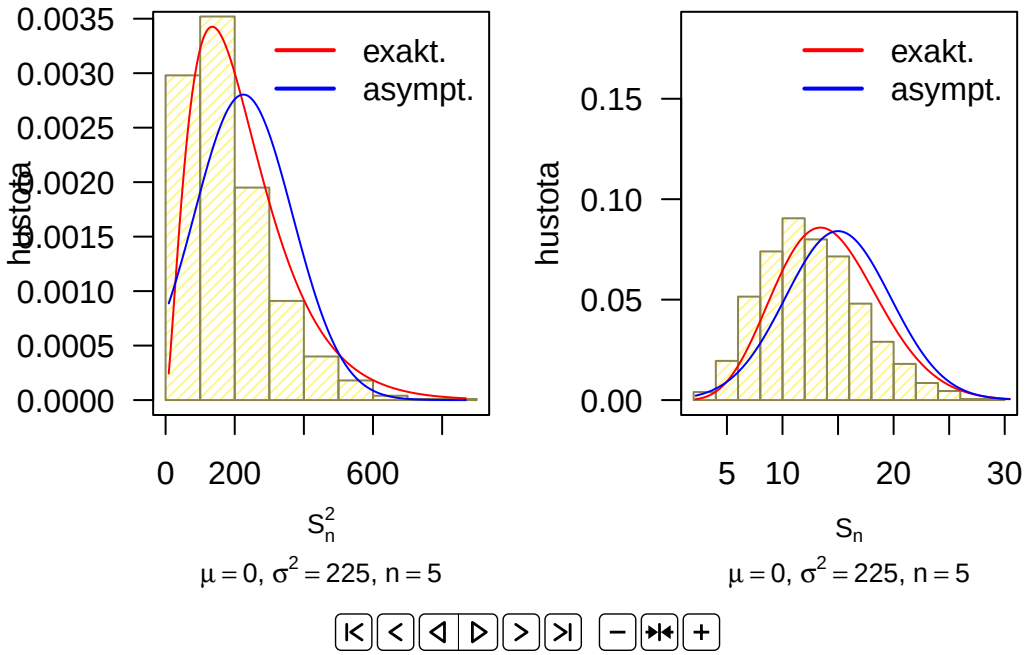
```
  zfit_e <- VGAM::dgengamma.stacy(xfit_Sn, scale = sqrt(2 * sigma ^ 2 / n),
                                  d = 2, k = n / 2) # ex. sm. odchylka
  zfit_a <- ... # asymptot. sm. odchylka

  par(mfrow = c(1, 2), mar = c(5, 4, 2, 1)) # nastaveni grafickeho okna
  hist(..., prob = T, ylim = c(0, as.numeric(max(yfit_e, yfit_a))), ...)
  lines(...) # cervene exaktni hustota
  lines(...) # modre asymptoticka hustota
  mtext(expression(S[n]^2), side = 1, line = 2.5, cex = 0.8)
  mtext(bquote(paste(mu == 0, ", ", sigma^2 == .(sigma^2), ", ", n == .(n))),
        side = 1, line = 3.7, cex = 0.8)
  legend(...) # legenda

 ... # totez pro sm. odchylku
}
```

Animace

```
```{r}
#| fig-show: animate
n <- c(5, 10, 50, 100, 500, 1000)
for (i in 1:length(n)) {
  RozdeleniSn2Sn(15, n = n[i])
}
```
```



Příklad 2

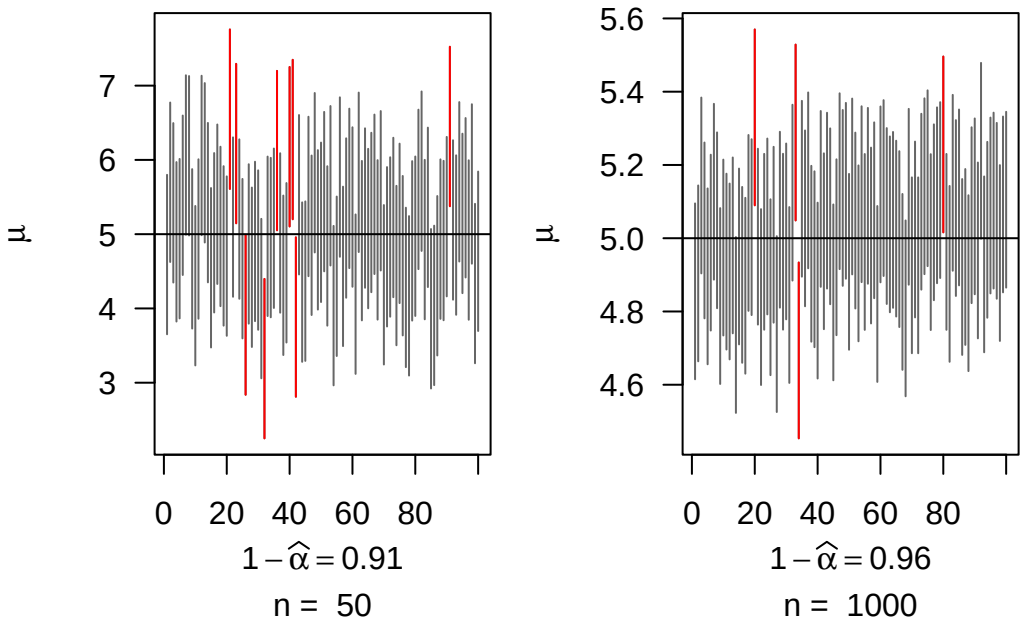
Odhad koeficientu spolehlivosti $1-\alpha$ Waldova empirického DIS pro parametr μ normálního rozdělení při známém rozptylu σ^2
Nechť $X \sim N(5, 15)$. Pomocí simulační studie ($M = 100$) stanovte Monte Carlo (MC) odhad koeficientu spolehlivosti (pravděpodobnosti pokrytí) 95 % Waldova exaktního empirického DIS pro parametr μ normálního rozdělení při známém rozptylu σ^2.

Vygenerujte $M = 100$ náhodných výběrů z normálního rozdělení $N(5, 10)$ a na základě každého náhodného výběru vypočítejte 95 % DIS pro parametr μ.
Všechny DIS vykreslete do jednoho grafu jako svislé šedé úsečky. Červenou barvou dále vyznačte v grafu ty DIS, které nepokrývají střední hodnotu $\mu = 5$ a černou barvou vyznačte horizontální referenční čáru v bodě μ.
Dále vypočítejte aktuální pravděpodobnost pokrytí 95 % DIS pro μ jako podíl $\frac{\sum_m I(\mu \in IS_m)}{M}$ a porovnejte ji s nominální pravděpodobností pokrytí $1-\alpha$.
Rozsah náhodných výběrů volte (a) $n = 50$; (b) $n = 1000$.

Simulaci proveďte také za předpokladu, že data pochází ze smíšeného rozdělení $X \sim pN(5, 15) + (1-p)N(5, 15^2)$, kde $p = 0.9$, resp. ze smíšeného rozdělení $X \sim pN(5, 15) + (1-p)N(3, 15)$, kde $p = 0.9$.

Výsledné grafy

```
par(mfrow = c(1,2))
d1 <- DisMu(n = 50)
d2 <- DisMu(n = 1000)
```



Univerzální postup je vytvořit funkci DisMu připravenou obecně pro směs rozdělení, na vstupu budou parametry vztahující se k jednotlivým rozdělením (resp. směsi rozdělení), počet simulací a hladina významnosti.

Na výstupu chceme mít

- obrázek dle požadavků zadání,
- tabulku s hodnotami spolehlivosti, pravděpodobnosti pokrytí a dolní a holní hranice intervalu pro α.

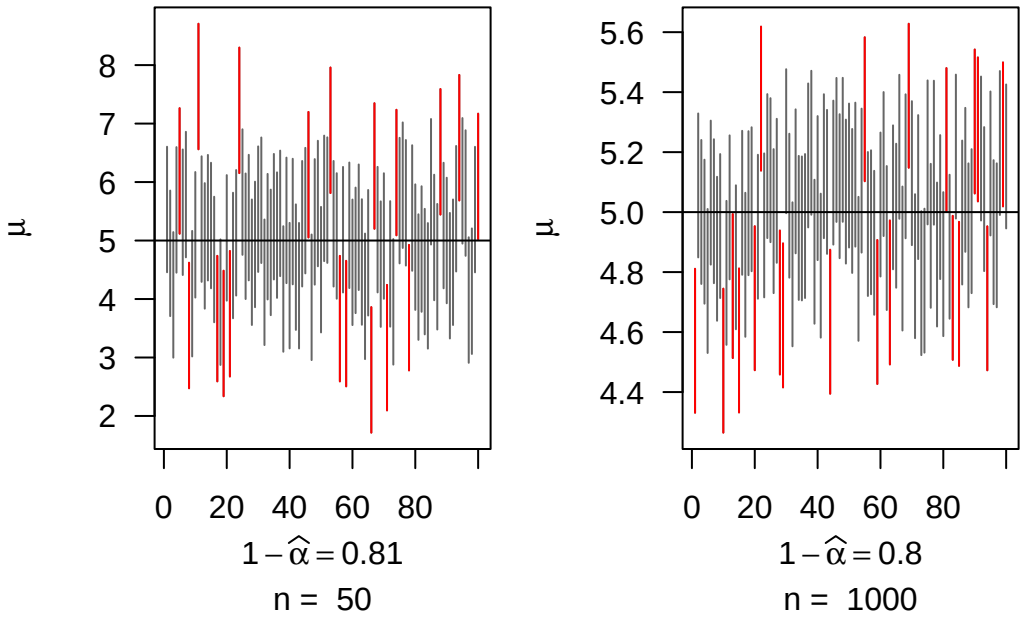
```r
DisMu <- function(n, mu = 5, mu2 = mu,
                  sigma = sqrt(15), sigma2 = sigma,
                  M = 100, p = 0.9, alpha = 0.05){
X <- matrix(NA, M, n) # vygenerovani smesi rozdeleni
  for (i in 1:M) {
    bin <- rbinom(n, 1, p)
    X[i, ][bin == 1] <- rnorm(sum(bin == 1), mu, sigma)
    X[i, ][bin == 0] <- rnorm(sum(bin == 0), mu2, sigma2)
  }
m <- ... # odhad mu
dh <- ... # dolni hranice Waldova IS pro mu
hh <- ... # horni hranice Waldova IS pro mu
rozhodnuti <- (dh < mu) & (mu < hh)
a_pst_pokryti <- sum(rozhodnuti) / M
a_alpha <- 1 - a_pst_pokryti

... # vykresleni grafu
# pridani usecek pomoci segments
segments(x, dh, x, hh, col = "grey40")
segments(x[rozhodnuti == 0], dh[rozhodnuti == 0], x[rozhodnuti == 0], hh[rozhodnuti == 0],
         col = "red")

# vytvoreni pozadovane tabulky
SE <- sqrt(a_alpha * (1 - a_alpha) / M)
dh_alpha <- 1 - a_alpha - qnorm(1 - alpha / 2) * SE #
hh_alpha <- 1 - a_alpha - qnorm(alpha / 2) * SE #
tab <- data.frame(spolehlivost = 1 - alpha, a_pst_pokryti, dh_alpha, hh_alpha)
return(tab)
}
```

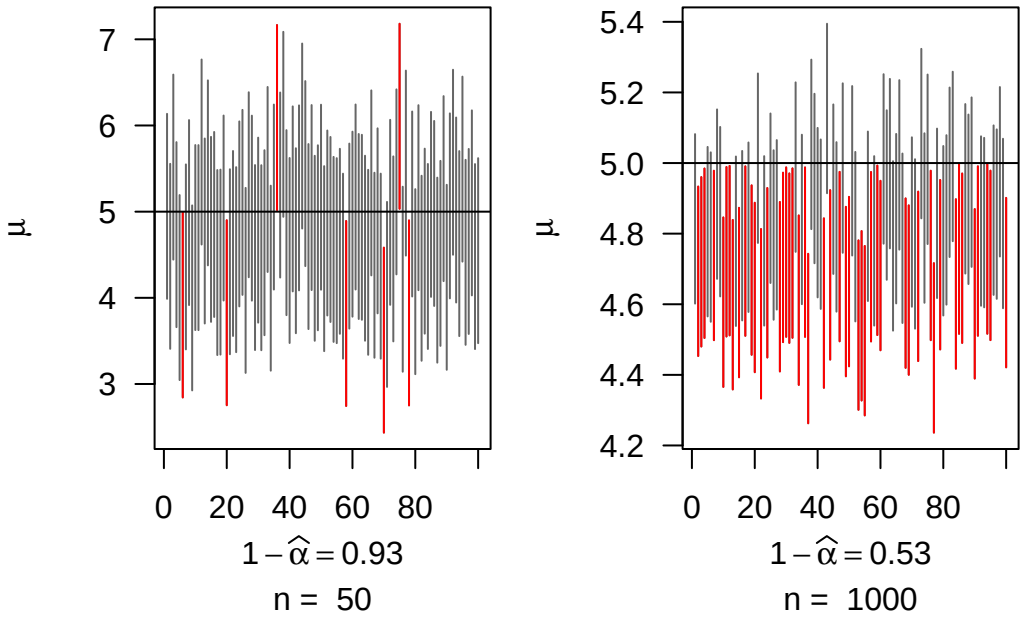
Směs rozdělení I.

```r
par(mfrow = c(1,2))
d3 <- DisMu(n = 50,  sigma2 = 15)
d4 <- DisMu(n = 1000, sigma2 = 15)
```



Směs rozdělení II.

```
par(mfrow = c(1,2))
d5 <- DisMu(n = 50,  mu2 = 3)
d6 <- DisMu(n = 1000, mu2 = 3)
```



Závěr

Zopakujeme-li několikrát simulační studii, vidíme, že pochází-li náhodné výběry z normálního rozdělení, pohybuje se aktuální pravděpodobnost pokrytí okolo nominální pravděpodobnosti pokrytí $1-\alpha$. V případě, že náhodné výběry pochází ze směsi dvou normálních rozdělení, které se liší pouze v rozptylech, dochází s rostoucím rozdílem mezi oběma rozptyly k postupnému snižování aktuální pravděpodobnosti pokrytí, a tedy ke zvýšení aktuální hladiny významnosti. Aby se však tento trend projevil, museli bychom hodnotu rozptylu σ_2^2 pozměnit opravdu výrazně (viz např. $\sigma_1^2 = 15$ vs $\sigma_2^2 = 15^2$). Podobná situace nastává v případě, že náhodné výběry pochází ze směsi dvou normálních rozdělení, které se liší střední hodnotou. Zde dochází k postupnému snižování aktuální pravděpodobnosti pokrytí, a tedy ke zvyšování aktuální hladiny významnosti s rostoucím rozsahem náhodných výběrů n.

Tabulka pomocí kable

```r
```{r}
#| label: tbl-dis-mu
#| tbl-cap: "Pravděpodobnost pokrytí"
library(knitr)
tab <- rbind(d1, d2, d3, d4, d5, d6)
```
```

```
names(tab) <- c("$1-\\alpha$", "$1-\\widehat{\\alpha}$",
                "$\\mathrm{dh}_\\alpha$", "$\\mathrm{hh}_\\alpha$")
row.names(tab) <- c("norm. rozd.,  $n = 50$", "norm. rozd.,  $n = 1000$",
                    "směs s různými $\\sigma_1^2$ a $\\sigma_2^2$, $n = 50$",
                    "směs s různými $\\sigma_1^2$ a $\\sigma_2^2$, $n = 1000$",
                    "směs s různými $\\mu_1$ a $\\mu_2$, $n = 50$",
                    "směs s různými $\\mu_1$ a $\\mu_2$, $n = 1000$")
kable(tab, digits = 4) # escape = F
```

	$1-\alpha$	$1-\hat{\alpha}$	dh_α	hh_α
norm. rozd., $n = 50$	0.95	0.91	0.8539	0.9661
norm. rozd., $n = 1000$	0.95	0.96	0.9216	0.9984
směs s různými σ_1^2 a σ_2^2, $n = 50$	0.95	0.81	0.7331	0.8869
směs s různými σ_1^2 a σ_2^2, $n = 1000$	0.95	0.80	0.7216	0.8784
směs s různými μ_1 a μ_2, $n = 50$	0.95	0.93	0.8800	0.9800
směs s různými μ_1 a μ_2, $n = 1000$	0.95	0.53	0.4322	0.6278

Table 1: Pravděpodobnost pokrytí

Pozn.: Nastavení pozice titulku tabulky provádíme v YAML hlavičce pro celý dokument pomocí `tbl-cap-location`.