

Test o rozdílu středních hodnot

Mgr. Zdeňka Geršlová

Příklad 1

Pravděpodobnost pokrytí klasického a věrohodnostního dvouvýběrového testu o střední hodnotě

Pomocí simulační studie ($M = 2000$) vypočítejte pravděpodobnost pokrytí 95,% DIS pro rozdíl $\mu_1 - \mu_2$, a to jako podíl $\frac{\sum_{m=1}^M I(|t_{W,m}| < t_{df}(1-\alpha/2))}{M}$, kde t_W, m jsou testovací statistiky

- (1) klasického dvouvýběrového t -testu;
- (2) dvouvýběrového t -testu s Welchovou aproximací;
- (3) věrohodnostního testu za předpokladu, že rozptyly σ_1^2 a σ_2^2 jsou neznámé ale shodné;
- (4) věrohodnostního testu za předpokladu, že rozptyly σ_1^2 a σ_2^2 jsou neznámé a různé.

Hodnoty parametrů volte následující:

- (a) $X_j \sim N(\mu_j, \sigma^2)$, kde $j = 1, 2$, $\mu_1 = 20$, $\mu_2 = 20$ a $\sigma^2 = 9^2$;
- (b) $X_j \sim N(\mu_j, \sigma_j^2)$, kde $j = 1, 2$, $\mu_1 = 20$, $\mu_2 = 20$, $\sigma_1^2 = 9^2$ a $\sigma_2^2 = 12^2$;
- (c) $X_j \sim pN(\mu_j, \sigma^2) + (1-p)N(\mu_j, \sigma_a^2)$, kde $j = 1, 2$, $\mu_1 = 20$, $\mu_2 = 20$, $\sigma^2 = 9^2$ a $\sigma_a^2 = 18^2$, $p = 0.8$;
- (d) $X_j \sim pN(\mu_j, \sigma_j^2) + (1-p)N(\mu_j, \sigma_{ja}^2)$, kde $j = 1, 2$, $\mu_1 = 20$, $\mu_2 = 20$, $\sigma_1^2 = 9^2$, $\sigma_2^2 = 12^2$, $\sigma_{1a}^2 = 18^2$, $\sigma_{2a}^2 = 22^2$, $p = 0.8$.

Rozsahy náhodných výběrů zvolte (i) $n_1 = n_2 = 5$; (ii) $n_1 = n_2 = 50$; (iii) $n_1 = n_2 = 100$.

Požadovaný výstup

Pro každou situaci (a)–(d) vykreslete spojitý diagram zachycující pravděpodobnost pokrytí pro DIS (1)–(4) při volbách rozsahů n_1 a n_2 (i)–(iii). Jednotlivé typy DIS v grafu barevně odlište.

Nakonec zhodnoťte uvedené typy DIS ((1)–(4)) podle pravděpodobnosti pokrytí a uveďte, který DIS má z hlediska pravděpodobnosti pokrytí nejlepší a který naopak nejhorší vlastnosti. V úvahu vezměte jednak změnu vzhledem k rozsahu náhodného výběru a jednak chování pravděpodobnosti pokrytí v různých situacích (a)–(d).

Výpočet aktuální pravděpodobnosti pokrytí

V případě Waldova DIS si můžeme vybrat, zda čítel vzorce výpočtu aktuální pravděpodobnosti pokrytí bude počet testovacích statistik $t_{W,m}$, které náležejí do kritického oboru W , nebo počet intervalů spolehlivosti IS_m , které pokrývají $\mu_1 - \mu_2 = \mu_0 = 0$. Oba postupy jsou ekvivalentní.

V případě věrohodnostního DIS je vhodnější počítat čítel jako počet testovacích statistik $t_{W,m}$ náležících do kritického oboru W . Při výpočtu přes IS_m bychom k přesnému stanovení hranic potřebovali funkci `uniroot()`, kde by ale byl problém s automatizováním počátečních podmínek. Druhou možností je spočítat hranice IS_m vygenerováním posloupnosti rozdílů $\mu_1 - \mu_2$, spočítat hodnoty ULR test. statistik pro každý rozdíl a hranice stanovit jako největší, resp. nejmenší hodnotu rozdílu, pro niž je $ULR_m < \chi_1^2(1 - \alpha)$. Tím ovšem získáme pouze přibližné hranice IS_m . V okrajových případech by se tak mohlo stát, že rozdíl $\mu_1 - \mu_2$ bude nesprávně zařazen do/mimo IS a aktuální pravděpodobnost pokrytí by nebyla vypočítaná správně. **Z toho důvodu použijeme ve všech případech kritérium pomoci test. statistiky a kritického oboru.**

Waldovy testovací statistiky

$$(1) T_W = \frac{\bar{X}_1 - \bar{X}_2 - \mu_0}{S_D} \sim t_{df},$$

$$df = n_1 + n_2 - 2 \text{ a } S_D^2 = \frac{n_1 + n_2}{n_1 n_2} S^2,$$

$$\text{kde } S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$(2) T_W = \frac{\bar{X}_1 - \bar{X}_2 - \mu_0}{S_D} \sim t_{df}, \text{ přičemž}$$

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}} \text{ a } S_D^2 = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

Věrohodnostní statistiky

$$(3) U_{LR} = n \ln\left(1 + \frac{t_W^2}{n-2}\right) \sim \chi_1^2, \text{ kde } n = n_1 + n_2$$

$$(4) U_{LR} = \frac{n_1 n_2 (\bar{x}_2 - \bar{x}_1)^2}{(n_1 + \gamma n_2)^2} \left(\frac{\gamma^2 n_2}{\hat{\sigma}_1^2} + \frac{n_1}{\hat{\sigma}_2^2} \right) \sim \chi_1^2,$$

$$\text{kde } \gamma = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2},$$

$$\hat{\sigma}_j^2 = \frac{1}{n_j} \sum_{i=1}^{n_j} (x_{ji} - \bar{x}_j)^2$$

Pozn.: Výpočet $\hat{\sigma}_j^2$ je uveden zvláště proto, abyste si uvědomili, že v něm používáme v čitateli n_j a nikoliv $n_j - 1$. Pro výpočet $\hat{\sigma}_j$ tedy můžeme použít upravenou směrodatnou odchylku získanou z funkce `sd` například takto:

```
s1 <- sqrt((n1 - 1) / n1 * s1^2) # s1 je sm. odchylka
```

Pokud bychom chtěli použít přímo samotný vzorec, je důležité si uvědomit, že chceme výpočet provést zvlášť pro každý řádek (resp. sloupec, podle toho, jak kdo generuje data).

Aktuální pravděpodobnost pokrytí

- $\frac{\sum_{m=1}^M I(|t_{W,m}| < t_{df}(1-\alpha/2))}{M}$ pro (1) a (2)
- $\frac{\sum_{m=1}^M I(U_{LR} < \chi_1^2(1-\alpha))}{M}$ pro (3) a (4)

Postup I.

Vytvoříme funkci `PokrytiRozdilMu`, která bude pro dané parametry směsi, rozsahy náh. výběrů, počet simulací a hladinu významnosti počítat pravděpodobnosti pokrytí všech testů (1) - (4).

V této funkci nejprve vytvoříme směs rozdělení, potom pro tuto směs spočítáme test. statistiky a aktuální pravděpodobnost pokrytí a výstupem bude vektor pravděpodobností pokrytí všech statistik (tj. 4-místný vektor).

```
PokrytiRozdilMu <- function(mu0, mu1, mu2 = mu1,
                             sig1, sig2 = sig1,
                             sig1a = sig1, sig2a = sig2,
                             n1 = 5, n2 = n1, M = 2000,
                             alpha = 0.05, p = 0.8) {
```

```

## mu0 ... rozdíl středních hodnot (pro nás příklad 0)
## mu1, mu2 ... jednotlivá mu
## sig1, sig2 ... sd jednotlivých vyberu
## sig1a, sig2a ... sd pro směsi norm. rozdělení
## tj. sigma_1a a sigma_2a v zadání příkladu
## n1, n2 ... rozsahy nah. vyberu
## M ... počet simulací
## alpha ... hl. významnosti
## p ... parametr směsi rozdělení

...

}

```

Postup II.

```

## uvnitř funkce
## generování směsi
X1 <- matrix(NA, M, n1)
X2 <- matrix(NA, M, n2)
for (i in 1:M) {
  bin1 <- rbinom(n1, 1, p)
  bin2 <- rbinom(n2, 1, p)
  X1[i, ][bin1 == 1] <- rnorm(sum(bin1 == 1), mu1, sig1)
  X1[i, ][bin1 == 0] <- rnorm(sum(bin1 == 0), mu1, sig1a)
  ... # doplňte X2 se správnými parametry
}

m1 <- apply(X1, 1, mean)
m2 <- ...
s1 <- apply(X1, 1, sd)
s2 <- ...
s <- ... # doplňte vzorec
sD <- ... # doplňte vzorec
tW_Wald <- (m1 - m2 - mu0) / sD
df_Wald <- ...
pokryti_Wald <- sum(abs(tW_Wald) < qt(1 - alpha / 2, df_Wald)) / M

... # točez pro ostatní statistiky

```

Postup III.

```
a5 <- PokrytiRozdilMu(mu0 = 0, mu1 = 20, sig1 = 9, n1 = 5)
a50 <- ...
a100 <- ...

## analogicky varianty (b) - (d)

## sestaveni tabulky
pst_pokryti <- rbind(a5, a50, a100, ...)

tab <- data.frame(mu1 = rep(20, 12), mu2 = rep(20, 12),
                  sig1 = rep(9, 12),
                  sig2 = ...,
                  ...,
                  n1 = rep(c(5, 50, 100), 4),
                  n2 = ...,
                  pst_pokryti)

row.names(tab) <- ...
names(tab) <- c("$\\mu_1$", "$\\mu_2$",
               "$\\sigma_{11}$", "$\\sigma_{21}$",
               "$\\sigma_{12}$", "$\\sigma_{22}$",
               "$n_1$", "$n_2$",
               "$Pr_{W}$", "$Pr_{W+WA}$", "$Pr_{V(SR)}$", "$Pr_{V(RR)}$")

kable(...)
```

Výsledky

	μ_1	μ_2	σ_{11}	σ_{21}	σ_{12}	σ_{22}	n_1	n_2	Pr_W	Pr_{W+WA}	$Pr_{V(SR)}$	$Pr_{V(RR)}$
$(a)_5$	20	20	9	9	NA	NA	5	5	0.96	0.97	0.91	0.89
$(a)_{50}$	20	20	9	9	NA	NA	50	50	0.95	0.95	0.95	0.94
$(a)_{100}$	20	20	9	9	NA	NA	100	100	0.96	0.96	0.95	0.95
$(b)_5$	20	20	9	12	NA	NA	5	5	0.94	0.95	0.90	0.87
$(b)_{50}$	20	20	9	12	NA	NA	50	50	0.95	0.95	0.95	0.95
$(b)_{100}$	20	20	9	12	NA	NA	100	100	0.95	0.95	0.95	0.95
$(c)_5$	20	20	9	9	18	18	5	5	0.94	0.95	0.91	0.88
$(c)_{50}$	20	20	9	9	18	18	50	50	0.95	0.95	0.94	0.94
$(c)_{100}$	20	20	9	9	18	18	100	100	0.95	0.95	0.95	0.95

	μ_1	μ_2	σ_{11}	σ_{21}	σ_{12}	σ_{22}	n_1	n_2	Pr_W	Pr_{W+W}	$Pr_{V(SR)}$	$Pr_{V(RR)}$
$(d)_5$	20	20	9	12	18	22	5	5	0.96	0.96	0.92	0.89
$(d)_{50}$	20	20	9	12	18	22	50	50	0.95	0.95	0.94	0.94
$(d)_{100}$	20	20	9	12	18	22	100	100	0.95	0.95	0.95	0.95

Tvorba grafu

```

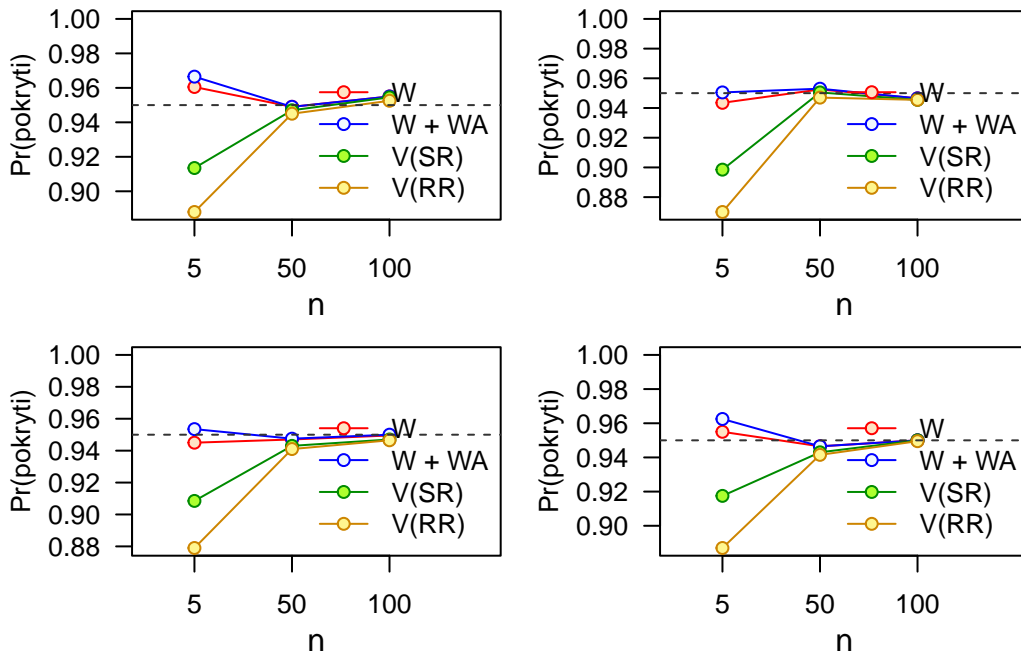
PokrytiPlot <- function(X, position = 'bottomright') {
  par(mar = c(3, 4, 1, 1))
  # graf varianty (a)
  plot(X[,1], type = 'o', xlim = c(0.5, 4), ylim = c(min(X), 1),
        pch = ..., col = ..., bg = ..., axes = F, ...)
  box(bty = 'o')
  axis(1, 1:3, labels = c(5, 50, 100))
  axis(2, las = 1)

  lines(...) # postupne pridame dalsi varianty (b) - (d)
  abline(h = 0.95, lty = 2, col = 'grey20')
  legend(position, lty = 1, pch = ..., bty = 'n',
         col = ..., pt.bg = ..., legend = ...)
}

par(mfrow = (2,2))
PokrytiPlot(cbind(a5, a50, a100))
... # analogicky pro dalsi varianty

```

Výsledné grafy



Závěr

Z hlediska pravděpodobnosti pokrytí vychází pro všechny situace (a)–(d) nejlépe Waldovy DIS pro shodné rozptyly. Tyto DIS nejsou konzervativní ani liberální již pro nízké rozsahy ($n \geq 5$) náhodných výběrů.

Pro vyšší rozsahy ($n \geq 50$) vychází dobře také Waldovy DIS pro různé rozptyly (pro nižší rozsahy $n < 50$ jsou konzervativní).

Věrohodnostní DIS vychází z hlediska pravděpodobnosti pokrytí dobře pro náhodné výběry s rozsahy $n \geq 100$. Pro $n \leq 100$ jsou věrohodnostní DIS mírně liberální, pro nízké rozsahy ($n \leq 5$) velmi liberální (tj. nevhodné).

Pozn.: Z hlediska pravděpodobnosti pokrytí není příliš výrazný rozdíl mezi intervaly spolehlivosti sestavenými za předpokladu shodných vs. různých rozptylů.

Praktický příklad

Dvouvýběrový test o rozdílu středních hodnot $\mu_1 - \mu_2$

Načtěte datový soubor `03-paired-mean-clavicle.txt`. Zmíněný soubor obsahuje osteometrická data o délce klíční kosti (clavicula) anglického souboru 50 mužských a 50 ženských dokumentovaných skeletů. Konkrétně jde o délku klíční kosti z pravé strany těla (`length.R`) a levé strany těla (`length.L`).

Načteme data, vybereme pouze proměnnou `length.R`, se kterou budeme dále pracovat, a uložíme si muže a ženy do separátních proměnných. Nezapomeňte zkontrolovat přítomnost NA hodnot.

```
data <- ...
muzi <- data$length.R[data$sex == 'm']
zeny <- ...
```

(A)

Na hladině významnosti $\alpha = 0.05$ otestujte nulovou hypotézu o shodě délky klíční kosti na pravé straně u mužů a u žen. K testování použijte

- (1) klasický dvouvýběrový t -test,
- (2) dvouvýběrový t -test s Welchovou aproximací;
- (3) věrohodnostní test za předpokladu, že rozptyly σ_1^2 a σ_2^2 jsou neznámé ale shodné;
- (4) věrohodnostní test za předpokladu, že rozptyly σ_1^2 a σ_2^2 jsou neznámé a různé.

Testování proveďte pomocí (i) kritického oboru, (ii) intervalu spolehlivosti, (iii) p -hodnoty. Před testováním ověřte předpoklad normality a předpoklad shody rozptylů obou náhodných výběrů.

(B)

Vygreslete grafy 95% věrohodnostních empirických intervalů spolehlivosti za předpokladu, že

- (a) rozptyly σ_1^2 a σ_2^2 jsou neznámé ale shodné,
- (b) rozptyly σ_1^2 a σ_2^2 jsou neznámé a různé.

Testy předpokladu normality

Pro ověření normality náh. výběru lze použít Shapiro-Wilkův, Andersonův-Darlingův nebo Lillieforsův test normality (oba později jmenované jsou v knihovně `nortest`). Test musíme provést zvlášť pro muže a pro ženy.

Pozn.: Rozsah 50 je na hraně kritérií pro použití testů.


```
nortest::ad.test(muzi)
```

Anderson-Darling normality test

```
data: muzi  
A = 0.36157, p-value = 0.4316
```

```
nortest::ad.test(zeny)
```

Anderson-Darling normality test

```
data: zeny  
A = 0.90842, p-value = 0.01908
```

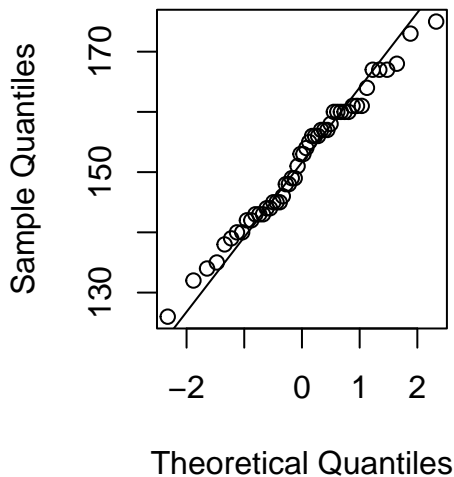
```
nortest::lillie.test(zeny)
```

Lilliefors (Kolmogorov-Smirnov) normality test

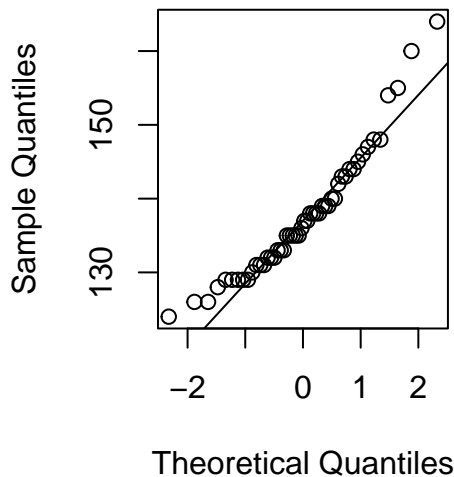
```
data: zeny  
D = 0.11944, p-value = 0.07197
```

Ověření normality graficky

Delka klicni kosti – muzi



Delka klicni kosti – zeny



Pozn.: Normalita byla pro ženy Anderson-Darlingovým testem zamítnuta, pro testování hypotézy bychom tedy měli zvážit použití neparametrických testů. Z kvantil-kvantilového grafu (a výsledku Lillieforsova testu) se ale zdá, že porušení není zásadní, proto použijeme pro účely výuky testy parametrické.

Ověření shodnosti rozptylů

Pomocí funkce `var.test()`

```
var.test(muzi, zeny)
```

```
F test to compare two variances
```

```
data: muzi and zeny
```

```
F = 1.6116, num df = 49, denom df = 49, p-value = 0.09812
```

```
alternative hypothesis: true ratio of variances is not equal to 1
```

```
95 percent confidence interval:
```

```
0.9145385 2.8399240
```

```
sample estimates:
```

```
ratio of variances
      1.611589
```

Výpočty pro případ (1) a (2)

Hranice IS pro (1) a (2):

$$DH = \bar{x}_1 - \bar{x}_2 - t_{df}(1 - \alpha/2)S_D, HH = \bar{x}_1 - \bar{x}_2 - t_{df}(\alpha/2)S_D$$

Pozn.: Nejprve je potřeba spočítat průměry (m_1 a m_2) a směrodatné odchytky - tentokrát je narozdíl od předchozího příkladu počítáme přímo z dat.

```
tW_Wald <- ...
df_Wald <- ...
W_hh_Wald <- qt(alpha / 2, df_Wald)
W_dh_Wald <- ...

dh_Wald <- m1 - m2 - qt(1 - alpha / 2, df_Wald) * sD_Wald
hh_Wald <- ...

p_Wald <- 2 * (1 - pt(abs(tW_Wald), df_Wald))

... # analogicky pro Welchovu aproximaci

## pro klasicky Walduv test lze pouzít i fci t.test()
## s nastavením var.equal = T,
## pro Welchovu aproximaci potom var.equal = F
```

Případ (3) a (4)

Věrohodnostní IS budeme chtít tentokrát počítat přesně, tj. využitím funkce `uniroot`, která bude hledat kořeny rovnice $U_{LR} - \chi_1^2(1 - \alpha)$. Proto si přichystáme funkci `ULRchisq`, která nám pro dané vektory, hodnotu μ_0 a příslušnou hladinu významnosti spočítá hodnotu $U_{LR} - \chi_1^2(1 - \alpha)$.

Pozn.: Funkci připravíme pro oba případy (stejně i různé rozptyly), postup viz další slide.

```
n <- ...
uLR_SR <- ...
W_dh_uLR_SR <- qchisq(1 - alpha, df = 1)
```

```

dh_uLR_SR <- uniroot(ULRchisq, c(0, 17), alpha = alpha, X = muzi, Y = zeny, equal = T)$root
hh_uLR_SR <- uniroot(ULRchisq, c(17, 30), ...)$root

p_uLR_SR <- 1 - pchisq(uLR_SR, 1)

... # analogicky pro ULR pri nestejnem rozptylu

```

Funkce ULRchisq

```

ULRchisq <- function(mu0, X, Y,
                    alpha = 0.05, equal = T) {
  ... # zakl. charakteristiky vektoru
  ... # pomocne vypocty vstupujici do vzorcu
  sD_Wald <- ...
  tW_Wald <- ...

  if(equal == T) {
    uLR <- ...} # ULR pro stejne rozptyly
  if(equal == F) {
    uLR <- ... # ULR pro ruzne rozptyly
  }
  uLR_chisq <- uLR - qchisq(1 - alpha, 1)
  return(uLR_chisq)
}

```

Výsledky

	m1	m2	stat	hh_W	dh_W	dh_IS	hh_IS	p
(1)	151.74	137.68	7.102	-1.984	1.984	10.131	17.989	0
(2)	151.74	137.68	7.102	-1.986	1.986	10.129	17.991	0
(3)	151.74	137.68	41.520	NA	3.841	10.182	17.938	0
(4)	151.74	137.68	51.467	NA	3.841	10.219	17.901	0

Vyhodnocení: Na základě jsme rozhodli, že H_0 zamítáme/nezamítáme.

Interpretace: Mezi délkou klíční kosti na pravé straně u mužů a žen existuje/neexistuje statisticky významný rozdíl.

(B) - vykreslení věrohodnostních IS

Postup je analogický jednovýběrovému případu. Volíme dostatečně hustou sekvenci bodů pro vykreslení IS, pro niž vypočítáme hodnoty ULR a zvýrazníme oblasti uvnitř IS:

```
mu0i <- seq(0, 30, length = 10000)
plot(mu0i, ..., # prislusne ULR
     type = 'l', ylim = c(0, 15), las = 1,
     xlab = '', ylab = expression(U[LR]))
lines(mu0i[mu0i > dh_uLR_SR & mu0i < hh_uLR_SR],
      uLR_SR_i[...], col = ..., lwd = 2)
## totez pro ruzne rozptyly
```

Výsledné grafy

