Test o pravděpodobnosti

Mgr. Zdeňka Geršlová

Příklad 1

Simulační studie rozdělení testovací statistiky pro test o p

Na základě simulační studie prověřte, že pokud má náhodná proměnná X asymptoticky binomické rozdělení $\text{Bin}(N, p)$, potom testovací statistika

$$Z_W = \frac{X/N - p}{\sqrt{p(1-p)/N}}$$

má asymptoticky normální rozdělení $N(0, 1)$.

Použijte (a) $p = 0.1$; (b) $p = 0.5$; (c) $p = 0.9$, a $N = 5$, 10, 15, 20, 25, 50, 100 a 200. Okomentujte výsledky ve spojitosti s Haldovou podmínkou $Np(1-p) > 9$.

Pro každou simulaci X vypočítejte $z_{W,m}$, $m = 1, 2, \dots, M$, kde $M = 1000$. Histogram vygenerovaných testovacích statistik v relativní škále superponujte s teoretickou křivkou hustoty Z_W.

Zhodnoťte, zda je Haldova podmínka dobrým ukazatelem situací, kdy je možné rozdělení testovací statistiky Z_W aproximovat normálním rozdělením $N(0, 1)$, i situací, kdy to naopak není vhodné.

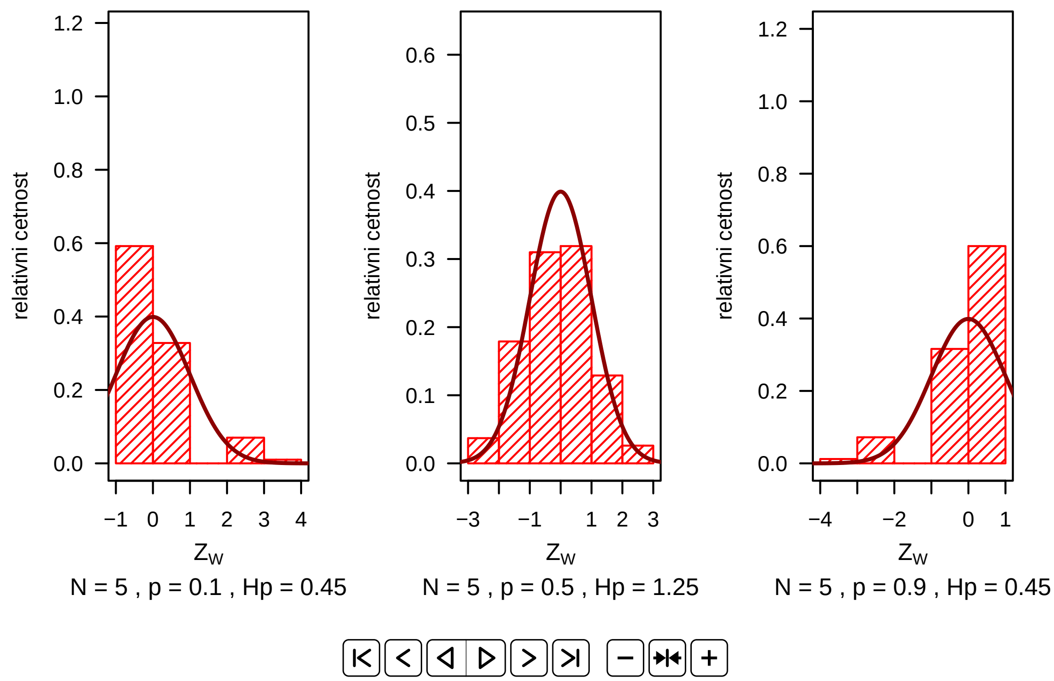
Funkce HistBinom

Vytvoříme funkci, která pro zadané parametry binomického rozdělení a počet simulací M vytvoří histogramy superponované křivkou teoretického rozdělení, které budou obarvené podle ne/splnění Haldovy podmínky - pro nesplnění chceme barvu histogramu červenou, při splnění zelenou.

```
HistBinom <- function(N, p, M = 1000){
  X <- replicate(M, rbinom(1, N, p))
  zW   <- ... # test. statistika
  xfit <- seq(min(zW) - 2, max(zW) + 2, length = 512)
```

```r
  yfit <- ... # teoreticka hustota
  d    <- max(yfit, hist(zW, plot = F)$density)
  Hp   <- ... # Haldova podm.

  ## nastaveni barvy podle Hp
    if (Hp < 9) {
   c_hist <- "red"
   c_lines <- "darkred"
 }
  if (Hp >= 9) {
    ...
 }

  hist(..., col = c_hist, ylim = c(0, d), ...,
       breaks = seq(floor(min(zW)), ceiling(max(zW)), by = 1) )
  # zaokrouhleni hodnot na nejblizsi nizsi/vyssi cele cislo
  # pro automaticke nastaveni hranic tridicich intervalu

  lines(..., col = c_lines, lwd = 2)
  ... # pridani hodnoty Hp do popisku

}
```

Výsledná simulace pro jednotlivá p



Závěr

Z animace vidíme, že Haldova podmínka je dobrým ukazatelem situací, kdy je možné rozdělení testovací statistiky Z_W aproximovat normálním rozdělením $N(0, 1)$.

Animace ukazují, že pro nižší rozsahy není aproximace normálním rozdělením dostatečně kvalitní. Pro $p = 0.5$, potřebujeme nejmenší minimální rozsah náhodného výběru, čím více se s p blížíme k 0, či 1, tím vyšší minimální rozsah náhodného výběru potřebujeme, aby byla Haldova podmínka dobré aproximace splněna a aproximace rozdělení testovací statistiky Z_W normálním rozdělením byla dostatečně kvalitní.

Příklad 2

Pravděpodobnost pokrytí Waldova DIS

Nechť $X \sim \text{Bin}(N, p)$, kde $N = 30$ a $p = 0.8$. Při experimentu bylo z $N = 30$ pokusů zanamenáno 24 úspěchů, tj. $x = 24$. Odhad pravděpodobnosti úspěchu $\hat{p} = \frac{24}{30} = 0.8$. Waldův 95 % empirický DIS pro p je rovný $(d, h) = (0.657, 0.943)$. Vypočítejte pravděpodobnost pokrytí tohoto intervalu.

Pravděpodobnost pokrytí Waldova 95 % DIS pro p vypočítáme následovně:

$$\Pr(pokryti) = \sum_j \Pr(X = Np_j : p \in \text{ Waldovu 95 \% DIS pro } p_j),$$

kde $p_j \in \mathcal{M}_J = \{\frac{1}{30}, \frac{2}{30}, \dots, 1 - \frac{1}{30}\}$, t.j. jde o součet takových funkčních hodnot pravděpodobnostní funkce v bodech Np_j, kde $p \in$ Waldovu 95 % DIS pro p_j. Výsledky uspořádejte do tabulky, jejíž sloupce budou x_j, p_j, dh_j (dolní hranice Waldova 95 % DIS pro p_j), hh_j (horní hranice Waldova 95 % DIS pro p_j), $\Pr(X = Np_j)$ a $p \in IS_j$ (indikace toho, zda p náleží nebo nenáleží do Waldova 95 % DIS pro p_j).

Celý postup následně zopakujte, tentokrát pro parametr $p = 0.79$. Pozorujte, jak se změnila pravděpodobnost pokrytí.

Hranice Waldova IS

Waldův $(1-\alpha)100\%$ DIS pro p :

$$\left(\frac{x}{N} - u_{1-\alpha/2}\sqrt{\frac{p(1-p)}{N}}, \frac{x}{N} - u_{\alpha/2}\sqrt{\frac{p(1-p)}{N}}\right)$$

Pozn.: Při výpočtu pravděpodobnosti pokrytí využíváme toho, že počet úspěchů v N pokusech je konečný.

Funkce PokrytiP

Vytvoříme funkci, která na základě parametrů binomického rozdělení a hladiny významnosti vypočítá pravděpodobnost pokrytí pro daný typ test. statistiky (v tomto příkladu Waldova, v následujícím přidáme skóre a věrohodnostní). Na výstupu funkce chceme mít jednak konkrétní pravděpodobnost pokrytí (jednu hodnotu pro zadané parametry), jednak tabulku, v níž budou jednotlivá x_j, p_j, dh_j, hh_j, $\Pr(X = Np_j)$ a $p \in IS_j$.

```
PokrytiP <- function(p, N, alpha = 0.05, type = 'wald'){
  xj <- ... # posloupnost celych cisel od 1 do (N - 1)
  pj <- ... # vektor pravdepodobnosti pj

  if (type == 'wald'){
    dhj <- ... # dolni hranice Waldova DIS pro dane xj a pj
    hhj <- ... # horni hranice Waldova DIS pro dane xj a pj
  }

  pr_x_Npj <- dbinom(xj, N, p)
```

```
  # pravdep., ze pocet uspechu je roven xj za predpokladu, ze se ridime Bin(N,p)
  p_in_ISj  <- (p > dhj & p < hhj) * 1
  # indikacni funkce, ktera urcuje, zda odhad p patri ci nepatri do Waldova DIS
  # chceme ciselnou hodnotu, ne TRUE/FALSE, proto nasobeni 1
  tab      <- data.frame(...)
  pst_pokryti <- ... # vypocet pst pokryti podle vzorce
  return(list(tab = tab, pst_pokryti = pst_pokryti))
}
```

Výsledná tabulka pro $p = 0.8$

Aktuální pravděpodobnost pokrytí $1 - \hat{\alpha}$ je rovna 0.9463279.

Výpis tabulky:

```
tab <- PokrytiP(p = 0.8, N = 30, type = 'wald')$tab
kable(tab, digits = 4, longtable = T)
```

xj	pj	dhj	hhj	pst	p_in_ISj
1	0.0333	-0.0309	0.0976	0.0000	0
2	0.0667	-0.0226	0.1559	0.0000	0
3	0.1000	-0.0074	0.2074	0.0000	0
4	0.1333	0.0117	0.2550	0.0000	0
5	0.1667	0.0333	0.3000	0.0000	0
6	0.2000	0.0569	0.3431	0.0000	0
7	0.2333	0.0820	0.3847	0.0000	0
8	0.2667	0.1084	0.4249	0.0000	0
9	0.3000	0.1360	0.4640	0.0000	0
10	0.3333	0.1646	0.5020	0.0000	0
11	0.3667	0.1942	0.5391	0.0000	0
12	0.4000	0.2247	0.5753	0.0000	0
13	0.4333	0.2560	0.6107	0.0000	0
14	0.4667	0.2881	0.6452	0.0000	0
15	0.5000	0.3211	0.6789	0.0002	0
16	0.5333	0.3548	0.7119	0.0007	0
17	0.5667	0.3893	0.7440	0.0022	0
18	0.6000	0.4247	0.7753	0.0064	0
19	0.6333	0.4609	0.8058	0.0161	1
20	0.6667	0.4980	0.8354	0.0355	1
21	0.7000	0.5360	0.8640	0.0676	1

xj	pj	dhj	hhj	pst	p_in_ISj
22	0.7333	0.5751	0.8916	0.1106	1
23	0.7667	0.6153	0.9180	0.1538	1
24	0.8000	0.6569	0.9431	0.1795	1
25	0.8333	0.7000	0.9667	0.1723	1
26	0.8667	0.7450	0.9883	0.1325	1
27	0.9000	0.7926	1.0074	0.0785	1
28	0.9333	0.8441	1.0226	0.0337	0
29	0.9667	0.9024	1.0309	0.0093	0

Výsledky pro $p = 0.79$

Aktuální pravděpodobnost pokrytí $1 - \hat{\alpha}$ je rovna 0.8875662.

xj	pj	dhj	hhj	pst	p_in_ISj
1	0.0333333	-0.0309007	0.0975674	0.0000000	0
2	0.0666667	-0.0225940	0.1559274	0.0000000	0
3	0.1000000	-0.0073516	0.2073516	0.0000000	0
4	0.1333333	0.0116915	0.2549751	0.0000000	0
5	0.1666667	0.0333080	0.3000253	0.0000000	0
6	0.2000000	0.0568645	0.3431355	0.0000000	0
7	0.2333333	0.0819845	0.3846822	0.0000000	0
8	0.2666667	0.1084244	0.4249090	0.0000000	0
9	0.3000000	0.1360176	0.4639824	0.0000000	0
10	0.3333333	0.1646465	0.5020202	0.0000000	0
11	0.3666667	0.1942261	0.5391072	0.0000002	0
12	0.4000000	0.2246955	0.5753045	0.0000016	0
13	0.4333333	0.2560114	0.6106552	0.0000086	0
14	0.4666667	0.2881453	0.6451880	0.0000419	0
15	0.5000000	0.3210806	0.6789194	0.0001788	0
16	0.5333333	0.3548120	0.7118547	0.0006706	0
17	0.5666667	0.3893448	0.7439886	0.0022092	0
18	0.6000000	0.4246955	0.7753045	0.0063821	0
19	0.6333333	0.4608928	0.8057739	0.0161231	1
20	0.6666667	0.4979798	0.8353535	0.0354709	1
21	0.7000000	0.5360176	0.8639824	0.0675636	1
22	0.7333333	0.5750910	0.8915756	0.1105586	1
23	0.7666667	0.6153178	0.9180155	0.1538207	1
24	0.8000000	0.6568645	0.9431355	0.1794575	1
25	0.8333333	0.6999747	0.9666920	0.1722792	1

xj	pj	dhj	hhj	pst	p_in_ISj
26	0.8666667	0.7450249	0.9883085	0.1325224	1
27	0.9000000	0.7926484	1.0073516	0.0785318	1
28	0.9333333	0.8440726	1.0225940	0.0336565	0
29	0.9666667	0.9024326	1.0309007	0.0092846	0

Komentář

Příklad ilustruje úskalí aktuální pravděpodobnosti pokrytí 95% Waldova empirického DIS. Pokud náhodná veličina $X \sim \text{Bin}(30, 0.8)$ a při experimentu dojde k zaznamenání 24 úspěchů ze 30, je aktuální pravděpodobnost pokrytí $(1 - \hat{\alpha}) = 0.9463 \doteq 0.95$.

Ovšem pokud náhodná veličina $X \sim \text{Bin}(30, 0.79)$ a při experimentu dojde k zaznamenání 24 úspěchů ze 30 (což by reálně s vysokou pravděpodobností skutečně nastalo), klesne aktuální pravděpodobnost pokrytí $(1 - \hat{\alpha})$ na hodnotu $0.8876 \doteq 0.89$, tj. aktuální pravděpodobnost pokrytí je o 0.06 nižší než nominální pravděpodobnost pokrytí.

Příklad 3

Pravděpodobnost pokrytí Waldova, skóre a věrohodnostního DIS
Nechť $X_i \sim \text{Bin}(N, p_i)$, $p_i \in \{0.001, 0.003, \dots, 0.997, 0.999\}$. Vypočítejte pravděpodobnosti pokrytí (a) Waldova, (b) skóre, (c) věrohodnostního 95 % empirického oboustranného intervalu spolehlivosti pro každé p_i. Nakreslete obrázek, kde na ose x budou p_i a na ose y bude pravděpodobnost pokrytí $\text{Pr}_i(\text{pokrytí})$. Zvolte (i) $N = 30$, (ii) $N = 100$, (iii) $N = 500$.

Pozn.: Pravděpodobnosti pokrytí 95 % DIS pro p_i vypočítáme jako

$$\text{Pr}_i(\text{pokrytí}) = \sum_j \text{Pr}(X = Np_j : p_i \in 95\,\%\ \text{DIS pro } p_j),$$

kde $p_j \in \mathcal{M}_J = \{\frac{1}{N}, \frac{2}{N}, \dots, 1 - \frac{1}{N}\}$, t.j. jde o součet takových funkčních hodnot pravděpodobnostní funkce v bodech Np_j, kde $p_i \in 95\,\%$ DIS pro p_j.

Zopakujte si definici konzervativního a liberálního IS a zamyslete se, zda je (a) Waldův; (b) skóre; (c) věrohodnostní DIS pro parametr p konzervativní, liberální, nebo ani jedno. Dále všechny tři DIS vzájemně porovnejte a uveďte, který DIS je z hlediska pravděpodobnosti pokrytí nejvhodnějším intervalovým odhadem parametru p a který je naopak nejméně vhodný.

Vzorce pro skóre a věrohodnostní IS

Explicitní vyjádření hranic skóre IS:

$$p\frac{N}{N+u_{\alpha/2}^2}+\frac{1}{2}\frac{u_{\alpha/2}^2}{N+u_{\alpha/2}^2}\pm u_{\alpha/2}\sqrt{\frac{1}{N+u_{\alpha/2}^2}\left[p(1-p)\frac{N}{N+u_{\alpha/2}^2}+\frac{1}{4}\frac{u_{\alpha/2}^2}{N+u_{\alpha/2}^2}\right]}$$

Věrohodnostní test. statistika

$$u_{LR}=2\left(x\ln\frac{x}{Np_0}+(N-x)\ln\frac{N-x}{N-Np_0}\right)$$

a $100(1-\alpha)\%$ věrohodnostní DIS počítáme opět jako

$$\{p_0:U_{LR}\leq\chi_1^2(\alpha)\}$$

Doplnění funkce PokrytiP

Do těla funkce vytvořené v př. 2 nyní přidáme výpočet hranic DIS pro případ (b) a (c), pro případ (c) převedeme nerovnost $U_{LR}\leq\chi_1^2(\alpha)$ na nerovnost $U_{LR}-\chi_1^2(\alpha)\leq 0$ a využijeme funkce `uniroot()` pro nalezení kořenů rovnice.

```
if(type == 's'){
    dhj <- ... # dolni hranice skore DIS pro pj
    hhj <- ... # horni hranice skore DIS pro pj
  }

  if(type == 'lr'){
    ## funkce pro vypocet u_LR statistiky s odectenim chi kvadratu
    ULRchisq <- function(x, N, p, alpha){
      uLR <- ... # vzorec pro u_LR
      # pozn. p0 ve vzorci je "nase" p ze vstupnich parametru rozdeleni
      uLR_chisq <- uLR - qchisq(1 - alpha, 1)
      return(uLR_chisq)}

    ## vypocet dh a hh verohodnostniho IS pomoci uniroot
    dhj <- hhj  <- NULL
    for(j in 1:length(pj)){
      dhj[j] <- uniroot(ULRchisq, c(1e-6, pj[j]), x = xj[j], N = N,
                        alpha = alpha)$root
```

```
      hhj[j] <- uniroot(ULRchisq, c(pj[j], 1-1e-6), x = xj[j], N = N,
                        alpha = alpha)$root}
  }
```

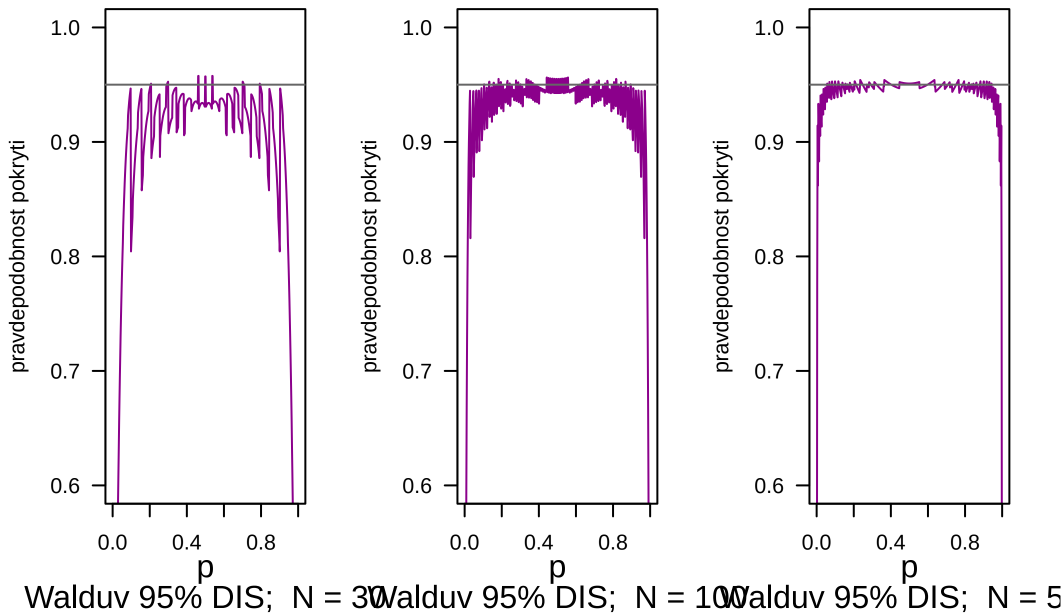
Funkce PlotPokrytiP

Konečně vytvoříme funkci `PlotPokrytiP`, která nám výsledky vykreslí (příp. lze upravit původní funkci přidáním argumentu plot = T/F)

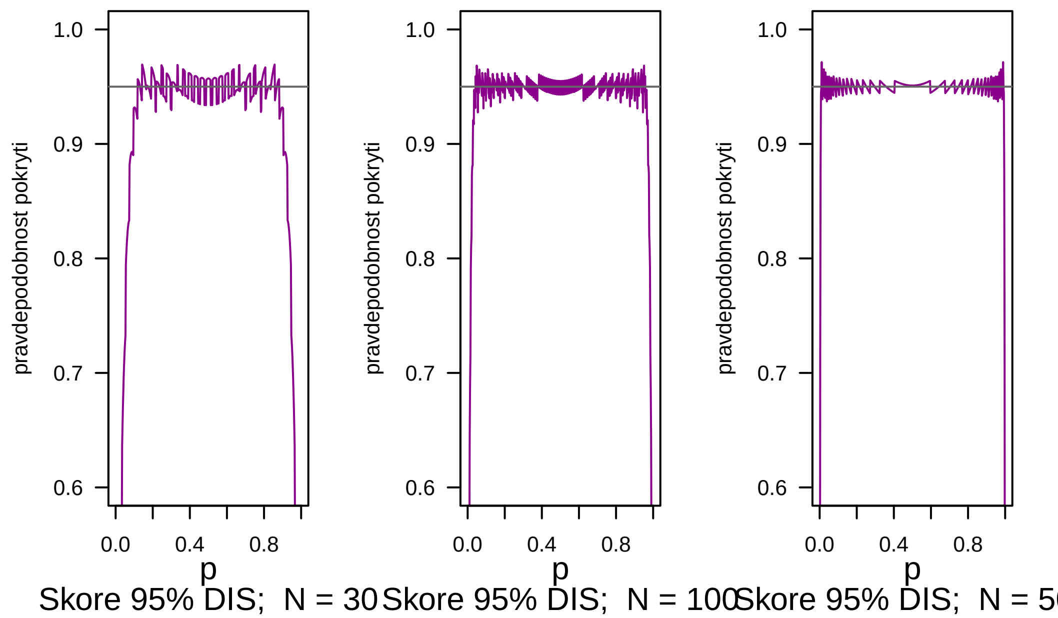
```
PlotPokrytiP <- function(N, pst, type, alpha = 0.05){
  pokryti <- NULL
  for(i in 1:length(pst)){
    pokryti[i] <- PokrytiP(pst[i], N = N, alpha = alpha, type = type)$pst_pokryti
  }

  # vytvoreni odpovidajicich popisku
  if (type == "wald") {
    text <- bquote(paste("Walduv ", .((1 - alpha) * 100), "% DIS;  N = ", .(N)))
  }
  if (type == "s") {
    text <- ...
  }
  if (type == "lr") {
    text <- ...
  }

  par(mar = c(5, 4, 2, 2))
  plot(..., type = 'l', ylim = c(0.6, 1),
       ...) # vykreslete graf, kde na ose x je sekvence pravdepodobnosti
  # a na ose y prislusne pokryti
  ...
  mtext(text, ...) # pridani vytvorenych popisku pod osu x
  ... # horizontalni cara v nominalni hladine vyznamnosti
}
```

Při vytváření grafů potom za argument pst vkládáme sekvenci od 0.001 do 0.999 s dostatečně malým krokem (např. 0.002 nebo 0.005, čím jemnější dělení, tím delší doba výpočtu).

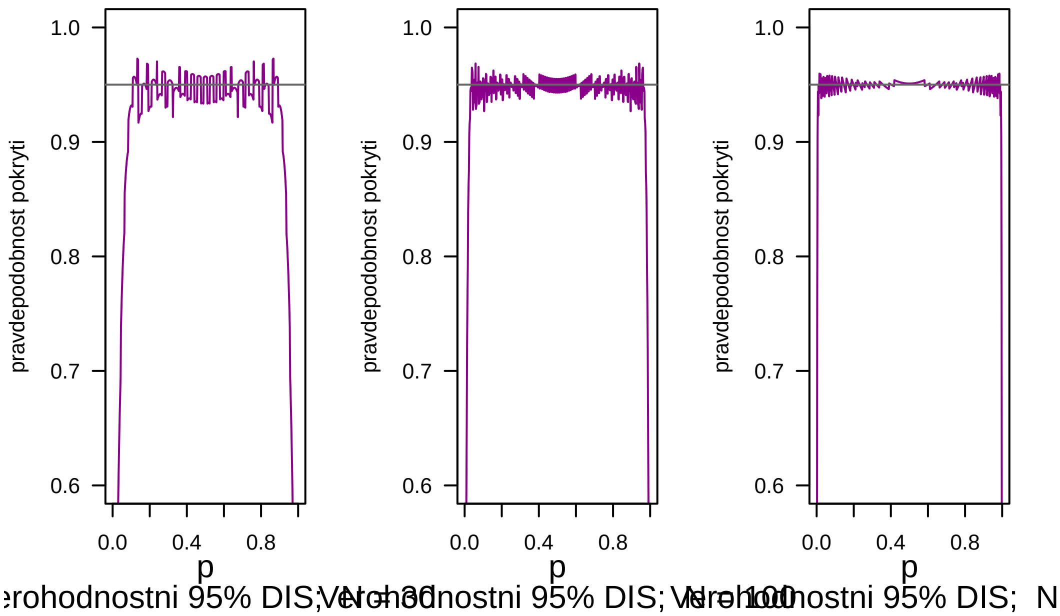
Graf pro Waldův DIS



Graf pro skóre DIS



Graf pro věrohodnostní DIS



Zhodnocení

Waldův 95% DIS pro $N = 5$ ve většině případů z hlediska pravděpodobnosti pokrytí naprosto selhává, aktuální pst pokrytí je nižší než nominální pst pokrytí a DIS je silně liberální. Pro rostoucí N se situace lepší, DIS nicméně povětšinou zůstává liberální, navíc výrazně selhává při nízkých hodnotách parametru p, zejména pro $p < 0.1$ nebo $p > 0.9$.

Skóre 95% DIS pro $N = 5$ dává dobré výsledky (není ani konzervativní ani liberální) pro $p \in (0.2; 0.8)$. Pro cca $p < 0.2$ a $p > 0.8$ je liberální. S rostoucím N však jeho kvalita velmi rychle roste, pro $N = 100$ funguje skvěle pro cca $p \in (0.05; 0.95)$, pro $N = 500$ funguje skvěle de facto na celém intervalu $(0; 1)$. Ze všech DIS je z hlediska pravděpodobnosti pokrytí nejlepší.

Věrohodnostní 95% DIS pro $N = 5$ dává dobré výsledky (není ani konzervativní ani liberální) pro cca $p \in (0.2; 0.8)$. Pro $p < 0.2$ a $p > 0.8$ je liberální. S rostoucím N však jeho kvalita velmi rychle roste, pro $N = 100$ funguje skvěle pro cca $p \in (0.1; 0.9)$, pro $N = 500$ funguje skvěle pro cca $p \in (0.02; 0.98)$.

Praktický příklad

Test o pravděodobnosti p

Mějme datový soubor `one-sample-probability-dermatoglyphs.txt` obsahující údaje o frekvenci výskytu dermatoglyfických vzorů *vír* (whorl), *smyčka* (loop) a *oblouček* (arch) na deseti prstech 470 jedinců (235 mužů a 235 žen) populace Bagathů z Araku Valley. Současně máme k dispozici hodnotu pravděpodobnosti výskytu dermatoglyfického vzoru *smyčka* u jedinců z populace Lambadis ($p_m = 0.5618$, $p_f = 0.6233$).

(A) Test H_0

Na hladině významnosti $\alpha = 0.05$ zjistěte, zda existuje rozdíl mezi frekvencemi výskytu dermatoglyfického vzoru *smyčka* u mužů bagathské populace z Araku Valley a mužů z populace Lambadis. Před testováním ověřte splnění Haldovy podmínky dobré aproximace.
Testování proveďte pomocí

(a) kritického oboru;
(b) intervalu spolehlivosti;
(c) p-hodnoty

při použití testovacích statistik
(1) $Z_W \sim N(0,1)$,
(2) $U_S = \dfrac{(p-p_0)^2}{p_0(1-p_0)/N} \sim \chi^2_\alpha(1)$,
(3) $U_{LR} \sim \chi^2_\alpha(1)$.

U každého typu testování (a)–(c) pro každý test (1)–(3) uveďte zdůvodněný závěr o H_0 a nakonec uveďte antopologickou interpretaci výsledku testování.

Pozn.: Při testování H_0 ověřujeme podmínku dobré aproximace ve tvaru $Np(1-p) > 9$, kde za p dosazujeme hodnotu p_0 z nulové hypotézy H_0.

(B) Zobrazení IS a sloupcového grafu

- Vykreslete graf zobrazující hranice a oblast 95 % empirického věrohodnostního intervalu spolehlivosti pro test o parametru p.
- Nakonec vykreslete sloupcový diagram relativních četností výskytu vzoru *smyčka* v populaci bagathských mužů z Araku Valley.

Načtení dat, ověření podmínek

- načteme datový soubor `one-sample-probability-dermatoglyphs.txt` - data jsou ve formě tabulky, kde máme frekvenci výskytu daného vzoru (sloupec `pattern`) u mužů (sloupec `m`) a žen (sloupec `f`)
- podle vzorce spočítáme Haldovu podmínku

Hodnota $Np_0(1 - p_0) = 578.5248$, takže podmínka dobré aproximace je/není splněna.

Výpočet statistik, kritických hodnot, IS a p-hodnot

- na základě vzorců a znalosti asymptotického rozdělení test. statistik spočítáme jednotlivé statistiky a IS (ty už jsme programovali v předchozím příkladu, pro výpočet věrohodnostních IS opět použijeme funkci `uniroot`), doplníme kritické hodnoty a p-hodnoty a vše sestavíme do přehledné tabulky

Výsledky

	p	stat	wd	wh	dh	hh	p.1
Z_W	0.53	-3.068	-1.96	1.960	0.51	0.55	0.001
U_S	0.53	9.524	NA	3.841	0.51	0.55	0.002
U_{LR}	0.53	9.481	NA	3.841	0.51	0.55	0.002

Vyhodnocení: Protože hodnota test. statistiky …. je/není uvnitř kritického oboru ….., zamítáme/nezamítáme H_0 o tom, že pravděpodobnost výskytu vzoru smyčka u mužů bagathské populace je stejná jako u mužů populace Lambadis.

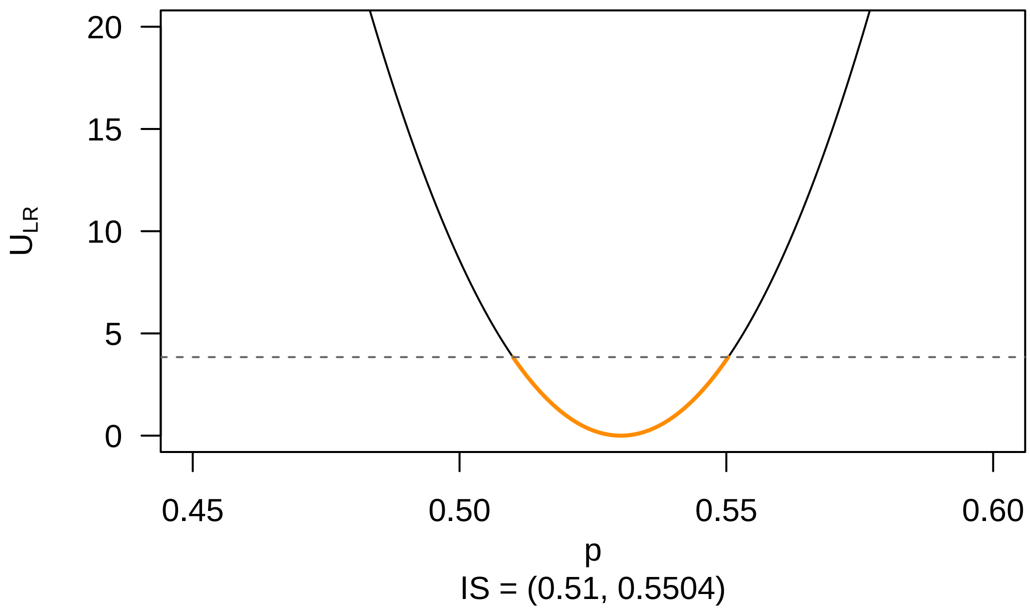
Protože hodnota …. spadá/nespadá do IS: ……………………………., zamítáme/nezamítáme H_0.

Protože p-hodnota ………… je větší/menší než ………….., zamítáme/nezamítáme H_0.

Na hladině významnosti 0.05 jsme tedy prokázali/neprokázali statisticky významný rozdíl v pravděpodobnosti výskytu vzoru smyčka u mužů populace Bagathů a Lambadis.

Vykreslení IS

- pro vykreslení věrohodnostního IS použijeme sekvenci p_i od 0.45 do 0.6 dostatečné délky, postupujeme stejně jako v předchozích cvičeních



Sloupcový diagram - postup

Jedním ze způsobů, jak vykreslit sloupcový diagram relativních četností, je vypočítané rel. četnosti uložit do matice (o rozměrech $k \times 1$, kde k je počet kategorií, v našem případě $k = 2$, protože chceme porovnat výskyt vzoru smyčka vs. jiných vzorů) a vykreslit pomocí funkce `barplot`

```
barplot(matrix(c(x/N, (N - x)/N), 2,1),
        col = ...,
        border = ...,
        xlim = c(0.2, 2.4),
        ... # nastaveni dalsich argumentu jako popisy os, density apod.
        )
legend(...) # pridani legendy k barplotu

## pridani linky do hodnoty pravdepodobnosti pop. lambadis
segments(0.2, p0, 1.2, p0, col = 'red', lwd = 2)
text(1.62, p0, bquote(paste(p[0] == .(p0))), col = 'red')
... # pridani hodnot jednotlivych skupin pomoci mtext
```

Výsledný graf