Testy o korelačním koeficientu

Mgr. Zdeňka Geršlová

Příklad 1

Konvergence ρ a ξ k normálnímu rozdělení pro $n \to \infty$
Proveďte simulaci pseudonáhodných čísel z $N_2(,)$, kde $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 1$, $M = 1\,000$, $\rho = 0.8$. Pro každé $m = 1, 2, \ldots, M$, vypočítejte realizaci výběrového korelačního koeficientu r_m a Fisherovy Z-transformace $z_{R,m}$. Zobrazte histogramy simulovaných r_m a $z_{R,m}$ a superponujte je teoretickými hustotami příslušných normálních rozdělení. Vytvořte animaci zobrazující rozdělení výběrového korelačního koeficientu R a Fisherovy Z-transformace pro různé rozsahy náhodného výběru $n \in \{5, 10, 15, \ldots, 65, 70\}$.

Porovnejte kvalitu konvergence R a Z_R k normálnímu rozdělení pro $n \to \infty$.

Příklad 2

Konvergence ρ a ξ k normálnímu rozdělení pro $\rho \to 1$

Vytvořte animaci zobrazující konvergenci rozdělení výběrového korelačního koeficientu R a Fisherovy Z-transformace k normálnímu rozdělení pro $\rho \to 0.9$. Hodnoty koeficientu ρ volte $\rho \in \{0.1, 0.2, \ldots, 0.9\}$, rozsah náhodného výběru zvolte (a) $n = 5$, (b) $n = 50$.

Zamyslete se nad změnou kvality konvergence R a Z_R k normálnímu rozdělení pro $\rho \to 0.9$ (je-li nějaká). Také navzájem porovnejte situace pro $n = 5$ a $n = 50$.

Postup

Příklady 1 a 2 vypracujeme společně. Vytvoříme funkci `HistRho`, která pro zadané parametry vykreslí histogramy výběrového korelačního koeficientu a Fisherovy Z-transformace superponované křivkou asymptotického rozdělení.

Pozn.: Protože tvoříme funkci pro oba příklady, musíme varianční matici dvourozměrného rozdělení sestavovat až uvnitř funkce, protože v příkladu 2 měníme hodnotu ρ.

```
HistRho <- function(n, rho, xlim1, xlim2, ylim1, ylim2, M = 1000,
                    mu, sigma1, sigma2){
  ## xlim1, xlim2, ylim1, ylim2 ... vstupni parametry pro zmenu rozsahu os
  ## histogramu: xlim1 = rozsah osy x prvniho histogramu atd.
  ## mu ... vektor strednich hodnot dvourozmer. norm. rozd.
  ## sigma1, sigma2 ... sd pro dvourozmer. norm. rozd.

  Sigma <- matrix(c(sigma1 ^ 2, rho * sigma1 * sigma2,
                    rho * sigma1 * sigma2, sigma2 ^ 2), 2, 2)

... # generovani dat
... # vypocet R a Z_R
... # posloupnosti pro vykresleni hustoty
... # histogram prolozeny krivkou hustoty
}
```

Generování dat

```
## prvni zpusob - pomoci for cyklu
## vygenerujeme vzdy jedno rozdeleni a spocitame R a Z_R
  R <- zR <- NULL
  for(i in 1:M) {
    XY <- MASS::mvrnorm(n, mu = mu, Sigma = Sigma)
    R[i] <- cor(XY[ ,1], XY[ ,2])
    zR[i] <- ... # vzorec pro Fisher. Z transformaci
  }

## pomoci replicate - ja osobne upřednostnuji
R <- replicate(M, cor(MASS::mvrnorm(n, mu = mu, Sigma = Sigma))[1, 2])
## vystupem z cor() aplikovaneho na mvnorm je cela korelacni matice,
## my chceme jen koeficient R, ktery je na pozici [1,2] (nebo [2,1])
## rovnou uz mame cely vektor
zR <- ... # vzorec pro Fisher. Z transformaci
```

Asymptoticky:

$R \sim N(\rho, \frac{(1-\rho^2)^2}{n-1})$,

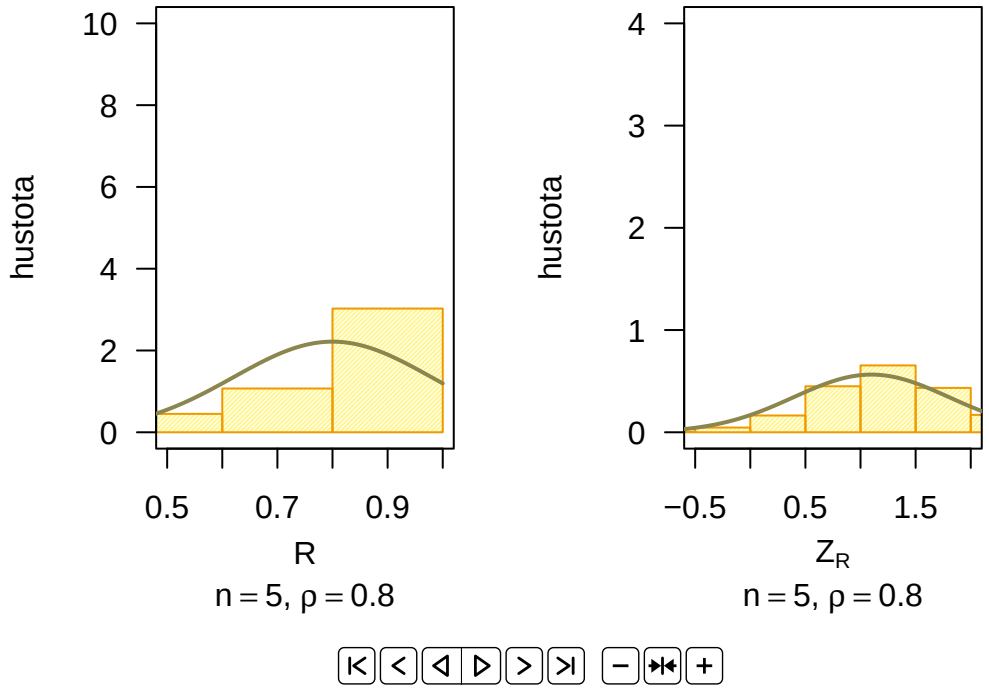
$Z_R = \frac{1}{2} \ln \frac{1+R}{1-R} \sim N(\frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \frac{1}{n-3})$

Vykreslení histogramu a křivky hustoty

```
## pro R i Z_R vykreslime histogram superponovany krivkou hustoty

xfit <- ... # sekvence pro vypocet hustoty
## pro R od -1 do 1, pro Z_R od min(Z_R - 1) do max(Z_R + 1)
yfit <- ... # hustota prislusneho rozdeleni

## parametry graf. okna:
par(mfrow = c(1, 2), mar = c(5, 4, 1, 2))

... # vykresleni histogramu R/Z_R
... # pridani krivky hustoty pomoci lines
```

Pro vytvoření animace potom definujeme příslušnou sekvenci n, resp. ρ a vykreslíme animaci pomocí `for` cyklu.

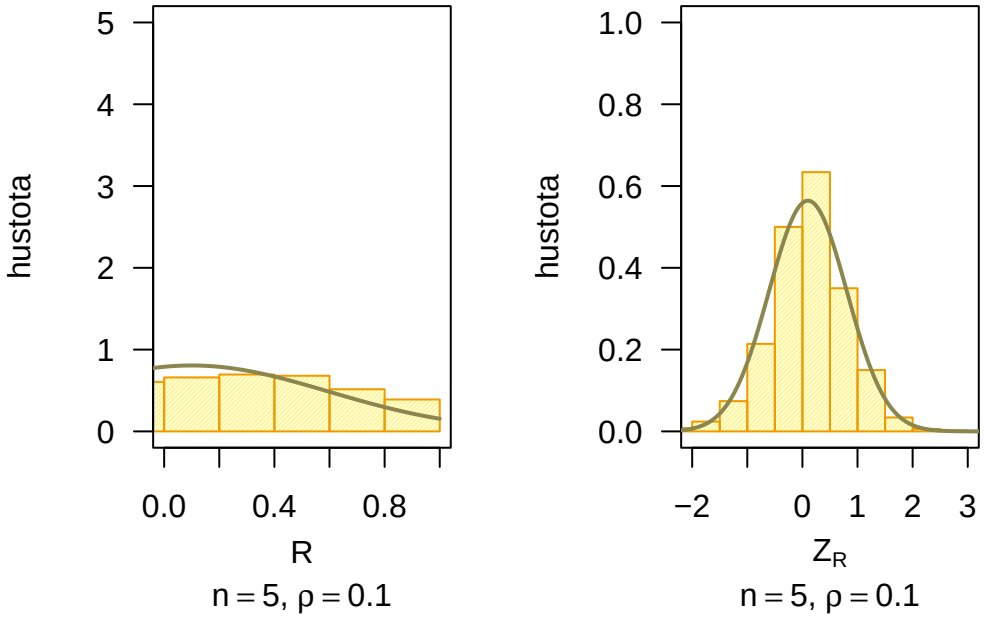
Výsledné animace pro příklad 1



Animace asymptotického rozdělení R a Z_R pro měnící se n

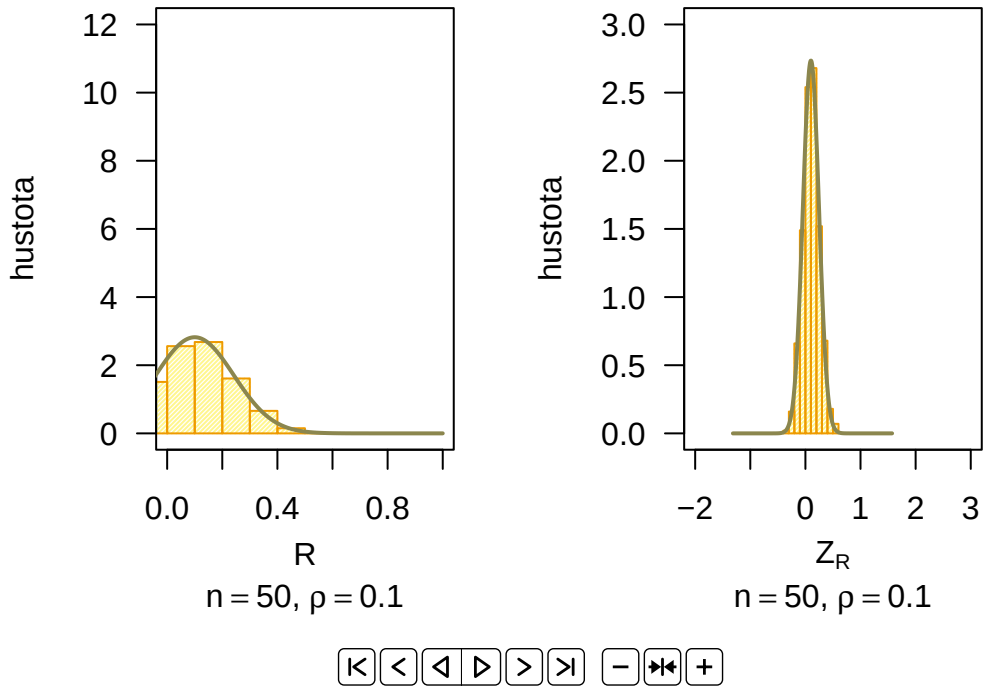
Všimněte si, že Fisherova Z-transformace konverguje k asymptotickému rozdělení rychleji.

Animace pro příklad 2 (n = 5)



Pro nízké rozsahy kvalita konvergence R není uspokojivá, pro Z_R se zhoršuje s rostoucím ρ.

Animace pro příklad 2 (n = 50)



Praktický příklad

Test o korelačním koeficientu ρ
Mějme datový soubor `one-sample-correlation-skull-mf.txt` obsahující údaje o největší výšce mozkovny `skull.pH` (v mm) a morfologické výšce tváře `face.H` (v mm) starověké egyptské mužské a ženské populace.

Na hladině významnosti $\alpha = 0.05$ testujte nulovou hypotézu o shodě korelačního koeficientu největší výšky mozkovny a morfologické výšky tváře u mužů s hodnotou 0.251 (hodnota korelačního koeficientu mezi stejnými proměnnými u novověké egyptské populace). Testování proveďte pomocí
(a) kritického oboru,
(b) intervalu spolehlivosti,
(c) p-hodnoty při použití
(1) Waldovy testovací statistiky Z_W,
(2) věrohodnostní testovací statistiky U_{LR}.

Dále vykreslete graf zobrazující 95 % věrohodnostní interval spolehlivosti pro korelační koeficient ρ získaný na základě U_{LR} testovací statistiky.

Načtení dat a test normality

Načtení dat provedeme klasickým způsobem, vybereme pouze muže. Pozor - soubor obsahuje množství NA hodnot, ty je nutné odstranit pomocí funkce `na.omit`.

Pro otestování předpokladu dvourozměrné normality použijeme testy z knihovny `MVN` - funkce `mvn`, typ testu měníme parametrem `mvnTest`. Často se používá Mardiův test, který testuje zvlášť šikmost a špičatost a dvourozměrnou normalitu zamítá už při porušení jedné z veličin.

Pozn.: Ve výstupu funkce ***mvn*** *kromě p-hodnoty testu najdete* YES/NO *- v tomto případě* YES *znamená, že daná testovaná veličina nevykazuje odchylku od normálního rozdělení.*

```
MVN::mvn(data.M, mvnTest = 'mardia')$multi
```

```
             Test         Statistic           p value Result
1 Mardia Skewness  4.15619988627931 0.385279036110575    YES
2 Mardia Kurtosis -1.28198025744202 0.199849571904474    YES
3             MVN              <NA>              <NA>    YES
```

```
## analogicky zkuste dalsi dva testy:
## Henze-Zirkler: mvnTest = 'hz'
## Royston: mvnTest = 'royston'
```

Grafické ověření dvourozměrné normality

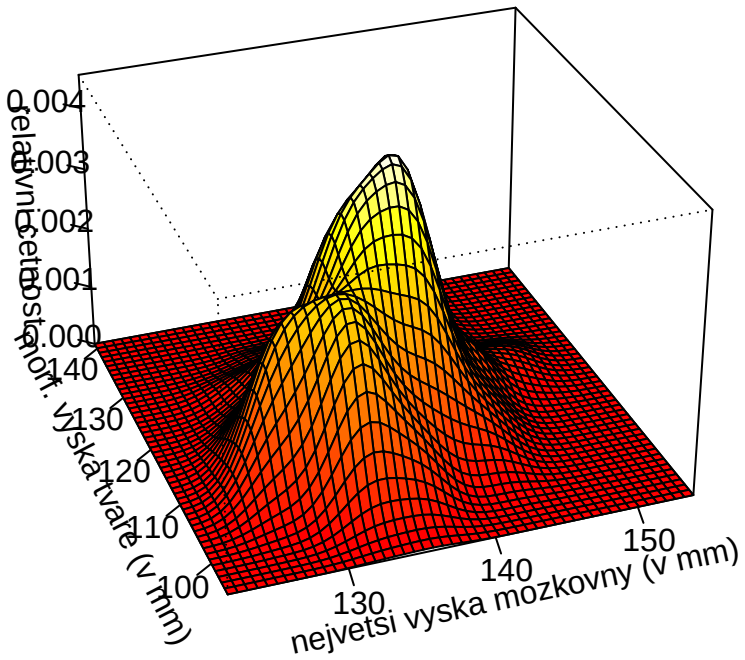
Pro 3D zobrazení dvourozměrné hustoty nejprve vypočítáme jádrový odhad hustoty pomocí funkce `kde2d` z knihovny `MASS` a poté ji zobrazíme pomocí funkce `persp`.

Pozn.: Z minulého semestru údajně znáte "trik" k získání středů intervalů pro dělení barevné škály, aby různé "výšky" hustoty barevně odpovídaly. Pokud nemáte požadavky na konkrétní počet intervalů, do kterých chcete škálu dělit, lze použít funkci ***persp3D*** *z knihovny* ***GA****.*

```
## prepocet stredu intervalu do skaly heat.colors
kd <- MASS::kde2d(skull.M, face.M, n = 50,
                  lims = c(min(skull.M) - 5, max(skull.M) + 5,
                           min(face.M) - 5, max(face.M) + 5))
x <- kd$x
y <- kd$y
Z <- kd$z
n <- dim(Z)[1]
stredy <- Z[-1, -1] + Z[-1, -n] + Z[-n , -1] + Z[-n , -n]
```

```r
vyska <- cut(stredy, 12)

par(mar = c(1, 1, 0, 1))
persp(x, y, Z, xlab = ..., ylab = ..., zlab = ...,
      theta = -20, phi = 30, col = heat.colors(12)[vyska])

## pouziti persp3D
par(mar = c(1, 1, 0, 1))
GA::persp3D(x, y, Z, xlab = ..., ylab = ..., zlab = ...,
      theta = -20, phi = 30, col.palette = heat.colors, border = 'black')
```

Graf dvourozměrné hustoty



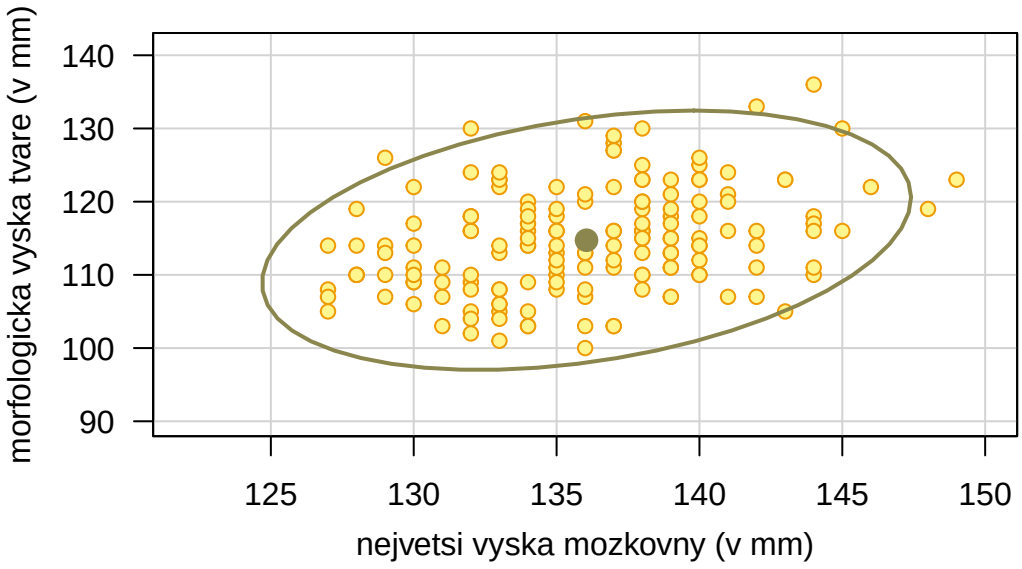
Elipsa spolehlivosti

Pro vykreslení elipsy spolehlivosti použijeme funkci `dataEllipse` z knihovny `car`. Pokud data pocházejí z dvourozměrného normálního rozdělení, pak 95% elipsa spolehlivosti by měla pokrývat alespoň 95% dat.

Pozn.: Prakticky ověřujeme tak, že vypočítáme 5% z rozsahu náhodného výběru a je-li počet bodů vně elipsy menší nebo roven tomuto číslu, považujeme předpoklad dvourozměrné normality za splněný.

```
par(mar = c(5, 4, 2, 1))
car::dataEllipse(skull.M, face.M, level = 0.95,
                 xlim = c(122, 150), # nastaveni o neco vice nez je range dat
                 ylim = c(90, 141),
                 xlab = '',
                 ylab = ...,
                 main = '', pch = ...,
                 col = c(...,...), # nejprve barva obrysu bodu, potom barva elipsy
                 bg = ..., # v pripade, ze chcete menit barvu vyplne bodu
                 lwd = 2, las = 1)
... # doplneni popisku
```

95% elipsa spolehlivosti



Testovací statistiky

Testujeme $H_0 : \rho = \rho_0$ vs. $H_1 : \rho \neq \rho_0$, kde $\rho_0 = 0.251$.

Odhad $\hat{\rho}$ na základě dat (fce `cor()`): $\hat{\rho} = 0.3306$.

Waldova statistika:

$Z_W = \sqrt{n-3}(Z_R - \xi_0) \sim N(0,1)$, kde $Z_R = \frac{1}{2}\ln\frac{1+R}{1-R}$ a $\xi_0 = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}$.

Věrohodnostní statistika:

$U_{LR} = n\ln\frac{(1-\rho_0\hat{\rho})^2}{(1-\rho_0^2)(1-\hat{\rho}^2)} \sim \chi^2_1(\alpha)$.

Výsledky: $z_W = 1.1047989$ a $u_{LR} = 1.2417565$.

Intervaly spolehlivosti

Hranice Waldova DIS pro ρ jsou: $\tanh(z_R - \frac{u_{\alpha/2}}{\sqrt{n-3}}), \tanh(z_R + \frac{u_{\alpha/2}}{\sqrt{n-3}})$.

Věrohodnostní DIS: $\{\rho_0 : U_{LR} < \chi^2_1(\alpha)\}$.

Kritické hodnoty a p-hodnoty spočítáme postupy analogickými předchozím jednovýběrovým případům (např. testu o střední hodnotě). Nezapomeňte, že v R při výpočtu kritické hodnoty $\chi^2_1(\alpha)$ zadáváme do funkce `qchisq` argument `1 - alpha`.

```
## sekvence pro vypocet hranic verohodnostniho DIS
rho.i <- seq(0.1, 0.6, by = 0.0001)
```

Tabulka výsledků

	statistika	W_{hh}	W_{dh}	IS_{dh}	IS_{hh}	p-hodnota
Wald	1.1048	-1.96	1.9600	0.1869	0.4606	0.2692
Věrohodnost	1.2418	NA	3.8415	0.1880	0.4596	0.2651

Protože hodnota test. statistiky …… spadá/nespadá do kritického oboru ……., zamítáme/nezamítáme H_0 o tom, že hodnota korelačního koeficientu největší výšky mozkovny a morfologické výšky tváře u mužů je rovna 0.251.

Protože hodnota $\rho_0 = 0.251$ náleží/nenáleží DIS ……., zamítáme/nezamítáme H_0 o tom, že ….

Protože p-hodnota ……. je menší/větší než hladina významnosti α, zamítáme/nezamítáme H_0 o tom, že ….

Interpretace: Mezi korelačním koeficientem největší výšky mozkovny a morfologické výšky

tváře mužů starověké a novověké egyptské populace existuje/neexistuje statisticky významný rozdíl.

Mezi největší výškou mozkovny a morfologickou výškou tváře mužů starověké egyptské populace existuje mírný stupeň přímé lineární závislosti ($\rho = 0.3306$).

Graf věrohodnostního DIS