7.1, 7.2

Nechť náh. vektor $(X,Y)^T \sim N_2(\mu, \Sigma)$, kde $\mu = (\mu_1,\mu_2)^T$ je vektor středních hodnot a $\Sigma = \begin{pmatrix} \sigma_1^2 & \varrho\sigma_1\sigma_2 \\ \varrho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ je variační matice. K vyjádření vztahu mezi náh. veličinami X a Y použijeme klasický / lineární Pearsonův korelační koeficient

$$R = \frac{1}{n-1} \frac{\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{S_X^2 S_Y^2}} = \frac{1}{n-1} \frac{\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})}{S_X S_Y},$$

kde $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i-\bar{X})^2$ a $S_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i-\bar{Y})^2$.

Z bakalářského studia víme, že Pearsonův korelační koeficient pochází asymptoticky z normálního rozdělení, konkrétně

$$R \overset{A}{\sim} N\Big(\underbrace{\varrho}_{\text{střední hodnota}}, \underbrace{\frac{(1-\varrho^2)^2}{n-1}}_{\text{rozptyl}}\Big).$$

Konvergence R k normalitě je však pomalá, proto používáme tzv. Fisherovu Z-proměnnou

$$Z_R = \frac{1}{2}\ln\frac{1+R}{1-R} \sim N\Big(\underbrace{\frac{1}{2}\ln\frac{1+\varrho}{1-\varrho}}_{\text{střední hodnota}}, \underbrace{\frac{1}{n-3}}_{\text{rozptyl}}\Big)$$

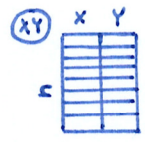
V příkladech 7.1 a 7.2 se zaměříme na rychlost konvergence statistik R a Z_R k normálnímu rozdělení při $n \to \infty$ (7.1) a při $\varrho \to 1$ (7.2). Příklady vyřešíme společně vytvořením funkce `rho.stat()`, která pro zadané $n, \varrho, \mu_1, \mu_2, \sigma_1, \sigma_2$ a M vykreslí dvojici histogramů (pro R + pro Z_R). Tuto funkci pak použijeme k vykreslení animací v obou příkladech.

```
rho.stat <- function(n, rho, mu1=..., mu2=..., sigma1=..., sigma2=..., M=...,
                     xlim1=c(0.5,1), xlim2=c(-0.5,2), ylim1=c(0,10), ylim2=c(0,4)) {
```
(xlim1: rozsah osy x prvního hist.; xlim2: rozsah osy x druhého hist.; ylim1: rozsah osy y prvního hist.; ylim2: rozsah osy y druhého hist.)

```
mu <- c(...)   ... vektor středních hodnot μ = (μ1, μ2)
Sigma <- matrix(...)   ... variační matice Σ = ((σ1², ϱσ1σ2), (ϱσ1σ2, σ2²)) ... ! po vytvoření matice Sigma si ji vypište a zkontrolujte ! často se v ní chybuje !
```

Generování dat: I. způsob - cyklem:

```
R <- zR <- NULL   ... příprava prázdných vektorů R a zR
for (i in 1:M) {
  XY <- MASS::mvrnorm(n=n, mu=μ, Sigma=Σ) ...
  X <- XY[,...]   ... vektor X (délka=n)
  Y <- XY[,...]   ... vektor Y (délka=n)
  R[i] <- cor(...)   ... R (1 číslo; délka R[i]=1)
  zR[i] <- zR   (1 číslo, délka zR[i]=1)
}   ... výstupem cyklu jsou dva vektory R a zR o délce M.
```



Generování dat: II. způsob - funkcí replicate():

```
R <- replicate(M, cor(mvrnorm(n=n, mu=μ, Sigma=Σ))[1,2])   ... vektor M=1000 korel. koef. R
zR <- 1/2 ln((1+R)/(1-R))   ... vektor M=1000 Fisherových Z-transformací.
```

(Výstupem je korelační matice!!! $\begin{pmatrix} 1 & R \\ R & 1 \end{pmatrix}$)

Generování dat (vyzkoušejte si oba způsoby) =)

Histogram R

```
xfit <- seq(...) ... posloupnost od -1 do 1 o délce minimálně 500
yfit <- dnorm(...) ... hustota N(ϱ, (1-ϱ²)²/(n-1)) nad posloupností xfit
par(mfrow=..., mar=...)
hist(..., prob=..., col=..., border=..., xlim=xlim1,
     ylim=ylim1, ...) ... histogram R se šrafovanými sloupci, rozsah osy x (xlim) je zadán argumentem xlim1,
                                                                   y (ylim)      -||-            ylim1
box(...)
lines(...) ... křivka hustoty N(ϱ, (1-ϱ²)²/(n-1)) se šířkou 2
mtext(...) ... popisek R
mtext(bquote(paste(...)), ...) .. automaticky se měnící popisek n=..., ϱ=...
```

Histogram Z_R

```
xfit <- seq(...) ... posloupnost od min(ZR)-1 do max(ZR)+1 o délce minimálně 500
yfit <- dnorm(...) ... hustota N(1/2 ln (1+ϱ)/(1-ϱ), 1/(n-3)) nad posl. xfit
hist(..., xlim=xlim2, ylim=ylim2, ...) ... histogram ZR
box(...)
lines(...) ... křivka hustoty N(1/2 ln (1+ϱ)/(1-ϱ), 1/(n-3))
mtext(...) ... popisek Z_R
mtext(...) ... automaticky se měnící popisek n=..., ϱ=...
}
```

S5.1: Animace:

```
n <- seq(...) ... posloupnost n zadaná podle zadání příkladu 5,10,15,...,70.
... úvodní nastavení animace
saveLatex(for(...){
          rho.stat(n=n[i], rho=0.8)
        },...)
```

Vytvořenou animaci si pomalu projděte a porovnejte kvalitu konvergence R a Z_R k normálnímu rozdělení. Do komentářů uveďte zhodnocení kvality konvergence a závěr, pro jak velké n je podle vašeho názoru již vhodné použít Pearsonův korelační koeficient R a Fisherovu Z-transformaci Z_R.

S5.2: Animace pro n=5:

```
rho <- seq(...) ... zadaná posloupnost ϱ = 0.1, 0.2, ..., 0.9
... úvodní nastavení animace
saveLatex(for(...){
          rho.stat(n=5, rho=rho[i], xlim1=c(0,1), xlim2=c(-2,3),
                   ylim1=c(0,5), ylim2=c(0,1))
        },...)
```

S5.2: Animace pro n=50:

Analogicky jako pro n=5. Hodnoty ostatních argumentů volte: xlim1=c(0,1), xlim2=c(-2,3), ylim1=c(0,12), ylim2=c(0,3).

Vytvořené animace si pomalu projděte a porovnejte kvalitu konvergence R a Z_R k normálnímu rozdělení. Do komentářů popište, jak se mění kvalita konvergence R a Z_R při $\varrho \to 0.9$ (mění-li se nějak) a je-li nějaký rozdíl mezi situací kdy n=5 a n=50.

V rámci tohoto příkladu si vyzkoušíme aplikaci testovacích statistik Z_W a U_{LR} na reálná data. Nejprve ověříme předpoklad dvourozměrné normality. Následně otestujeme H_0 ze zadání, a to všemi třemi způsoby (krit. oborem, IS a p-hodnotou) pro každou test. statistiku Z_W a U_{LR}. Nezapomeneme vždy uvést závěr o H_0 a nakonec interpretaci závěru testování. Nakonec vykreslíme hranici a oblast věrohodnostního 95% empirického IS.

Načtení a příprava dat

```
data <- read.delim(...)
data.M <- data[date$sex==..., c('skull.pH','face.H')]
data.M <- na.omit(...)

skull.M <- data.M$...
face.M <- data.M$...
```

Test(y) dvourozměrné normality

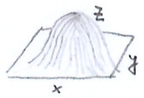
```
# H0: ...
# H1: ...
MVN::mvn(data.M, mvnTest = 'mardia')$multi ... MARDIŮV TEST (← preferovaný) .. testuje zvlášť aspekt šikmosti a špičatosti
MVN::mvn(  -||-  ,  -||-  = 'hz')$...     ... HENZE-ZIRKLERŮV TEST
MVN::mvn(  -||-  ,  -||-  = 'royston')$... ... ROYSTONŮV TEST
# Závěr: Protože..., H0... na hladině významnosti....
# Interpretace:
```

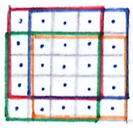
VYZKOUŠEJTE VŠECHNY 3 TESTY

3D (persp) graf

```
kd <- MASS::kde2d(skull.M, face.M, n=50, lims = c(120, 155, 90, 145)) ... jádrový odhad dvourozměrné hustoty (nad osou x v rozmezí 120-155 a nad osou y v rozmezí 90-145).
x <- kd$x ... vektor hodnot x vytvořený fcí kde2d
y <- kd$y ...  -||-   y   -||-   kde2d
Z <- kd$z ... matice dvourozměrné hustoty f(x,y) vytvořená fcí kde2d.
N <- dim(Z)[1]   (50)
stredy <- Z[-1,-1] + Z[-1,-n] + Z[..,..] + Z[...,...] ... analogická finta na získání středů jako v S1 1.
vyska <- cut(stredy, 12) ... rozdělení středů do 12 kategorií
par(mar=...)   (0 / 1 1 / 1)
persp(... , theta = -20, phi = 30, col = heat.colors(12)[vyska], ...) ... 3D jádrový odhad hustoty
```

(col: obarvení každé kategorie jednou z 12 barev škály heat.colors().)





Tečkový graf s 95% elipsou spolehlivosti

```
par(...)   (2 / 4 1 / 5)
car::dataEllipse(skull.M, face.M, level = 0.95,
                 xlim = c(122, 150), ylim = c(90, 141),
                 pch = 21, bg = 'khaki1', col = c('orange2', 'khaki4'), ...) ... Graf elipsy spolehlivosti
mtext(...) ... popisek 'nejvetsi vyska mozkovny (v mm)'
mtext(...) ... popisek 'Mardiův test ...' (nemusí být automatizovaný).
```

(level: 95% elipsa spolehlivosti; xlim: rozsah osy x; ylim: rozsah osy y; bg: barva výplně bodů; 'orange2': barva obrysů bodů; 'khaki4': barva elipsy)

Poznámka: Interpretace normality na základě tečkového grafu s elipsou spolehlivosti.

Předpokládáme, že pokud data pochází z dvourozměrného normálního rozdělení N_2, potom 95% elipsa spolehlivosti pokrývá alespoň 95% dat. (tj. nejvýše 5% dat leží mimo elipsu spolehlivosti).

Zjistěte tedy rozsah náhodného výběru n, spočítejte 5% z n, dále spočítejte, kolik bodů leží vně elipsy spolehlivosti a do komentářů uveďte zdůvodněný závěr o dvourozměrné normalitě dat na základě tohoto grafu.

Test o korelačním koeficientu ϱ:

H0:

H1:

Příprava hodnot

rho0 <- ... ϱ_0

n <- length(...) ... počet pozorování (rozsah náh. výběru)

alpha <- ... α

r <- cor(...) ... r ... realizace Pearsonova korelačního koeficientu R pro skull.M a face.M

zR <- $\frac{1}{2}\ln\frac{1+R}{1-R}$ ... z_R

ksi0 <- $\frac{1}{2}\ln\frac{1+\varrho_0}{1-\varrho_0}$ ... ξ_0

Funkce počítající hodnotu test. stat. U_{LR}

```
uLR.stat <- function(r, rho0, n){
    uLR <- n( ln (1-rho0*r)^2 / ((1-rho0^2)(1-r^2)) )
    return(...)
}
```

Testování kritickým oborem

Testovací statistiky

zW <- $\sqrt{n-3}\,(z_R-\xi_0)$ ... z_W

uLR <- uLR.stat(...) ... hodnota test. statistiky U_{LR}

Hranice kritických oborů

q1 <- ... $u_{\alpha/2}$

q2 <- ... $u_{1-\alpha/2}$

q <- ... $\chi^2_1(1-\alpha)$

Pro každý test uveďte rozhodnutí o H0

Testování IS:

Intervaly spolehlivosti ξ

dh.zR <- $z_R - \frac{u_{1-\alpha/2}}{\sqrt{n-3}}$ ... dolní hranice Waldova IS pro ξ_0

hh.zR <- $z_R - \frac{u_{\alpha/2}}{\sqrt{n-3}}$ ... horní hranice Waldova IS pro ξ_0

Intervaly spolehlivosti R

dh.R <- tanh(dh.zR) ... zpětně transformovaná dolní hranice Waldova IS (pro ϱ)

hh.R <- tanh(...) ... zpětně transformovaná horní hranice Waldova IS (pro ϱ)

čím více bodů, tím přesnější budou hranice věrohodnostního IS

Intervaly U_{LR}

rho.i <- seq(...) ... posloupnost ϱ_i od 0.1 do 0.6 se vzdálenostmi mezi body 0.0001 (5001 hodnot)

uLR.i <- uLR.stat(r, ...) ... vektor 5001 test. statistik $U_{LR,i}$ pro posloupnost ϱ_i (rho.i)

dh.uLR <- min(... [...]) ... dolní hranice věroh. 95% empirického IS

hh.uLR <- max(... [...]) ... horní hranice -||-

} princip je analogický jako v 6.6 a 5.8

Pro každý test uveďte rozhodnutí o H0:

Testování p-hodnotou

p-hodnoty

p.zW <- 2·min(pnorm(...), 1-pnorm(...))

p.uLR <- 1-pchisq(...)

Pro každý test uveďte rozhodnutí o H0.

Interpretace výsledků testování: Uveďte antropologický závěr (interpretaci výsledků testování). Co jsme se dozvěděli o vztahu nejvyšší výšky mozkovny a morfologické výšky tváře u mužů starověké egyptské populace?

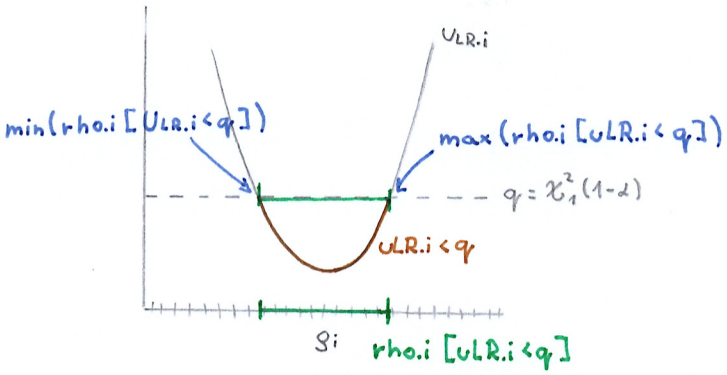
Na závěr uveďte také kompletní interpretaci Pearsonova korelačního koeficientu r = 0.3306 tak, jak jsme si ji uváděli na začátku semestru.

Tabulka výsledků

```
tab <- data.frame(...
                  IS.dh=c(dh.R, dh.uLR),
                  ...)
round(tab, 4) ... zaokrouhlení na 4 des. místa
```

Graf s věrohodnostním IS

```
par(...)  mfrow=c(1,1), mar=c(4,4,1,1)
plot(rho.i, uLR.i, ylim=c(0,5),...) ... křivka U_LR ∪
mtext(...) ... popisek ϱ
abline(...) ... vodorovná šedá přerušovaná referenční čára v hodnotě q = χ²₁(1-α) ---
lines(rho.i[...], uLR.i[...],...) ... barevně vypracěná oblast U_LR v IS ∪
mtext(...) ... popisek IS = (..., ...), analogický zápis jako v 5.8 a 6.6.
```



7.4

Odvození:

(Odvození je analogické jako ve cvičení 04. Doporučuji vyjít ze rde uvedených vzorců síly, vyzkoušel si odvození samostatně a rde uvedené odvození použít jako kontrolu. =))

1. $H_{01}: \vartheta = \vartheta_0$ $\quad H_{11}: \vartheta \neq \vartheta_0$ (oboustranná alt.)

Exaktní síla: $\beta^*_{11}(\xi) = \Phi(u_{\alpha/2} + \sqrt{n-3}\,(\xi_0 - \xi)) + \Phi(u_{\alpha/2} - \sqrt{n-3}\,(\xi_0 - \xi))$

Aproximativní síla: $\tilde{\beta}^*_{11}(\xi) = \Phi(u_{\alpha/2} + \sqrt{n-3}\,|\xi_0 - \xi|)$

Minimální rozsah n:

$\tilde{\beta}^*_{11}(\xi) = \Phi(u_{\alpha/2} + \sqrt{n-3}\,|\xi_0 - \xi|) \quad / u()$

$u_{\tilde{\beta}^*_{11}(\xi)} = u_{\alpha/2} + \sqrt{n-3}\,|\xi_0 - \xi|$

$$\left(\frac{u_{\tilde{\beta}^*_{11}(\xi)} - u_{\alpha/2}}{|\xi_0 - \xi|}\right)^2 + 3 = n \quad \Rightarrow \quad n \geq \left(\frac{u_{\tilde{\beta}^*_{11}(\xi)} - u_{\alpha/2}}{|\xi_0 - \xi|}\right)^2 + 3 \quad / u_{\alpha/2} = -u_{1-\alpha/2}$$

$$n \geq \left(\frac{u_{\tilde{\beta}^*_{11}(\xi)} + u_{1-\alpha/2}}{|\xi_0 - \xi|}\right)^2 + 3$$

2. $H_{02}: \vartheta \leq \vartheta_0$ $\quad H_{12}: \vartheta > \vartheta_0$ (pravostranná alt.)

Exaktní síla: $\beta^*_{12}(\xi) = 1 - \Phi(u_{1-\alpha} + \sqrt{n-3}\,(\xi_0 - \xi)) = \Phi(u_\alpha - \sqrt{n-3}\,(\xi_0 - \xi))$

Minimální rozsah n:

$\beta^*_{12}(\xi) = 1 - \Phi(u_{1-\alpha} + \sqrt{n-3}\,(\xi_0 - \xi))$

$1 - \beta^*_{12}(\xi) = \Phi(u_{1-\alpha} + \sqrt{n-3}\,(\xi_0 - \xi)) \quad / u()$

$u_{1-\beta^*_{12}(\xi)} = u_{1-\alpha} + \sqrt{n-3}\,(\xi_0 - \xi)$

$$\left(\frac{u_{1-\beta^*_{12}(\xi)} - u_{1-\alpha}}{\xi_0 - \xi}\right)^2 + 3 = n \quad \Rightarrow \quad n \geq \left(\frac{u_{1-\beta^*_{12}(\xi)} - u_{1-\alpha}}{\xi_0 - \xi}\right)^2 + 3 \quad / u_\alpha = -u_{1-\alpha}$$

$$n \geq \left(\frac{u_{1-\beta^*_{12}(\xi)} + u_\alpha}{\xi_0 - \xi}\right)^2 + 3$$

3. $H_{03}: \vartheta \geq \vartheta_0$ $\quad H_{13}: \vartheta < \vartheta_0$ (levostranná alt.)

Exaktní síla: $\beta^*_{13}(\xi) = \Phi(u_\alpha + \sqrt{n-3}\,(\xi_0 - \xi))$

Minimální rozsah n:

$\beta^*_{13}(\xi) = \Phi(u_\alpha + \sqrt{n-3}\,(\xi_0 - \xi)) \quad / u()$

$u_{\beta^*_{13}(\xi)} = u_\alpha + \sqrt{n-3}\,(\xi_0 - \xi)$

$$\left(\frac{u_{\beta^*_{13}(\xi)} - u_\alpha}{\xi_0 - \xi}\right)^2 + 3 = n \quad \Rightarrow \quad n \geq \left(\frac{u_{\beta^*_{13}(\xi)} - u_\alpha}{\xi_0 - \xi}\right)^2 + 3 \quad / u_\alpha = -u_{1-\alpha}$$

$$n \geq \left(\frac{u_{\beta^*_{13}(\xi)} + u_{1-\alpha}}{\xi_0 - \xi}\right)^2 + 3$$

1. Nejprve si naprogramujeme funkci min.n( ), která pro zadané hodnoty ϱ, ϱ_0, α, β^* a pro zvolený typ alternativy vrátí minimální rozsah náh. výběru n.

```
min.n <- function(rho, rho0, alpha=..., sila=0.8..., alternative=...){
    ksi <- ... ξ = 1/2 ln (1+ϱ)/(1-ϱ)
    ksi0 <- ... ξ0 = 1/2 ln (1+ϱ0)/(1-ϱ0)
    if(alternative == 'two.sided') {n <- ( (u_{β*_{1-sila/2}}(ξ) - u_{α/2}) / |ξ0 - ξ| )^2 + 3 )}
    if(            == 'greater'  ) {n <- ( (u_{1-β*_{1-sila}}(ξ) - u_{1-alpha}) / (ξ0 - ξ) )^2 + 3}
    if(            == 'less'     ) {n <- ( (u_{β*_{sila}}(ξ) - u_{alpha}) / (ξ0 - ξ) )^2 + 3
                                   <- ((qnorm(sila) - qnorm(alpha)) / (ksi0 - ksi))^2 + 3 }
    return(round(n))
}
```

2. Vykreslíme graf pro (a) oboustrannou alternativu:

```
rho <- c(seq(...), seq(...)) ... spojení dvou posloupností: První od -0.95 po -0.1 po kroku 0.05; druhá od 0.1 do 0.95 po kroku 0.05
n11 <- min.n(rho=rho, rho0=0, alternative='two.sided') ... vektor minimálních rozsahů n pro každou hodnotu ϱ z vektoru rho při volbě oboustranné alt.
par(...) 4 1/5 1
plot(rho, n11, xlim=c(-1,1), ...) ... čárkový graf: body s jinou barvou výplně a obrysu
mtext(expression(rho - rho[0]), ...) ... popisek ϱ-ϱ0
mtext(...) ... popisek 'oboustranná alternativa'
```

3. Vykreslíme graf ((b) pro pravostrannou alt. resp. (c) pro levostrannou alt.)

```
rho <- seq(...) ... pos. od 0.1 do 0.95 po 0.05 resp. od -0.95 do -0.1 po 0.05.
n13
n12 <- min.n(...) ... vektor minimálních rozsahů n pro každou hodnotu ϱ z vektoru rho při volbě pravostr. (resp. levostranné) alt.
plot(rho, n12, xlim=c(-1,1), ...) ... čárkový graf: body s jinou barvou výplně a obrysu
segments(-2, 0, 0, 0, ...) ... šedá přerušovaná čára značící situaci, pro které nemělo smysl minimální rozsah n počítat
mtext(expression(...)) ... popisek ϱ-ϱ0
mtext(...) ... popisek s typem alternativy
```

4. Spočítáme tabulku:

```
n11 <- min.n(rho=0.60, rho0=0.85, alpha=..., sila=..., alternative=...)
n12 <- min.n(...)
n13 <- min.n(...)
alpha <- c( )   0.10  0.05  0.01   (délka = 3)
sila <- c( ) ... vektor zadaných sil (délka = 3)
beta <- 1-sila ... vektor pravd. chyb II. druhu (délka = 3)
rho0 <- c( ) ... vektor ϱ0 (3)
rho  <- c( ) ... vektor ϱ (3)
n    <- c(n11, n12, n13) ... vektor minimálních rozsahů n

tab <- data.frame(alpha, sila, ..., n, row.names = c('pravostranna', ...)) ... tabulka výsledků

round(...) ... výpis tabulky zaokrouhlení na 4 des. místa
```