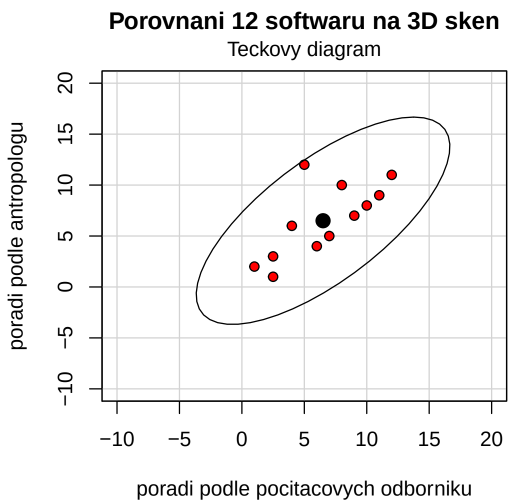
12 Jednoduchá korelační analýza

Příklad 12.1. Testování nezávislosti ordinálních veličin 12 různých softwarových firem nabízí speciální programové vybavení pro 3D skenování lidského těla. Jednotlivé programy byly posouzeny odbornou komisí složenou z počítačových odborníků a komisí složenou z antropologů. Úkolem bylo doporučit vhodný program na základě stanovení pořadí jednotlivých programů. Výsledky posouzení:

Produkt firmy číslo	1	2	3	4	5	6	7	8	9	10	11	12
Pořadí dle programátorů	6	7	1	8	4	2.5	9	12	10	2.5	5	11
Pořadí dle antropologů	4	5	2	10	6	1	7	11	8	3	12	9

Vypočtěte Spearmanův koeficient pořadové korelace a na hladině významnosti $\alpha = 0.05$ testujte hypotézu, že hodnocení obou komisí jsou nezávislá. Data jsou uložena v souboru 3D-sken.txt.

Ověření dvourozměrné normality pomocí tečkového diagramu



Testování hypotézy o nezávislosti

- H_0 : ....
- H_1 : ....

```
## [1] "Spearmanuv koeficient: 0.7145"
## [1] "Asymptoticka varianta testu: T0= 3.2298"
## [1] -2.228139
## [1] 2.228139
## [1] "Asymptoticka varianta testu: p-hodnota= 0.009024"
```

Spearmanův koeficient pořadové korelace nabývá hodnoty r_S =................................, tedy mezi hodnocením obou komisí existuje .... stupeň .... závislosti.

1. **Testování kritickým oborem**
   Tento postup používáme přednostně, protože $n = 12 < 20$. Testovací statistikou je v tomto případě přímo hodnota Spearmanova koeficientu pořadové korelace r_S =................................. .
   Kritický obor má tvar W =.. Protože r_S............W, H_0 o pořadové / lineární nezávislosti .... na hladině významnosti α = ...

2. **Testování kritickým oborem - Asymptotické varianta testu**
   Tento postup používáme v případě, že $n > 20$. To v našem případě není splněno, řešení si tedy uvádíme jen pro příklad.
   Testovací statistika T_0 =................................. Kritický obor má tvar W =.. .
   Protože T_0 .... W, H_0 o pořadové / lineární nezávislosti .... na hladině významnosti $\alpha =$ ....

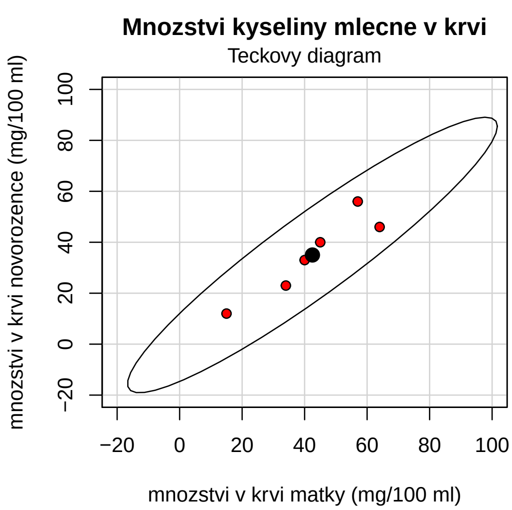
3. **Testování p-hodnotou - Asymptotická varianta testu**
   Tento postup používáme v případě, že $n > 20$. To v našem případě není splněno, řešení si tedy uvádíme jen pro příklad.
   Protože p-hodnota .... je .... než $\alpha = 0.05$, H_0 o pořadové / lineární nezávislosti .... na **asymptotické** hladině významnosti α =.............................. .

Příklad 12.2. Testování nezávislosti intervalových veličin
Zjišťovalo se, kolik mg kyseliny mléčné je ve 100 ml krve matek prvorodiček (veličina X) a u jejich novorozenců (veličina Y) těsně po porodu. Byly získány tyto výsledky:

Číslo matky	1	2	3	4	5	6
x_i	40	64	34	15	57	45
y_i	33	46	23	12	56	40

Pomocí tečkového diagramu otestujte dvourozměrnou normalitu dat. Vypočtěte výběrový korelační koeficient, sestrojte 95 % interval spolehlivosti pro korelační koeficient a na hladině významnosti $\alpha = 0.05$ testujte hypotézu o nezávislosti výsledků obou měření. Data jsou uložena v souboru kyselina_mlecna.txt.

Ověření dvourozměrné normality pomocí tečkového diagramu



Testování hypotézy o nezávislosti

- H_0 : ....
- H_1 : ....

```
##       cor
## 0.9348324
## [1] "T0=5.2653"
## [1] -2.776445
## [1] 2.776445
## [1] "IS= -0.8114 ; 0.8114"
## [1] "p-hodnota= 0.006232"
## [1] "Asymptotický IS= 0.5108 ; 0.993"
```

Výběrový korelační koeficient korelace nabývá hodnoty r_{12} = ..., tedy mezi množstvím kyseliny mléčné ve 100 ml krve rodiček a jejich novorozenců existuje .... stupeň .... závislosti.

1. **Testování kritickým oborem**
   Testovací statistika T_0 nabývá hodnoty ...., kritický obor má potom tvar ....
   Protože T_0 .... W, H_0 o nezávislosti .... na hladině významnosti $\alpha =$ ....
2. **Testování IS**
   Interval spolehlivosti pro ρ má tvar ....
   Protože ...., H_0 o nezávislosti .... na hladině významnosti $\alpha =$ ....
3. **Testování p-hodnotou**
   Protože p-hodnota .... je .... než $\alpha = 0.05$, H_0 o nezávislosti .... na hladině významnosti $\alpha =$............................ .

S rizikem omylu nejvýše 5 % jsme tedy prokázali, že mezi oběma koncentracemi .... pořadová / lineární závislost.

Příklad 12.3. Porovnání dvou korelačních koeficientů V psychologickém výzkumu bylo vyšetřeno 426 hochů a 430 dívek. Ve skupině hochů činil výběrový koeficient korelace mezi verbální a performační složkou IQ 0.6033, ve skupině dívek činil 0.5833. Za předpokladu dvourozměrné normality dat testujte na hladině významnosti $\alpha = 0.05$ hypotézu, že korelační koeficienty se neliší.

Testování hypotézy o dvou korelačních koeficientech

- H_0 : ....
- H_1 : ....

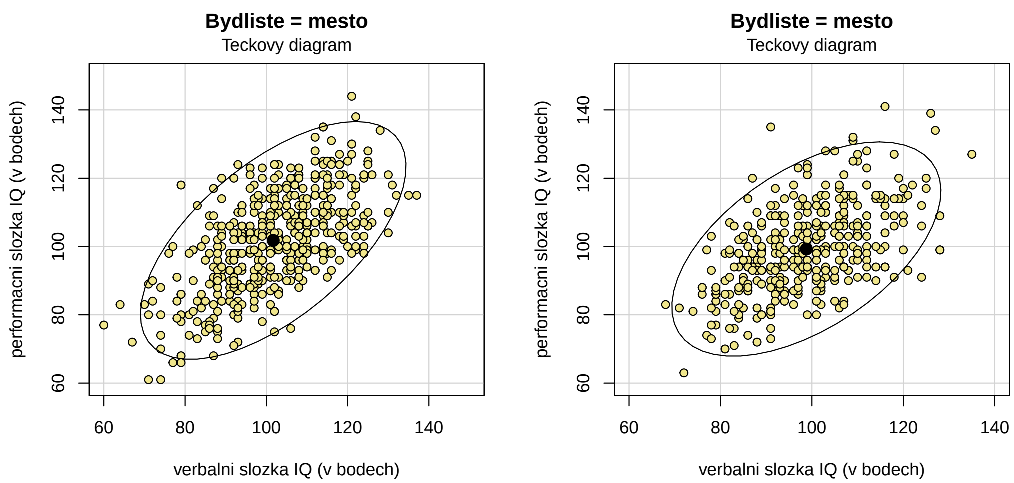
```
R1 <- 0.6033
R2 <- 0.5833
n1 <- 426
n2 <- 430
ksi <- 0
Z1 <- 1/2 * log((1 + R1)/(1 - R1))
Z2 <- 1/2 * log((1 + R2)/(1 - R2))
Zw <- (Z1 - Z2 - ksi) / sqrt(1/(n1 - 3) + 1/(n2 - 3))

(p.val <- 2 * min(pnorm(Zw), 1 - pnorm(Zw)))

## [1] 0.6527169
```

Testování pomocí p-hodnoty
Protože p-hodnota .... je .... než $\alpha = 0.05$, tedy H_0 o shodě dvou koeficientů korelace .... na **asymptotické** hladině významnosti $\alpha =$............................ .

Příklady k samostatnému řešení

Příklad 12.4. Načtěte datový soubor IQ.txt. Za předpokladu dvourozměrné normality dat (orientačně ověřte pomocí dvourozměrného tečkového diagramu) testujte na hladině významnosti $\alpha = 0.1$ hypotézu, že korelační koeficienty mezi verbální a performační složkou IQ jsou stejné u dětí z města a venkova.

Grafické ověření normality




```
## [1] 0.0780111
```

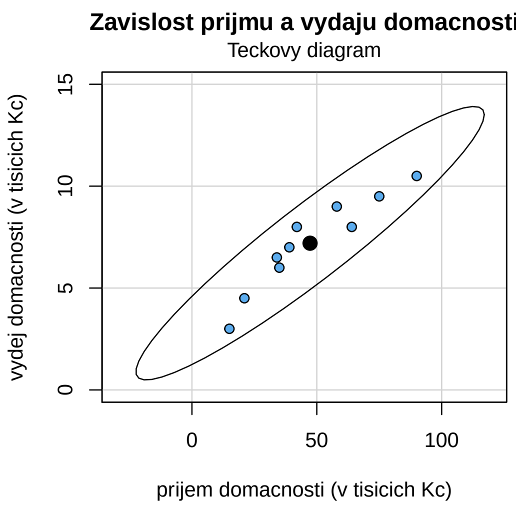
Výsledek p-hodnota$= 0.07801$, tedy s rizikem omylu nejvýše 10% jsme prokázali, že korelační koeficienty se liší.

Příklad 12.5. V náhodném výběru 10 dvoučlenných domácností byl zjišťován měsíční příjem (veličina X, v tisících Kč) a vydání za potraviny (veličina Y, v tisících Kč).

x_i	15	21	34	35	39	42	58	64	75	90
y_i	3	4.5	6.5	6	7	8	9	8	9.5	10.5

Vypočtěte a interpretujte výběrový koeficient korelace. Na hladině významnosti $\alpha = 0.05$ testujte hypotézu o nezávislosti veličin X, Y. Sestrojte 95 % asymptotický interval spolehlivosti pro ρ. Data jsou uložena v souboru prijem_vydani.txt.

Grafické ověření normality



Výsledek

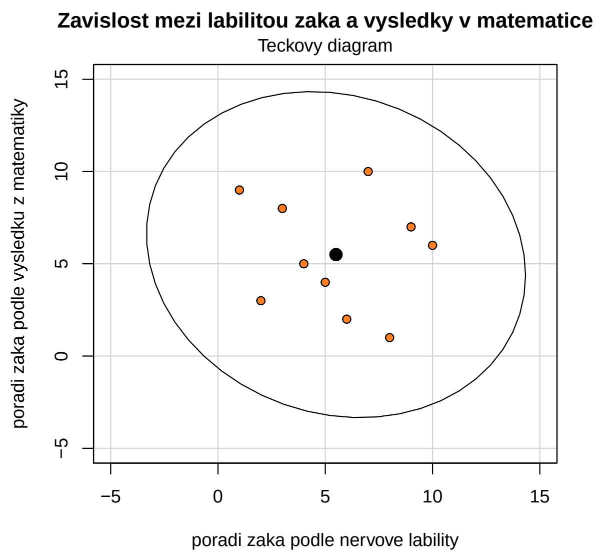
$r_{12} = 0.9405$, mezi měsíčními příjmy a výdaji tedy existuje velmi vysoký stupeň přímé lineární závislosti. p-hodnota$= 5.095\,\mathrm{e}-05$, tedy H_0 zamítáme na hladině významnosti $\alpha = 0.05$. S pravděpodobností alespoň 0.95 platí: $0.7623 < \rho < 0.9862$.

Příklad 12.6. Bylo sledováno 10 žáků. Na základě psychologického vyšetření byli tito žáci seřazeni podle nervové lability (čím byl žák labilnější, tím dostal vyšší pořadí R_i). Kromě toho sledování žáci dostali pořadí Q_i na základě svých výsledků v matematice (nejlepší žák v matematice dostal pořadí 1). Výsledky jsou uvedeny v tabulce:

Pořadí Ri	1	2	3	4	5	6	7	8	9	10
Pořadí Qi	9	3	8	5	4	2	10	1	7	6

Vypočtěte vhodný korelační koeficient a jeho hodnotu řádně interpretujte. Na hladině významnosti $\alpha = 0.05$ testujte hypotézu, že nervová labilita a výsledky v matematice jsou nezávislé. Data jsou uložena v souboru nervova_labilita.txt.

Grafické ověření normality



Výsledek: Spearmanův koeficient pořadové korelace $r_S = -0.127$, tedy mezi nervovou labilitou žáka a jeho výsledky v matematice existuje nízký stupeň nepřímé pořadové závislosti. p-hodnota= 0.7329, a tedy H_0 nezamítáme na hladině významnosti $\alpha = 0.05$.