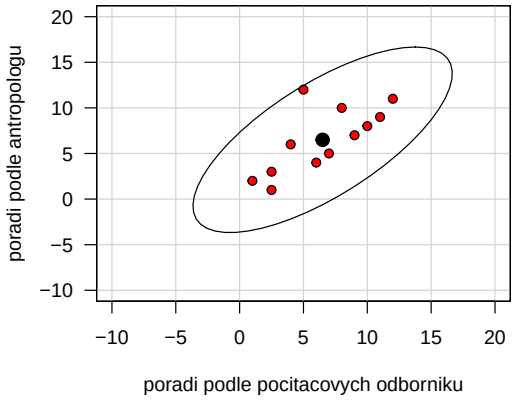
11 Korelační analýza

Příklad 11.1. Testování nezávislosti ordinálních veličin
12 různých softwarových firem nabízí speciální programové vybavení pro 3D skenování lidského těla. V rámci recenze byly jednotlivé programy posuzovány jednak odbornou komisí složenou z počítačových odborníků a jednak komisí složenou z antropologů. Výsledky posouzení jsou uvedeny v následující tabulce.

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.

Řešení příkladu 11.1
Ověření dvourozměrné normality pomocí tečkového diagramu



Testování hypotézy o nezávislosti

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

```
[1] "Spearmanuv koeficient: rucni vypocet:  0.715035"
```

```
[1] "Spearmanuv koeficient: cor.test():  0.714537"
```

```
[1] "Asymptoticka varianta testu: T0 =  3.2344"
```

```
[1] -2.228139
```

```
[1] 2.228139
```

```
[1] "Asymptoticka varianta testu: p-hodnota =  0.008954"
```

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

a) Testování pomocí kritického oboru
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 $\alpha =$ ..... .

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

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

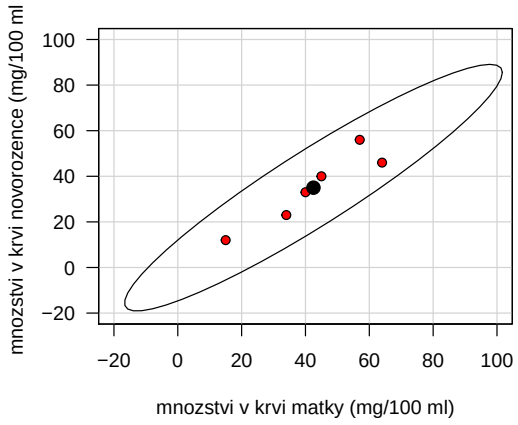
Interpretace výsledků testování: S rizikem omylu nejvýše 5 % jsme prokázali, že mezi hodnoceními obou komisí existuje / neexistuje statisticky významná pořadová / lineární závislost.

Příklad 11.2. 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.

Řešení příkladu 11.2
Ověření dvourozměrné normality pomocí tečkového diagramu



Testování hypotézy o nezávislosti

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

```
      cor
0.9348324
```

```
[1] "T0 = 5.2653"
```
9

```
[1] -2.776445
```
10

```
[1] 2.776445
```
11

```
[1] "IS =  -0.8114 ; 0.8114"
```
12

```
[1] "p-hodnota =  0.006232"
```
13

```
[1] "Asymptoticky IS =  0.5108 ; 0.993"
```
14

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

a) Testování pomocí kritického oboru
Testovací statistika T_0 nabývá hodnoty ....., kritický obor má potom tvar ..... .
Protože T_0 ..... W, H_0 o nezávislosti ..... na hladině významnosti $\alpha =$ ..... .

b) Testování pomocí IS
Interval spolehlivosti pro ρ má tvar .....
Protože ....., H_0 o nezávislosti ..... na hladině významnosti $\alpha =$ ..... .

c) Testování pomocí p-hodnoty
Protože p-hodnota ..... je ..... než $\alpha = 0.05$, H_0 o nezávislosti ..... na hladině významnosti $\alpha =$..... .

d) *Testování pomocí asymptotického IS
Asymptotický interval spolehlivosti pro ρ má tvar ..... .
Protože ....., H_0 o nezávislosti ..... na hladině významnosti $\alpha =$ ..... .

Interpretace výsledků testování: S rizikem omylu nejvýše 5 % jsme prokázali, že mezi oběma koncentracemi existuje / neexistuje statisticky významná pořadová / lineární závislost.

Příklad 11.3. 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.

Řešení příkladu 11.3

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

a) Testování pomocí kritického oboru

```
[1] 0.449991
```
15

```
[1] -1.959964
```
16

Testovací statistika Z_W nabývá hodnoty ....., kritický obor má potom tvar ..... .
Protože Z_W ..... W, H_0 o nezávislosti ..... na hladině významnosti $\alpha =$ ..... .

b) Testování pomocí p-hodnoty

```
[1] 0.6527169
```
17

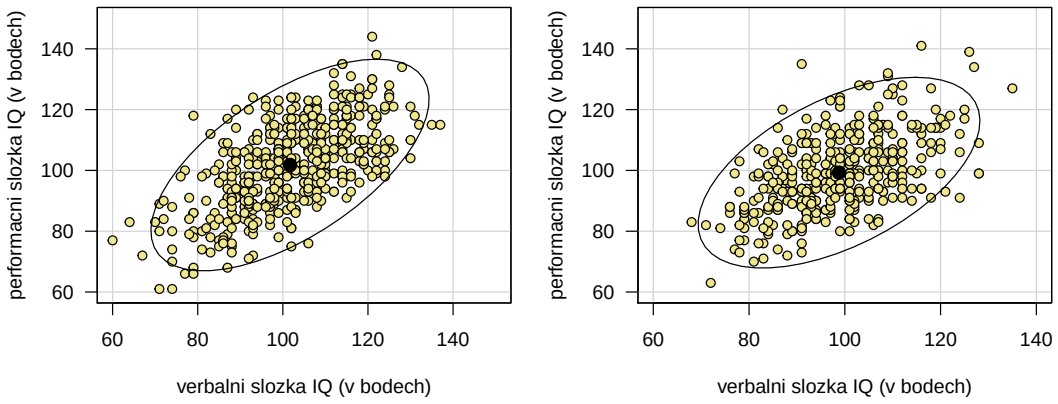
Protože p-hodnota ..... je ..... než $\alpha = 0.05$, H_0 o shodě dvou koeficientů korelace ..... na **asymptotické** hladině významnosti $\alpha =$..... .

Interpretace výsledků testování: Oba korelační koeficienty se statisticky významně liší / neliší.

Příklady k samostatnému řešení

Příklad 11.4. 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.

Řešení příkladu 11.4




```
[1] 0.0780111
```

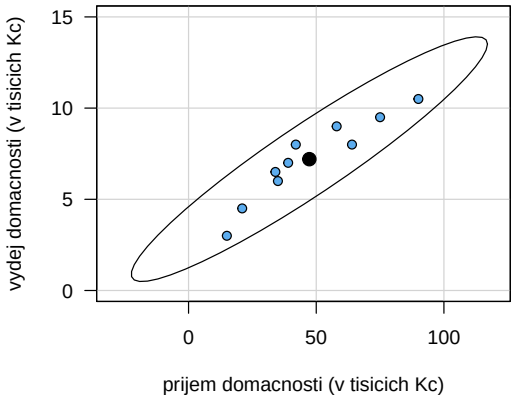
Výsledek: p-hodnota= 0.07801, tedy s rizikem omylu nejvýše 10% jsme prokázali, že korelační koeficienty se liší.

Příklad 11.5. 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.

Řešení příkladu 11.5



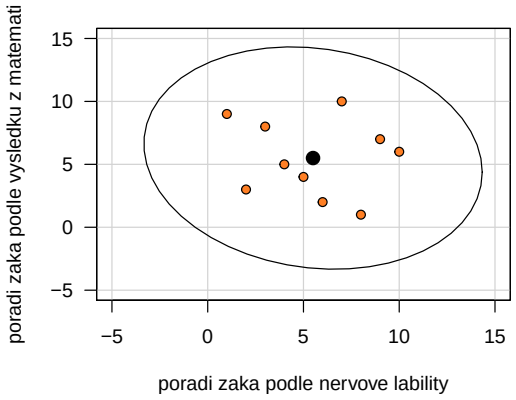
Výsledek: $r_{12} = 0.9405$, mezi měsíčními příjmy a výdaji tedy existuje velmi vysoký stupeň přímé lineární závislosti. p-hodnota= $5.095\,\mathrm{e} - 05$, tedy H_0 zamítáme na hladině významnosti $\alpha = 0.05$. S pravděpodobností alespoň 0.95 platí: $0.7623 < \rho < 0.9862$.

Příklad 11.6. 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 sledovaní 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.

Řešení příkladu 11.6



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 $\alpha = 0.05$.