Cvičení 4.: Korelace, Bayesův vzorec, opakované nezávislé pokusy

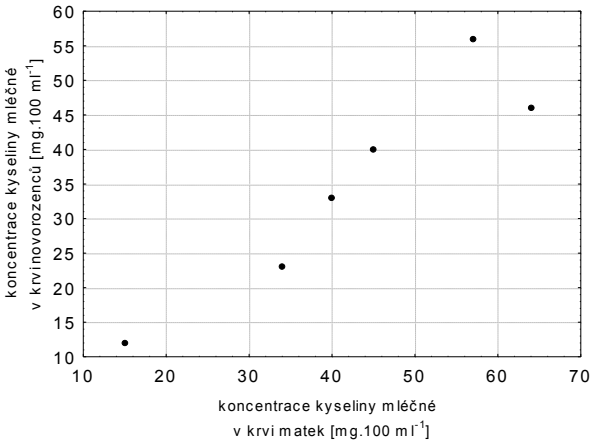
Úkol 1.: Zjišťovalo se, kolik mg kyseliny mléčné je ve 100 ml krve matek prvorodiček (veličina X) a u jejich novorozenců (veličina Y) těsně po porodu. Byly získány tyto výsledky:

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

Nakreslete dvourozměrný tečkový diagram a vypočtěte Pearsonův koeficient korelace znaků X, Y.

Řešení:

Dvourozměrný tečkový diagram



Výpočet korelace:

Statistiky – Základní statistiky/tabulky – Korelační matice – OK – 1 seznam proměnných – X, Y – OK, na záložce Možnosti zrušíme volbu Včetně průměrů a sm. odch. – Výpočet.¨

Proměnná	X	Y
X	1,00	0,93
Y	0,93	1,00

Vidíme, že mezi X a Y existuje silná přímá lineární závislost.

Úkol 2.: Načtěte soubor korkoef.sta, který obsahuje proměnné x,y1,y2,y3,y4, x4. Vypočtěte Pearsonovy korelační koeficienty dvojic proměnných (x,y1), (x,y2), (x,y3), (x4,y4) a pro každou z uvedených dvojic proměnných nakreslete dvourozměrný tečkový diagram. Pro které dvojice proměnných se hodí Pearsonův korelační koeficient jako vhodná míra těsnosti lineární závislosti?

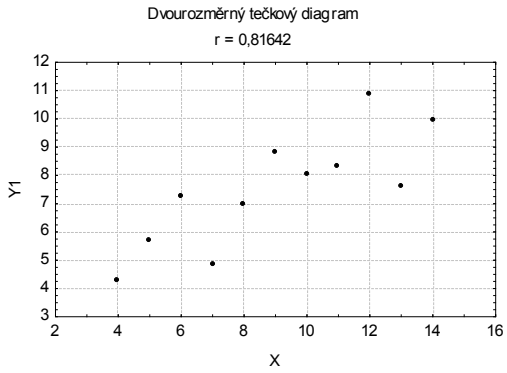
Řešení:

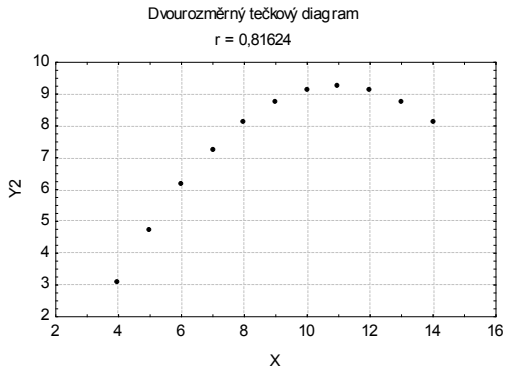
Variable	Correlations (korkoef)	
	X	Y1
X	1,000000	0,816421
Y1	0,816421	1,000000

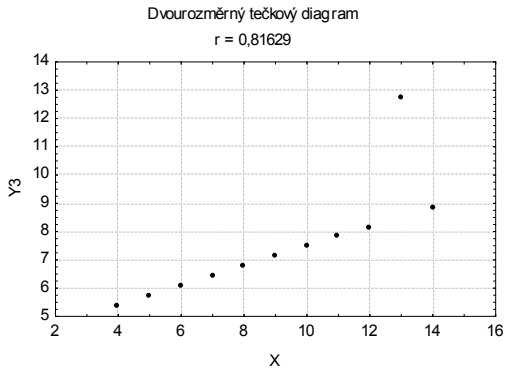
Variable	Correlations (korkoef)	
	X	Y2
X	1,000000	0,816237
Y2	0,816237	1,000000

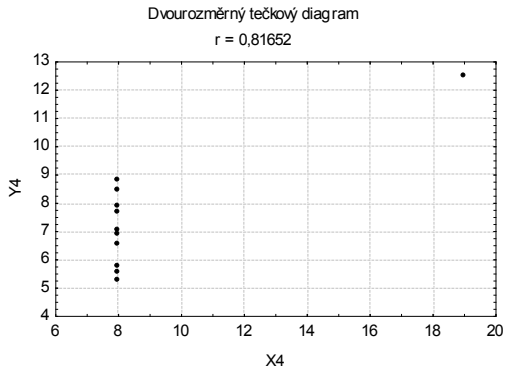
Variable	Correlations (korkoef)	
	X	Y3
X	1,000000	0,816287
Y3	0,816287	1,000000

Variable	Correlations (korkoef)	
	X4	Y4
X4	1,000000	0,816521
Y4	0,816521	1,000000









Komentář: Ve všech čtyřech případech nabývá koeficient korelace hodnoty 0,816, což by svědčilo o vysokém stupni těsnosti lineárního vztahu mezi sledovanými dvojicemi veličin. Při pohledu na dvourozměrné tečkové diagramy je však zřejmé, že pouze v prvním případě je použití Pearsonona korelačního koeficientu oprávněné.

Bayesův vzorec
Jevy $H_1, \ldots, H_n$ tvoří úplný systém hypotéz, tj. navzájem se vylučují a přitom vyčerpávají všechny možnosti. Jev A s hypotézami nesouvisí.
Počítáme pravděpodobnost některé hypotézy za podmínky, že nastal jev A:

$$P(H_k / A) = \frac{P(H_k) P(A/H_k)}{P(A)}, \; k = 1, \ldots, n, \text{ kde } P(A) = \sum_{i=} (H_i) P(A/H_i).$$

Vzorový příklad: U jistého druhu elektrického spotřebiče se s pravděpodobností 0,01 vyskytuje výrobní vada. U spotřebiče s touto výrobní vadou dochází v záruční lhůtě k poruše s pravděpodobností 0,5. Výrobky, které tuto vadu nemají, se v záruční lhůtě porouchají s pravděpodobností 0,01. Jaká je pravděpodobnost, že výrobek, který se v záruční lhůtě porouchá, bude mít dotyčnou výrobní vadu?

Řešení:
H_1 - výrobek má dotyčnou výrobní vadu
H_2 - výrobek nemá tuto výrobní vadu
A - výrobek se v záruční době porouchá
Pak je: $P(H_1) = 0{,}01$, $P(H_2) = 0{,}99$, $P(A/H_1) = 0{,}5$, $P(A/H_2) = 0{,}01$
$P(A) = P(H_1).P(A/H_1) + P(H_2).P(A/H_2) = 0{,}01.0{,}5 + 0{,}99.0{,}01 = 0{,}0149$

$$P(H_1 / A) = \frac{P(H_1) P(A / H_1)}{P(A)} = \frac{0{,}01 \cdot 0{,}5}{0{,}0149} = 0{,}3386$$

Příklady k samostatnému řešení:
1. Ve společnosti je 45% mužů a 55% žen. Výšku nad 190 cm má 5% mužů a 1% žen. Náhodně vybraná osoba je vyšší než 190 cm. Jaká je pravděpodobnost, že je to žena?
Návod: A … osoba měří více než 190 cm, H_1 … osoba je žena, H_2 … osoba je muž.
Výsledek: 0,1964

2. Potřebu smrkových sazenic kryje lesní závod produkcí dvou školek. První školka kryje 75% výsadby, přičemž ze 100 sazenic je 80 první jakosti. Druhá školka kryje výsadbu z 25%, přičemž na 100 sazenic připadá 60 první jakosti. Jaká je pravděpodobnost, že náhodně vybraná sazenice první jakosti pochází z produkce první školky?
Návod: A … sazenice je 1. jakosti, H_1 … sazenice pochází z 1. školky, H_2 … sazenice pochází z 2. školky.
Výsledek: 0,8

Opakované nezávislé pokusy - binomické rozložení pravděpodobností

Opakované nezávisle provádíme týž náhodný pokus a sledujeme nastoupení jevu, kterému říkáme úspěch. V každém z těchto pokusů nastává úspěch s pravděpodobností ϑ, $0 < \vartheta < 1$.

Pravděpodobnost, že v prvních n pokusech úspěch nastane právě x-krát ($0 \le x \le n$):

$$P_n(x) = \binom{n}{x} \vartheta^x (1-\vartheta)^{n-x} .$$

K výpočtu v systému STATISTICA slouží funkce Binom(x; ϑ ; n)

Pravděpodobnost, že v prvních n pokusech úspěch nastane nejvýše x_1-krát ($0 \le x_1 \le n$):

$$\sum_{x=0}^{x_1} P_n(x) .$$

K výpočtu v systému STATISTICA slouží funkce IBinom(x_1; ϑ ; n)

Pravděpodobnost, že v prvních n pokusech úspěch nastane aspoň x_0-krát ($0 \le x_0 \le n$):

$$\sum_{x=x_0}^{n} P_n(x) .$$

Výpočet lze provést takto: 1 - IBinom(x_0 - 1; ϑ ; n)

Pravděpodobnost, že v prvních n pokusech úspěch nastane aspoň x_0-krát a nejvýše x_1-krát:

$$\sum_{x=x_0}^{x_1} P_n(x) .$$

Výpočet lze provést takto: IBinom(x_1; ϑ ; n) - IBinom(x_0 - 1; ϑ ; n)

Příklad na binomické rozložení pravděpodobností: Pojišťovna zjistila, že 12% pojistných událostí je způsobeno vloupáním. Jaká je pravděpodobnost, že mezi 30 náhodně vybranými pojistnými událostmi bude způsobeno vloupáním
nejvýše 6,
aspoň 6,
právě 6,
od dvou do pěti?

Řešení:
Počet pokusů: n = 30, pravděpodobnost úspěchu: $\vartheta = 0{,}12$
ad a)

$$\sum_{x=0}^{x_1} P_n(x) = \sum_{x=0}^{6} P_{30}(x) = \sum_{x=0}^{6} \binom{30}{x} 0{,}12^x 0{,}88^{30-x} = \text{IBinom}(6;0{,}12;30) = 0{,}9393$$

S pravděpodobností 93,93% bude mezi 30 náhodně vybranými pojistnými událostmi způsobeno vloupáním nejvýše 6 událostí.

ad b)

$$\sum_{x=0}^{\quad} (x) = \sum_{x=}^{} P_{30}(x) = 1 - \sum_{x=}^{} P_{30}(x) = 1 - \sum_{x=}^{} \binom{30}{x} 0{,}12^x 0{,}88^{30-} = 1 - \text{IBinom}(5;0{,}12;30) = 0{,}1431$$

S pravděpodobností 14,31% bude mezi 30 náhodně vybranými pojistnými událostmi způsobeno vloupáním aspoň 6 událostí.

ad c)

$$P_n(x) = P_{30}(5) = \binom{30}{5} 0{,}12^6 0{,}88^{24} = \text{Binom}(6;0{,}12;30) = 0{,}0825$$

S pravděpodobností 8,25% bude mezi 30 náhodně vybranými pojistnými událostmi způsobeno vloupáním právě 6 událostí.

ad d)

$$\sum_{x=0}^{\quad} (x) = \sum_{x=}^{} P_{30}(x) = \sum_{x=}^{} P_{30}(x) - \sum_{x=}^{} P_{30}(x) = \sum_{x=}^{} \binom{30}{x} 0{,}12^x 0{,}88^{30-} - \sum_{x=}^{} \binom{30}{x} 0{,}12^x 0{,}88^{30-} =$$

$$= \text{IBinom}(5;0{,}12;30) - \text{IBinom}(1;0{,}12;30) = 0{,}7469$$

S pravděpodobností 74,69% bude mezi 30 náhodně vybranými pojistnými událostmi způsobeno vloupáním od 2 do 5 událostí.

Návod: Otevřeme nový datový soubor se čtyřmi proměnnými a o jednom případu.
Do Dlouhého jména 1. proměnné napíšeme =IBinom(6;0,12;30).
Do Dlouhého jména 2. proměnné napíšeme =1-IBinom(5;0,12;30).
Do Dlouhého jména 3. proměnné napíšeme =Binom(6;0,12;30).
Do Dlouhého jména 3. proměnné napíšeme =IBinom(5;0,12;30)-IBinom(1;0,12;30).

Příklad k samostatnému řešení: V rodině je 10 dětí. Za předpokladu, že chlapci i dívky se rodí s pravděpodobností 0,5 a pohlaví se formuje nezávisle na sobě, určete pravděpodobnost, že v této rodině je
a) právě 5 chlapců
b) nejméně 3 a nejvýše 8 chlapců.
Výsledek: ad a) 0,246, ad b) 0,935