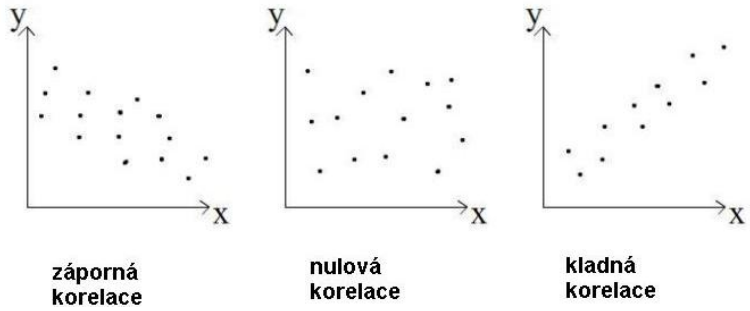
I Pearsonův korelační koeficient

I.I Úvod

Předpokládejme, že náhodně vybereme n objektů (nebo osob) ze zkoumané populace. Často se stává, že na každém z objektů měříme ne pouze jednu, ale několik kvantitativních proměnných. Uvažujme tedy pár takovýchto proměnných, mohlo by být zajímavé zjistit, zda mezi nimi existuje lineární vztah; tedy zjistit, jestli jsou korelované.
Typy korelace bychom mohli kategorizovat podle toho, co se stane s první proměnou, když druhá poroste:

- Kladná korelace – první proměnná má tendenci také růstů;
- Záporná korelace – první proměnná má tendenci klesat;
- Nulová korelace – první proměnná nemá tendenci ani růst, ani klesat.

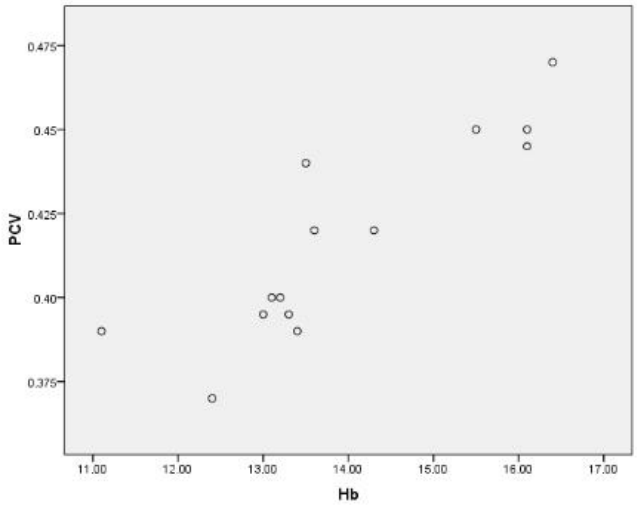
Začátkem každé takové analýzy by tedy měla být konstrukce a následné prozkoumání bodového grafu. Následují příklady záporné, nulové a kladné korelace.



Motivační příklad. Podívejme se nyní na konkrétní příklad. Následující data obsahují úrovně hemoglobinu (Hb) a celkové objemy buněk (PCV) od 14 dárců krve ženského pohlaví. Zajímalo by nás, zda existuje nějaký vztah mezi proměnnými Hb a PCV v populaci žen.

Hb	PCV	Hb	PCV
15.5	0.450	13.1	0.400
13.6	0.420	16.1	0.445
13.5	0.440	16.4	0.470
13.0	0.395	13.4	0.390
13.3	0.395	13.2	0.400
12.4	0.370	14.3	0.420
11.1	0.390	16.1	0.450

Bodový graf naznačuje vztah mezi PVC a Hb, kdy se vyšší hodnoty HB spojují s vyššími hodnotami PCV. Zdá se, že mezi těmito proměnnými existuje pozitivní korelace. Také si všimněme, že vztah mezi těmito proměnnými se zdá být lineární.



I.II Korelační koeficient

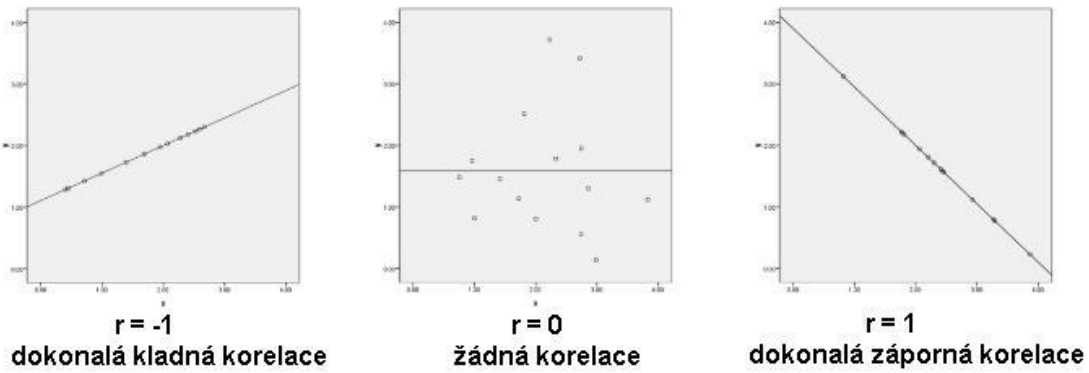
Pearsonův korelační koeficient je statistický ukazatel síly lineárního vztahu mezi párovými daty. Jedná se o výběrový korelační koeficient. Označme ho r, pro jeho hodnoty platí:

$$-1 \leq r \leq 1$$

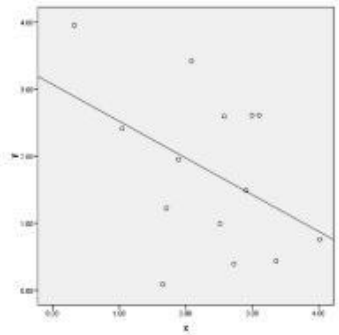
Poznámky:

- Kladné hodnoty r znamenají kladnou lineární korelaci;
- Záporné hodnoty r znamenají negativní lineární korelaci;
- Hodnota r nula znamená, že mezi proměnnými neexistuje lineární korelace;
- Čím je hodnota blíže 1 nebo -1, tím silnější lineární korelace je.

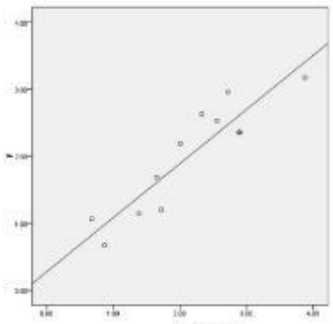
Na schématu jsou ukázky vzorku dat a hodnoty jejich příslušných korelačních koeficientů. První tři reprezentují „extrémní“ hodnoty korelací, a to -1 , 0 a 1:



Jestliže $r = \pm 1$, potom řekneme, že máme dokonalou korelaci, kdy jsou body poskládané v dokonale rovné přímce. Nicméně výběrový soubor, se kterým většinou pracujeme vypadá spíše, jako následující soubory:



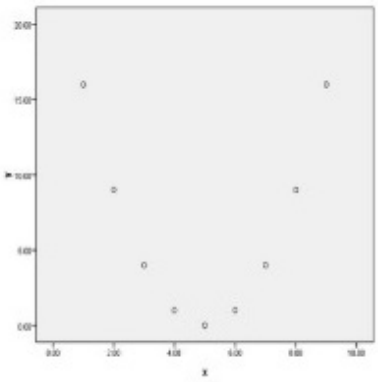
r = -0,45
středně silná
záporná korelace



r = 0,92
silná kladná
korelace

Poznámky:

- Korelační koeficient nesouvisí se sklonem proložené přímky kromě znamínka + a -
- Korelační koeficient je měřítkem lineárního vztahu a tedy hodnota r = 0 neznamená, že mezi proměnnými není žádný vztah. Například následující bodový graf má r = 0, což značí nulovou korelaci, nicméně proměnné mají dokonale kvadratický vztah.



r = 0
dokonale kvadratický vztah

Korelace je míra souvislosti a tak je možné sílu korelace popsat i verbálně. Použijeme Evansovu (1996) příručku, kterou navrhl pro absolutní hodnotu r:

- 0,00 - 0,19 „velmi slabá“
- 0,20 - 0,39 „slabá“
- 0,40 - 0,59 „střední“
- 0,60 - 0,79 „silná“
- 0,80 - 1,00 „velmi silná“

Například hodnota korelace r = 0,42 by byla „slabá kladná korelace“.

I.III Předpoklady

Výpočet Pearsonova korelačního koeficientu a další testy významnosti vyžadují následující předpoklady o datech:

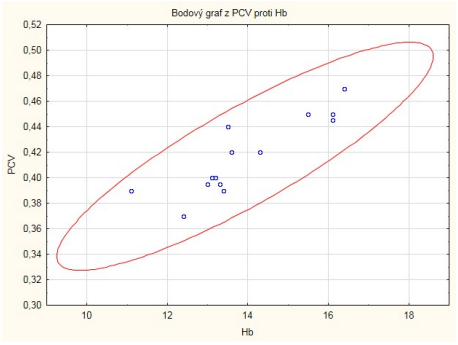
- Intervalový nebo poměrový charakter;
- Lineární vztah;
- Dvojrozměrné normální rozložení.

Pearsonův koeficient korelace je citlivý na zešikmení rozložení dat a na odlehlé hodnoty, při ověřování podmínek, bychom tedy měli dát důraz hlavně na tyto předpoklady.
Pokud naše data nesplňují výše uvedené předpoklady, použijeme Spearmanův koeficient pořadové korelace!

Příklad
Orientační ověření prvních dvou předpokladů jsme provedli výše, nyní se zabývejme ověřením předpokladu dvourozměrné normality. Pro orientační posouzení dvojrozměrné normality datového souboru může sloužit například 95% konfidenční elipsa (pokud alespoň 95% hodnot našeho výběrového souboru leží uvnitř této elipsy budeme předpokládat, že předpoklad dvourozměrné normality není porušen).

Podívejme se také na koeficienty šikmosti a zjistěme, zda naznačují zešikmení některé z proměnných.

Obě proměnné mají koeficienty šikmosti skutečně kladné. Rychlým posouzením závažnosti problému je ověřit, jestli jsou absolutní hodnoty koeficientů zešikmení menší než dvojnásobek příslušné směrodatné odchylky. V obou případech je toto splněno, což je v souladu s předpokladem



	Popisné statistiky		
Proměnná	N platných	Sm.odch.	Šikmost
Hb	14	1,557329	0,262206
PCV	14	0,029782	0,299063

normality dat. Nemáme tudíž žárné obavy ohledně normality dat a můžeme přistoupit ke korelační analýze.
Pro data o Hemoglobinu/PCV, software STATISTICA poskytuje následující výstup:
Hodnota Pearsonova korelačního koeficientu potvrzuje to, co bylo zřejmé z grafu, tedy že mezi

	Korelace N=14
Proměnná	PCV
Hb	,8770
	p=,000

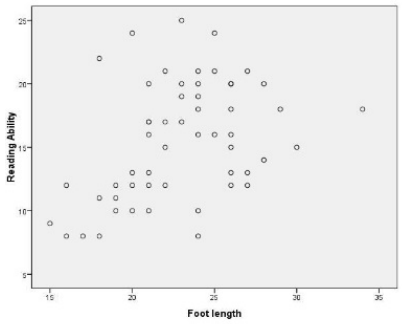
těmito proměnnými se zdá být pozitivní korelace.
Nicméně je nutné provést test významnosti, abychom rozhodli, jestli je možné na základě tohoto vzorku usuzovat na existenci lineární korelace v celé populaci.
Abychom toho dosáhli, provedeme test s nulovou hypotézou $H_0 : \rho = 0$, že v populaci žádná korelace není, proti alternativní hypotéze, $H_1 : \rho \neq 0$, která říká, že v populaci tato korelaci existuje. Na základě datového souboru zamítneme či nezamítneme nulovou hypotézu.
Tedy nulová hypotéza říká, že v populaci lineární korelace neexistuje proti alternativní hypotéze, že lineární korelace v populaci existuje. STATISTICA vypočítala p-hodnotu tohoto testu jako 0,000 a tedy můžeme říct, že máme velmi silný důkaz ve prospěch H1, tedy máme silný důvod věřit, že Hb a PCV jsou v ženské populaci lineárně korelované.
Významný Pearsonův koeficient s hodnotou 0.877 potvrzuje to, co bylo zřejmé z grafu; zdá se, že mezi těmito proměnnými je velmi silná pozitivní korelace. Tedy vyšší hodnoty Hb se spojují s vyššími hodnotami PCV.

Toto by mohlo být formálně zapsáno takto:
„Pearsonova korelace byla použita k určení vztahu mezi 14 hodnotami Hb a PCV měřených na ženách. Byla zjištěna velmi silná pozitivní korelace mezi Hb a PCV (r=0,88, N = 14, p< 0,001).“

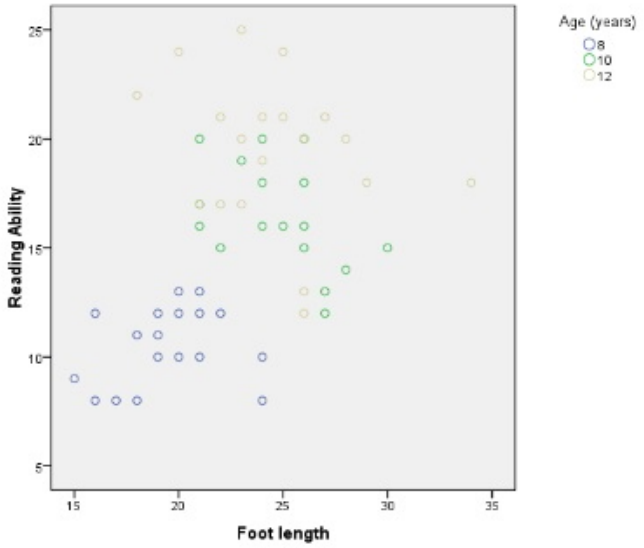
I.IV Upozornění

Existence silné korelace neznamená kauzální vztah mezi proměnnými. Například nemůžeme předpokládat, že hodnoty Hb určují hodnoty PCV nebo naopak.

Také bychom si měli být vědomi, že je možná existence skryté nebo intervenující proměnné. Například můžeme uvažovat vztah mezi schopností číst (Reading Ability) a velikostí nohou (Foot lenght) u dětí. Bodový graf a analýza korelace dat ukazují, že mezi nimi existuje velmi silná korelace (r= 0,88, N=54, p=0,003):
Nicméně pokud vezmeme v úvahu věk dítěte, vidíme, že tato očividná korelace by mohla být

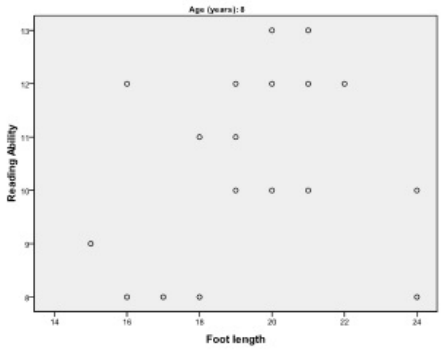


jen zdánlivá.
Pokud nyní znovu prozkoumáme data podle věkových skupin, skutečně zjistíme, že v každé



skupině není viditelná žádná korelace mezi schopností číst a velikostí nohou u dětí (tento příklad byl zpracován v softwaru SPSS).

Age (years) = 8

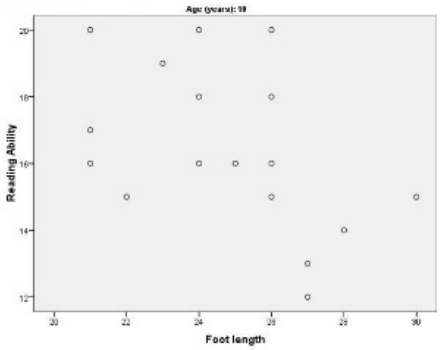


Correlations[a]

		Reading Ability	Foot length
Reading Ability	Pearson Correlation	1	.210
	Sig. (2-tailed)		.403
	N	18	18
Foot length	Pearson Correlation	.210	1
	Sig. (2-tailed)	.403	
	N	18	18

a. Age (years) = 8

Age (years) = 10

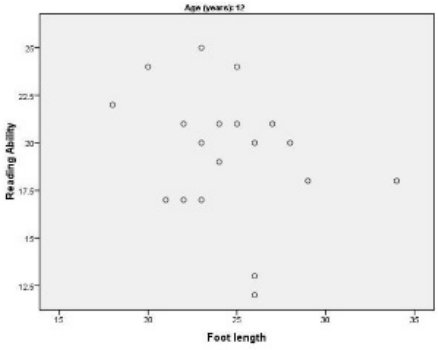


Correlations[a]

		Reading Ability	Foot length
Reading Ability	Pearson Correlation	1	-.465
	Sig. (2-tailed)		.060
	N	17	17
Foot length	Pearson Correlation	-.465	1
	Sig. (2-tailed)	.060	
	N	17	17

a. Age (years) = 10

Age (years) = 12



Correlations[a]

		Reading Ability	Foot length
Reading Ability	Pearson Correlation	1	-.290
	Sig. (2-tailed)		.228
	N	19	19
Foot length	Pearson Correlation	-.290	1
	Sig. (2-tailed)	.228	
	N	19	19

a. Age (years) = 12