7 Analýza závislostí

Statistická analýza se zřídka zabývá pouze jednou izolovanou proměnnou. Častěji se zajímáme o srovnání několika rozdělení, o změny proměnné v čase nebo vztahy mezi proměnnými. V předchozí kapitole jsme se zabývali porovnáváním rozdělení jedné proměnné mezi dvěma skupinami nebo mezi dvěma různými časovými okamžiky. V této kapitole se koncentrujeme na základní metody pro studium závislosti mezi proměnnými. Takové metody budou užitečné, jestliže nás zajímá:

- predikce budoucích zisků z prodeje produktu v závislosti na jeho ceně;
- závislost redukce váhy na počtu týdnů, kdy se jedinec se podrobí dietnímu režimu s příjmem určitého energetické hodnoty;
- jak výška dítěte v šesti letech predikuje jeho výšku v šestnácti letech;
- jak ovlivňuje spotřeba alkoholu snížení tělesné teploty apod.

Vztahy, které jsme uvedli, nemají ryze funkčně deterministický charakter. Proto je nutné použít pro jejich analýzu statistické metody. Příslušná oblast statistiky se nazývá korelační a regresní analýza. Stejně jako v jednorozměrné analýze začneme s grafickou analýzou, pak přejdeme k shrnujícím numerickým charakteristikám dat. Při výkladu budeme klást důraz na ty vztahy mezi proměnnými, jež lze popsat přímkou (lineární závislostí). Výklad později rozšíříme na jednoduché nelineární vztahy. Postupy této kapitoly tvoří základ pro metody vícerozměrné statistiky (kap. 10, resp. kap. 13). Hodnocením vztahů mezi kategoriálními proměnnými se budeme zabývat v následující kapitole 8.

Korelační analýza zkoumá vztahy proměnných graficky a pomocí různých měr závislosti, které nazýváme korelační koeficienty. **Regresní analýza** dává odpovědi na otázky typu: jaký vztah existuje mezi proměnnými X a Y (lineární, kvadratický atd.), lze proměnnou Y odhadnout pomocí proměnné X a s jakou chybou? Statistická analýza v těchto souvislostech má následující cíle:

a) poskytnout číselné míry vztahu dvou proměnných podobným způsobem, jako průměr a směrodatná odchylka popisují chování jedné proměnné;
b) najít vzorce pro optimální predikci proměnné, kterou považujeme za závisle proměnnou;

c) ohodnotit chybu predikce;
d) ověřovat různé hypotézy o zkoumaném vztahu.

Korelační a regresní analýza má intelektuální kořeny v práci Francise Galtona (1894). Inspirován částečně dílem svého bratrance Charlese Darwina *O původu druhů* snažil se Galton odhalit dědičné vlastnosti talentu, pohybových schopností a intelektu. Jeho výzkum začal studiem velikosti a váhy bílého hrachu ve dvou generacích. Ten vedl k odhalení fenoménu regrese, kterou Galton nazýval „reverse" (návrat). Popíšeme ho podrobněji v kapitole 7.4. Galton zjistil, že potomci extrémně velikých hrachů nebyli v průměru tak extrémně velicí, jako jejich „rodiče". Později zaznamenával fyzické charakteristiky tisíců dobrovolníků a nalezl podobnou „regresi" k průměru při mezigeneračním srovnání. Jeho koncepty pak rozvinul jeho žák K. Pearson a další statistici. Tradiční název „regrese" se stále používá v důsledku tradice, ačkoli se tato statistická technika spíše používá pro predikci. Metodu nejmenších čtverců pro odhad regresních funkcí používali již A. M. Legendre a C. F. Gauss.

7.1 Zobrazení dvojrozměrných dat

Základní postup dvojrozměrné analýzy dat je podobný jako v jednorozměrném případě:

1. Nejdříve se pokusíme zobrazit data graficky.
2. Hledáme základní konfigurace a tendence v datech.
3. Přidáváme numerické charakteristiky různých aspektů dat.
4. Často se nám podaří vystihnout stručným způsobem základní konfiguraci dat pomocí pravděpodobnostního modelu.

Data máme v podobě určitého počtu číselných dvojic údajů (x_i, y_i), které jsme získali měřením proměnných X a Y. Vždy je na místě provést před dalším zpracováním jejich grafickou interpretaci. Vynesením dat do souřadnicového systému (např. na milimetrový papír nebo zobrazením na displeji počítače pomocí vhodného programu) získáme základní představu o společném rozdělení obou proměnných. Každý bod odpovídá jednomu páru měření. Takovému grafickému znázornění říkáme dvojrozměrný bodový graf. Jeho prohlídkou odhadneme, zda je mezi proměnnými přesná funkcionální závislost, případně volnější vztah, jejž nazýváme statistická závislost, anebo jestli jsou na sobě evidentně nezávislé. Na obrázku 7.1 jsou znázorněny hodnoty měření výšky a váhy deseti studentů. Graf zobrazuje údaje z tabulky 7.1. Data se také zobrazují pomocí tzv. korelační tabulky (tab. 7.2). Dochází přitom k určité ztrátě informací rozdělením dat do intervalů.

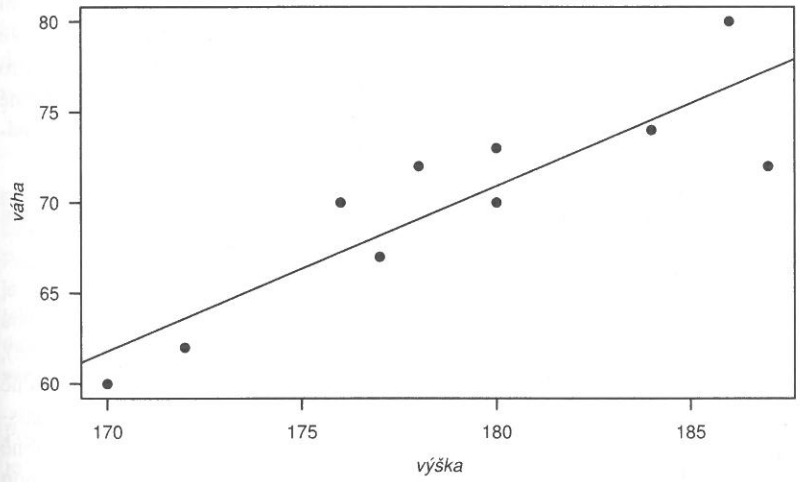
Tab. 7.1 Příklad dat, jejichž závislost chceme posoudit

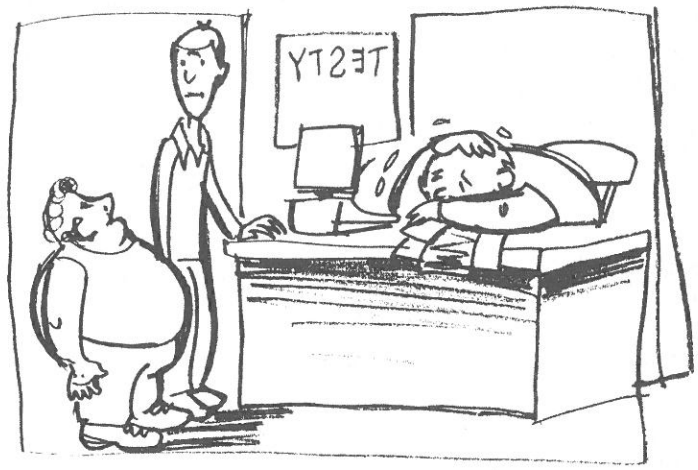
Výška [cm]	187	170	180	184	178	180	172	176	186	177
Váha [kg]	72	60	73	74	72	70	62	70	80	67

Tab. 7.2 Příklad korelační tabulky – korelace zjištěných hodnot výšky a váhy deseti studentů

Váha [kg]	**Výška [cm]**			**Celkem**
	do 170	**170–180**	**180–190**	
do 60	1	0	0	1
60–70	0	4	0	4
70–80	0	2	3	5
Celkem	1	6	3	10

Obr. 7.1 Bodový graf pro posouzení závislosti mezi váhou a výškou studentů





„Prý jsme mu zkazili jeho pozitivní korelaci mezi výškou a váhou."

Cílem regresní a korelační analýzy je popis statistických vlastností vztahu dvou nebo více proměnných. Dvojrozměrný bodový graf nebo korelační tabulka dávají první představu o rozdělení sledovaných proměnných. Graf často indikuje překvapivé vlastnosti dat jako nelinearitu vztahu, nehomogenitu nebo přítomnost odlehlých hodnot. Na obrázku 7.1 je rovněž vynesena přímka, která byla proložena body metodou nejmenších čtverců. Vliv třetí proměnné na rozložení bodů můžeme zachytit různým tvarem nebo barvou bodů v závislosti na hodnotě této proměnné (např. u dat o výšce a váze bychom mohli použít různé značky pro body odpovídající chlapcům a dívkám, pokud bychom tuto informaci o proměnné *pohlaví* měli k dispozici). Některé možné konfigurace dat v grafu popíšeme v následujícím odstavci.

7.2 Korelační analýza

V nejobecnějším smyslu, slovo „korelace" označuje míru stupně asociace dvou proměnných. Říká se, že dvě proměnné jsou korelované (resp. asociované), jestliže určité hodnoty jedné proměnné mají tendenci se vyskytovat společně s určitými hodnotami druhé proměnné. Míra této tendence může sahat od neexistence korelace (všechny hodnoty proměnné Y se vyskytují stejně pravděpodobně s každou hodnotou proměnné X) až po absolutní korelaci (s danou hodnou

proměnné X, se vyskytuje právě jedna hodnota proměnné Y). Pro měření korelace byla navržena řada koeficientů. Liší se podle typů proměnných, pro které se využívají. Statistické usuzování o korelačních koeficientech se opírá o teorii pravděpodobnosti pro společné rozdělení dvou nebo více náhodných proměnných.

Při zkoumání korelačních vztahů má rozhodující význam kvalitativní rozbor příslušného materiálu. Nemá smysl měřit závislost tam, kde na základě logické úvahy nemůže existovat. Často je zbytečné měřit závislosti i z jiných důvodů. Je to zejména tehdy, když je korelace způsobena: a) formálními vztahy mezi proměnnými; b) nehomogenitou studovaného základního materiálu; c) působením společné příčiny.

Formální korelace vzniká např. tehdy, když se zjišťuje korelace procentuálních charakteristik, jež se navzájem doplňují do 100 % (např. korelace procentního zastoupení bílkovin a tuku v potravinách).

Jestliže populace, kterou studujeme, obsahuje subpopulace, pro něž se průměrné hodnoty proměnných X a Y liší, vypočtené korelační vztahy jsou touto **nehomogenitou** silně ovlivněny a jejich hodnoty nepopisují skutečný vztah mezi uvažovanými proměnnými. Nehomogenita materiálu se projeví na bodovém grafu tak, že shluky bodů pro subpopulace se budou nacházet v různých oblastech souřadnicového systému. Na obrázku 7.2 je modelově ukázáno působení nehomogenity. Ta má za důsledek, že korelačním koeficientem hodnotíme bez diferenciace najednou dva shluky bodů, které přísluší k různým populacím. Na obrázku a) to vede k nenulovém korelačními koeficientu i přesto, že v obou shlucích jsou proměnné nekorelované, naopak proměnné na obrázku b) jsou v obou shlucích proměnné korelované, ale celková korelace je nulová.

Příkladem **korelací způsobených společnou příčinou** jsou vztahy mezi některými mírami těla, např. mezi délkou pravé a levé ruky. Jiným známým příkladem jsou zdánlivé korelace způsobené časovým faktorem nebo faktorem modernizace u dvou řad údajů.

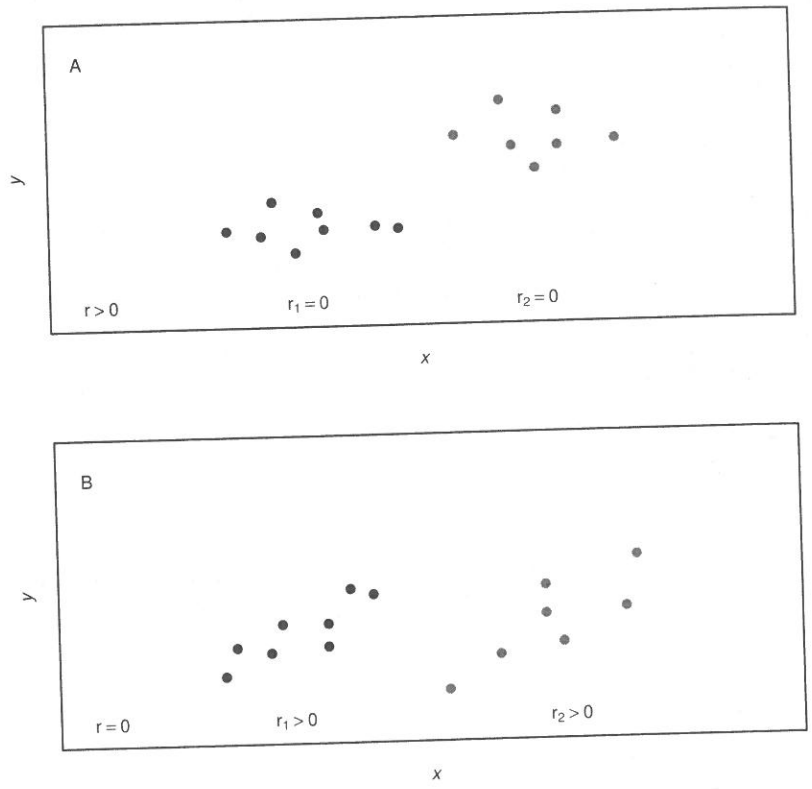
PŘÍKLAD 7.1

Zdánlivé korelace

Počet televizních přístrojů na osobu koreluje s očekávanou délkou života. Ve státech, kde je mnoho televizních přístrojů, dosahují obyvatelé vysokého věku. Je možné změnou počtu televizních přístrojů dosáhnout prodlužení věku v oblastech světa, kde je nižší očekávaná délka života?

Podobným korelacím se někdy říká „nesmyslné" korelace. Hodnota korelace je vysoká. Nesmyslný by byl závěr o příčinném působení. Korelační závislost

Obr. 7.2 Příklad kladné (A) a nulové (B) korelace, které jsou způsobené nehomogenitou dat

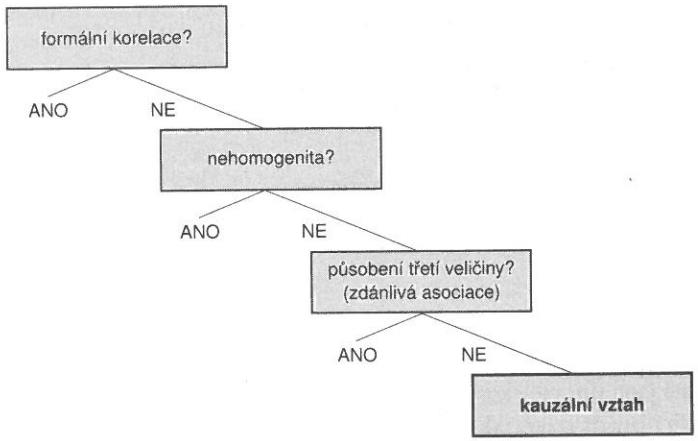


Korelační koeficient r je vypočtený pro všechny body, koeficienty r_1 a r_2 odděleně pro každý shluk zvlášť

je zdůvodněna proměnnou „národní důchod", jež je společnou příčinou obou proměnných.

Kromě tohoto působení proměnné jako „společné příčiny" mohou působit matoucí (rušivé) proměnné, které korelují jak s cílovou proměnnou, tak s proměnnou ovlivňující. Proměnná v tomto případě znesnadňuje interpretaci, protože nelze rozlišit vliv matoucí a sledované ovlivňující proměnné na cílovou proměnnou. Uvádíme pořadí, v němž máme vylučovat nezajímavé korelace, než se dostaneme do fáze, kdy by velká korelace mohla indikovat kauzální vztah (obr. 7.3).

Obr. 7.3 Postup pro ověření kauzálního vztahu



7.2.1 Pearsonův korelační koeficient

Přes některé své nedostatky zůstává Pearsonův korelační koeficient r nejdůležitější mírou síly vztahu dvou náhodných spojitých proměnných X a Y. Počítáme jej z n párových hodnot $\{(x_i, y_i)\}$ změřených na n jednotkách náhodně vybraných z populace. Korelační koeficient r nabývá hodnot z intervalu $[-1; 1]$. Jestliže má hodnotu 1 nebo -1, pak y-souřadnici bodu lze přesně spočítat pomocí lineárního vztahu z jeho x-souřadnice. Korelační koeficient r počítáme pomocí tzv. kovariance s_{xy} a směrodatných odchylek s_x a s_y obou proměnných:

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Vzorec s kovariancí pomáhá porozumět tomu, že r má kladnou hodnotu, pokud asociace proměnných je pozitivní. Dejme tomu, že studujeme korelaci výšky a váhy studentů. Jedinci, kteří mají hodnotu výšky nad průměrem, mívají nadprůměrnou i hodnotu váhy. Oba rozdíly od průměru, jež spolu násobíme při výpočtu kovariance, budou mít u vyšších a těžších jedinců kladnou hodnotu. Jedinci, kteří mají menší výšku, mají obvykle i menší váhu. U nich jsou oba

rozdíly od průměrů záporné, a proto je součin rozdílů od průměru rovněž kladný. Protože je většina sčítanců kladných, musí být kladná i výsledná hodnota kovariance a tedy i korelačního koeficientu. Tuto interpretaci lze ještě lépe pochopit při výpočtu r pomocí standardizovaných hodnot. Platí totiž vzorec

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \frac{\sum_{i=1}^{n} x_i' y_i'}{n-1},$$

kde x' a y' označují standardizované hodnoty.

Důležité vlastnosti Pearsonova korelačního koeficientu r shrneme pomocí několika tvrzení:

1. Platí $-1 \leq r \leq 1$.
2. Jestliže $|r| = 1$, leží všechny body na nějaké přímce.
3. Jestliže $r = 0$, nazýváme X a Y nekorelované proměnné. Dvě náhodné proměnné jsou tím více korelovány, čím blíže je hodnota r k číslům 1 nebo -1. V tom případě lze vztah obou proměnných dobře vyjádřit přímkou.
4. Jestliže $r < 0$, resp. $r > 0$, tak se Y v průměru zmenšuje, resp. zvětšuje při zvětšování proměnné X. Říkáme, že je asociace je záporná, resp. kladná.
5. Pearsonův korelační koeficient vyjadřuje pouze sílu *lineárního* vztahu. Špatně měří jiné vztahy, ať jsou jakkoli silné.
6. Korelační koeficient se nezmění, když změníme jednotky měření proměnných X a Y.
7. Podobně jako průměr nebo směrodatná odchylka, je korelační koeficient r velmi ovlivněn odlehlými hodnotami.
8. Korelační koeficient r nerozlišuje mezi závisle a nezávisle proměnnou.
9. Korelační koeficient r není úplným popisem dat i při velmi silném lineárním vztahu. Pro úplnější popis potřebujeme znát rovnici přímky, která vyjadřuje tvar vztahu.
10. Pokud jedna z proměnných nemá náhodný charakter (její hodnoty jsou pevně určeny), není vhodné korelační koeficient použít.
11. Korelace, ať je jakkoli silná, *neznamená* sama o sobě průkaz příčinného vztahu, tedy toho, že změny proměnné X skutečně působí změny proměnné Y.

Mezi proměnnými mohou existovat nejrůznější vztahy a máme i různé způsoby, jak je měřit. Některé z nich popíšeme v dalších odstavcích. Ačkoli korelační koeficient se používá velmi často, je nutné mít na paměti jeho omezení.

PŘÍKLAD 7.2

Výpočet korelačního koeficientu

Budeme hodnotit závislost výšky a váhy, jejichž hodnoty jsme naměřili u 10 studentů. Vypočítáme korelační koeficient pro párové hodnoty, které jsou uvedeny spolu s potřebnými dopočítanými hodnotami v tabulce 7.3. Hodnoty jsou zobrazeny na obr. 7.1 (s. 239).

Součet v posledním sloupci je základem pro výpočet kovariance

$$\text{cov}(x, y) = s_{xy} = 259/(10-1) = 28{,}8.$$

Dále jsme zjistili: $\bar{x} = 1790/10 = 179$; $\bar{y} = 700/10 = 10$; $s_x = 5{,}61$; $s_y = 5{,}83$. Korelační koeficient má tedy hodnotu $r = 28{,}8/(5{,}61 \times 5{,}83) = 0{,}88$.

Tab. 7.3 Příklad postupu výpočtu korelačního koeficientu

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
	187	72	8	2	16
	170	60	−9	−10	90
	180	73	1	3	3
	184	74	5	4	20
	178	72	−1	2	−2
	180	70	1	0	0
	172	62	−7	−8	56
	176	70	−3	0	0
	186	80	7	10	70
	177	67	−2	−3	6
Součet	1790	700	0	0	259

Někdy se zařazují hodnoty korelace do pásem podle síly asociace. V tabulce 7.4 uvádíme jeden z návrhů. Interpretace hodnot korelačního koeficientu není tak přímočará, jako je tomu u většiny jednorozměrných charakteristik. Proto se doporučuje dopočítat další charakteristiky, jako jsou parametry proložené přímky nebo směrodatná chyba odhadu při regresi (viz další kapitola).

Tab. 7.4 Pásma síly asociace podle velikosti korelačního koeficientu r

Síla asociace	$\|r\|$
malá	0,1–0,3
střední	0,3–0,7
velká	0,7–1,0

Hodnota korelačního koeficientu je bohužel silně ovlivňována odlehlými hodnotami ve výběru. Zkreslení také nastane, když se při výběru objektů omezíme pouze na ty, jejichž hodnota proměnné X nebo Y musí ležet v určitém intervalu. Korelační koeficient r má pak tendenci být menší než korelace r' vypočítaná bez omezení kladeného na data. Pro úpravu zkresleného korelačního koeficientu vlivem omezení rozsahu měření proměnné X použijeme vzorec

$$r' = \frac{Ur}{\sqrt{(U^2 - 1)r^2 + 1}},$$

kde $U = s/s'$ je poměr směrodatné odchylky s měření X ve studii a směrodatné odchylky s' v populaci bez restrikce.

Korelační koeficient je také ovlivněn nepřesností metod, kterými měříme obě proměnné. Jestliže známe r_{yy} a r_{xx} koeficienty spolehlivosti měření obou proměnných (jedná se korelace opakovaných měření), lze se přiblížit hodnotě korelačního koeficientu bezchybně změřených proměnných $r_{x'y'}$ pomocí úpravy

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}.$$

PŘÍKLAD 7.3

Význam exploračního zobrazení dvojrozměrných dat

Jednoduchým příkladem toho, jakou důležitou roli hraje explorační zobrazení dat, je zkoumání čtyř sérií modelových dat podle Anscomba (1973), které uvádí tabulka 7.5. Základní statistické charakteristiky proměnných X a Y a jejich korelační koeficient mají pro první sérii dat hodnoty $\bar{x} = 9{,}0$; $s_x = 3{,}31$; $\bar{y} = 7{,}5$; $s_y = 2{,}03$ a $r = 0{,}816$. Pokud spočteme tyto charakteristiky pro ostatní série, zjistíme, že jsou stejné. Pokud však všechny čtyři série zobrazíme graficky (viz obr. 7.5a-d, s. 260), výsledek je dost překvapivý.

Tab. 7.5 Série modelových dat se stejnými základními statistickými charakteristikami a korelačními koeficienty

x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,1	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,1	4	5,39	19	12,5
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

7.2.2 Pravděpodobnostní rozdělení dvou náhodných proměnných

Teorie pravděpodobnosti popisuje nejen rozdělení jedné náhodné proměnné, ale i společná pravděpodobnostní rozdělení dvou nebo více náhodných proměnných. Této teorie je zapotřebí tehdy, když chceme navrhovat pravděpodobnostní modely vztahu proměnných a zdůvodnit procedury pro statistické usuzování v korelační a regresní analýze. V našem jednoduše pojatém výkladu budeme postupovat tak, abychom mohli získané výsledky využít i v kapitole o analýze závislosti kategoriálních proměnných.

Zatím jsme se seznámili s jednou dvojrozměrnou charakteristikou, s Pearsonovým korelačním koeficientem r. Teoretickou hodnotu Pearsonova korelačního koeficientu v populaci označujeme ρ. Získali bychom ji výpočtem z údajů o všech prvcích populace. Výběrový koeficient r je bodovým odhadem této hodnoty. S rostoucím rozsahem výběru n se hodnota výběrového korelačního koeficientu r_n blíží ke své teoretické hodnotě ρ.