8 Průzkumová analýza dat

Cílem **průzkumové analýzy dat** (také známé pod zkratkou EDA - z anglického názvu exploratory data analysis) **je nalezení zvláštností statistického chování dat a ověření jejich předpokladů pro následné statistické zpracování** (MELOUN - MILITKÝ 1994).

Proč tyto vlastnosti potřebujeme zkoumat? Většina běžně používaných statistických metod předpokládá určité vlastnosti zpracovávaných souborů nebo výběrů, nejdůležitější z nich jsou následující:

- minimální rozsah výběru,
- normalita (tj. splnění předpokladu, že výběr pochází ze základního souboru s normálním rozdělením),
- absenci silně vychýlených hodnot,
- vzájemná nezávislost prvků výběru.

Splnění těchto podmínek podmiňuje použití nejznámějších a nejpoužívanějších statistických charakteristik, tzv. momentových – aritmetického průměru, rozptylu, směrodatné odchylky, koeficientů špičatosti a šikmosti. Pouhé okulární posouzení - zvláště u velkých souborů dat - není průkazné a mnohdy ani technicky možné. Grafické a početní metody průzkumové analýzy dat mohou rozhodování o splnění různých předpokladů objektivizovat. Mnohé soubory měřených dat jsou zcela unikátní a často nelze (jak z technických, tak i z ekonomických důvodů) měření opakovat nebo doplnit. V těchto případech nám průzkumová analýza dat může poskytnout velmi cenné informace ještě před provedením vlastní statistické analýzy, upozornit na možné problémy a pomoci při volbě nejvhodnějších metod zpracování (neboť i statistická analýza stojí čas a peníze - a v neposlední řadě značnou práci - a chybně stanovené metody analýzy nebo její nesprávné provedení může mnohdy zcela znehodnotit důležitý a nákladný výzkumný nebo komerční projekt).

Průzkumová analýza dat je relativně moderní statistickou disciplínou, jejíž rozvoj je spojen s rozšířením výpočetní techniky. Většina postupů průzkumové analýzy dat je totiž založena na grafických metodách, které je možné efektivně provádět jen s použitím speciálních statistických programů. Výhodou těchto metod (oproti metodám početním) je jejich názornost, relativní nevýhodou je nutnost určité zkušenosti při jejich interpretaci. Proto je nejvhodnější kombinovat početní (testy) a grafické metody.

Průzkumová analýza dat využívá především robustních kvantilových charakteristik (o nich podrobněji v kapitole 4.1 v I. dílu). Základem pro konstrukci kvantilových charakteristik je **pořádková statistika,** což jsou vzestupně uspořádané prvky souboru $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$. Pokud budou v dalším textu indexy označující jednotlivé prvky v závorce - $x_{(1)}$ - bude se jednat o pořádkovou statistiku. Z takto upraveného souboru je možné konstruovat kvantilové charakteristiky. Obecně platí, že střední hodnota i-té pořádkové statistiky je rovna $100P_i$ procentnímu kvantilu, což je hodnota pod kterou leží $100P_i$ procent prvků souboru. Určitým kvantilem je tedy každý prvek souboru. Hodnota P_i se nazývá **pořadová pravděpodobnost.** Obecně se P_i stanoví takto

$$P_i = \frac{i}{n+1}.$$

Pro účely průzkumové analýzy dat se obvykle P_i volí (MELOUN - MILITKÝ 1994)

$$P_i = \frac{i - \frac{1}{3}}{n + \frac{1}{3}}$$

V průzkumové analýze dat se používá vybraných kvantilů pro pořadové pravděpodobnosti $P_i = 2^{-i}$ pro i = 1, 2, 3, 4. Vzhledem k tomu, že se tyto vybrané kvantily obvykle označují písmeny, nazývají se **písmenové hodnoty**. Jejich přehled je v tabulce 8.1 .

i	*i-tý* kvantil	P_i	Písmeno
1	medián	$2^{-1} = 1/2$	M
2	kvartily	$2^{-2} = 1/4$	F
3	oktily	$2^{-3} = 1/8$	E
4	sedecily	$2^{-4} = 1/16$	D

Tabulka 8.1 - Přehled základních kvantilů používaných v průzkumové analýze dat a jejich písmenové ekvivalenty

Pro odhad písmenových hodnot se používá technika **pořadí a hloubek.** Každá z uspořádaných hodnot $x_{(i)}$ je určena trojicí $\{K_i, R_i, H_i\}$, kde je

$K_i = i$ rostoucí pořadí (pořadové číslo pořádkové statistiky počítané od nejmenšího prvku);

$R_i = n + 1 - i$ klesající pořadí (kde n je celkový počet prvků);

$H_i = \min\{ K_i, R_i\}$ hloubka pořádkové statistiky (je to menší z hodnot K_i, R_i).

Potom platí, že hloubka mediánu je

$$H_M = \frac{n+1}{2}.$$

Pokud tato hodnota není celé číslo, provádí se lineární interpolace mezi dvěma prostředními prvky souboru. Hloubky dolních písmenových hodnot jsou

$$H_L = \frac{1 + \mathrm{int}(H_{L-1})}{2},$$

kde L je obecné označení kvantilu (L = M, F, E, D), int (x) je celočíselná část x. Označení L - 1 značí vždy „předchozí“ kvantil, tj. D - 1 = E, E - 1 = F, F - 1 = M.

Pokud je H_L celé číslo, potom platí, že dolní kvantil se rovná

$$L_D = x_{(H_L)}$$

a horní kvantil

$$L_H = x_{(n+1-H_L)}$$

Příklad 8.1

Vyčíslete písmenové hodnoty pro zadanou číselnou řadu o 19 prvcích.

$x_{(i)}$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
R_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
K_i	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
H_i	1	2	3	4	5	6	7	8	9	10	9	8	7	6	5	4	3	2	1

Tabulka 8.2 - Metoda pořadí a hloubek

V tabulce 8.2 jsou vyčísleny hodnoty pořádkové statistiky, rostoucího a klesajícího pořadí a hloubky pro jednoduchou číselnou řadu čísel 1 - 19. Vidíme, že největší hloubku (10) má prostřední prvek souboru - medián. Jeho hloubka je (19 + 1)/2 = 10. Ostatní kvantily se získají podle výše uvedených vzorců. Např. pro kvartil platí - (1 + 10)/2 = 5,5, tj. musíme interpolovat mezi 5. a 6. prvkem. To je hodnota dolního kvartilu, horní kvartil je roven 19 +1 - 5,5 = 14,5, tj. interpolujeme mezi 14. a 15. prvkem. Podobně vypočítáme oktil s použitím hloubky kvartilu a sedecil s využitím hloubky oktilu. Tabulka 8.3 uvádí příslušné písmenové hodnoty.

Kvantil	Dolní kvantil	Horní kvantil
Medián - M	10.000	10.000
Kvartil - F	5.500	14.500
Oktil - E	3.250	16.750
Sedecil - D	2.125	17.675

Tabulka 8.3 - Hodnoty písmenových hodnot pro zadanou číselnou řadu

8.1 Základní grafické metody průzkumové analýzy dat

Mezi základní úkoly průzkumové analýzy dat patří posouzení:

- stupně symetrie a špičatosti rozdělení,
- lokálních koncentrací dat,
- vybočujících měření,
- shody s teoretickým rozdělením (zpravidla s normálním).

Nejběžnějšími prostředky pro splnění těchto úkolů jsou speciální grafické metody, především

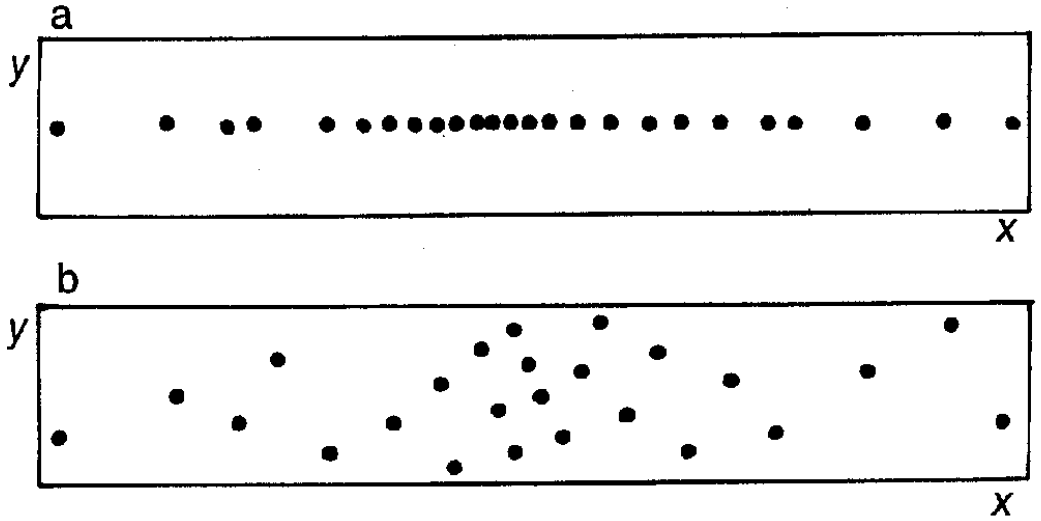
- diagram rozptýlení,
- rozmítnutý diagram rozptýlení,
- krabicový graf,
- vrubový krabicový graf,
- graf hustoty pravděpodobnosti,
- graf rozptýlení s kvantily.

Grafické metody mají oproti početním testům (např. testům normality, nezávislosti, apod.) určité výhody i nevýhody. Na jedné straně nedávají jednoznačné rozhodnutí o přijetí nebo odmítnutí určité hypotézy jako testy, o míře nesouladu s teoretickým rozdělením musí rozhodnout analytik na základě svých znalostí, ale na druhé straně jejich rozborem je možné postihnout příčiny nesouladu s určitým rozdělením (např. vliv šikmosti, špičatosti, odlehlých hodnot, je možné i detekovat směs více rozdělení apod.). Například při posuzování normality je statistický test na dané hladině významnosti průkazný, ale pouze nám zamítne nebo nezamítne nulovou hypotézu (tj. že výběr pochází nebo nepochází z normálního rozdělení), ale neanalyzuje příčiny. Vhodná grafická metoda průzkumové analýzy dat - v tomto případě např. kvantilový nebo rankitový graf - takto jednoznačnou informaci neposkytne (o míře normality musí rozhodnout hodnotitel), ale na druhé straně poskytne mnoho informací o možných příčinách nenormality (např. vybočující měření, šikmost apod.).Uvádí se také (MELOUN - MILITKÝ 1994), že grafické metody jsou citlivější, „přísnější“, než obvykle používané testy, kde jejich schopnost detekce závisí především na síle testu. Proto se doporučuje při posuzování výběrů pomocí grafických metod průzkumové analýzy dat obě skupiny metod kombinovat a závěry dělat až na základě posouzení výsledků obou skupin.

8.1.1 Graf rozptýlení

je v podstatě vynesení hodnot souboru na číselnou osu. I takto jednoduché grafické znázornění má daleko vyšší vypovídací hodnotu než pouhá řada čísel. Je možné rychle odhalit lokální koncentrace dat (velké nakupení hodnot v určitém úseku číselné osy) a podezřelé vybočující hodnoty (extrémně nízké nebo vysoké). Grafické schéma je na obrázku 8.1 .

Rozmítnutý graf rozptýlení je podobný jako předchozí a má i stejné použití. Body jsou však pomocí generátoru náhodných čísel ve vhodném měřítku „rozhozeny“ ve směru osy Y, aby v místech s velkou koncentrací hodnot nedocházelo k jejich splývání. Grafické schéma je na obrázku 8.1 .



Obrázek 8.1 – Schéma grafu rozptýlení a rozmítnutého grafu rozptýlení.

8.1.2 Krabicový graf

je jedním z nejběžnějších způsobů grafického znázornění dat. Je součástí většiny moderních statistických programů. Také se někdy můžeme setkat s názvem „vousatá krabička“ (z angl. názvu „box and whisker plot“). Umožňuje především

- znázornění robustního odhadu polohy – mediánu,
- posouzení symetrie rozdělení,
- identifikaci podezřelých odlehlých měření.

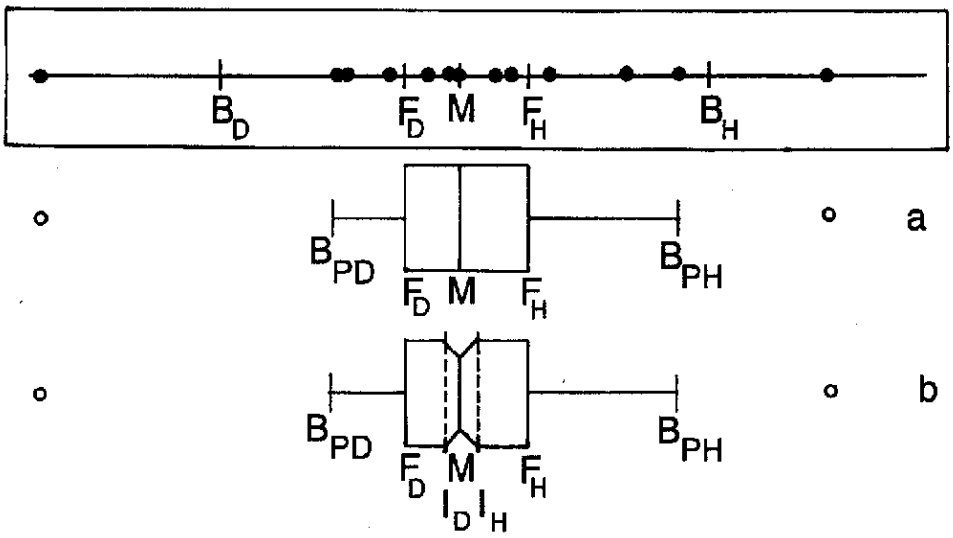
Jeho základem je obdélník s vhodně zvolenou šířkou a délkou rovnou **interkvartilovému rozpětí R_F** = F_H - F_D (tj. rozdílu horního a dolního kvartilu). Uvnitř obdélníku („krabičky“) je čára představující polohu **mediánu M**. Od obou protilehlých stran obdélníku pokračují úsečky („vousy“), které jsou ukončeny **přilehlými hodnotami - horní B_{PH} a dolní B_{PD}**. Přilehlé hodnoty jsou ty prvky souboru, které leží nejblíže **vnitřních hradeb souboru** - dolní hranice hradby B_D a horní hranice B_H. Tyto hodnoty se vypočítají $B_H = F_H + 1.5R_F$, resp. $B_D = F_D - 1.5R_F$. Samotné vnitřní hradby nejsou v grafu zpravidla znázorněny. **Koncové body úseček jsou tedy nejmenší a nejvyšší „bezproblémové“ hodnoty souboru**. **Body ležící mimo vnitřní hradby jsou považovány za „podezřelé“** (odlehlé, vybočující) a jsou graficky znázorněny (křížky, kolečky apod.) v příslušných vzdálenostech. Grafické schéma je na obrázku 8.2 .

8.1.3 Vrubový krabicový graf

je variantou předchozího grafu. Na „krabičce“ se vytvoří zářez, jehož šířka je rovna intervalu spolehlivosti mediánu (dolní hranice I_D, horní hranice I_H). Hranice se vypočítají podle vzorců

$$I_H = M + \frac{1{,}57 \cdot R_F}{\sqrt{n}} \qquad I_D = M - \frac{1{,}57 \cdot R_F}{\sqrt{n}}$$

Ostatní charakteristiky jsou stejné jako u krabicového grafu. Grafické schéma je na obrázku 8.2 .



Obrázek 8.2 - Obecné schéma krabicového (a) a vrubového krabicového grafu. (b). Nahoře je pro srovnání diagram rozptýlení s vyznačenými důležitými body pro konstrukci krabicových grafů. Prázdnými kolečky jsou vyznačena „vybočující“ měření. Symboly: M- medián, $F_{D(H)}$ – dolní

(horní) kvartil, $I_{D(H)}$ – dolní (horní) hranice intervalu spolehlivosti mediánu, $B_{D(H)}$ – dolní(horní) vnitřní hradba souboru (podle MELOUN - MILITKÝ 1994).

8.1.4 Graf rozptýlení s kvantily

je jeden z nejuniverzálnějších a také nejpoužívanějších průzkumových grafů. Na ose X se vynáší pořadová pravděpodobnost, na ose Y pořádková statistika. Základní tvar grafu vznikne spojením bodů $\{P_i, x_{(i)}\}$ lineárními úseky. Základní tvar pro normální rozdělení je sigmoidální, nejprve konkávní, potom konvexní. Ke zvýšení přehlednosti a vypovídací schopnosti grafu se zakreslují **kvantilové obdélníky** (pro kvartil, oktil a sedecil) a poloha mediánu. Každý obdélník má na ose X souřadnice dané hodnotami dolního a horního příslušného kvantilu (kvartil 0.25 a 0.75; oktil 0.125 a 0.875 a sedecil 0.0625 a 0.9375). Na ose Y jsou vynášeny příslušné pořádkové statistiky (tedy vzestupně uspořádané hodnoty). Vodorovné hrany kvantilových obdélníků nám tedy na ose Y ukáží hodnoty příslušných kvantilů. Bývá zde též zakreslen medián M včetně svého intervalu spolehlivosti.

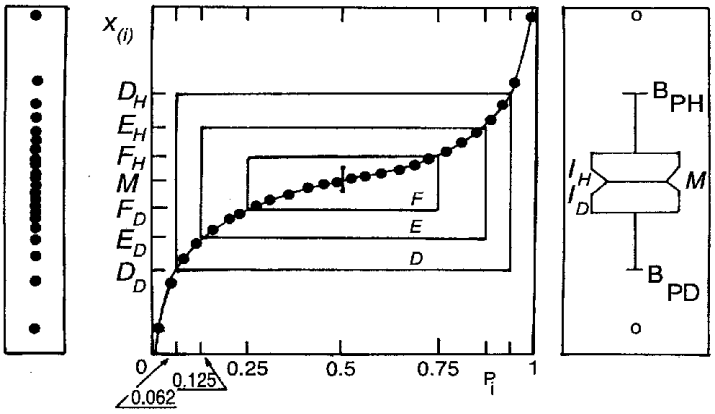
Pomocí grafu rozptýlení s kvantily se posuzuje zejména:

- sešikmenost rozdělení,
- modalita (unimodální - vícemodální rozdělení),
- odlehlé hodnoty.

Sešikmenost rozdělení se posuzuje podle vzájemné polohy kvantilových obdélníků. Symetrické rozdělení je charakterizováno tím, že jednotlivé obdélníky jsou symetricky jeden uvnitř druhého. Nejlepší kontrola je podle vzdálenosti dolních a horních stran příslušných obdélníků. Pokud se jedná o výrazně levostranné rozdělení (sešikmené k nižším hodnotám), potom jsou vzdálenosti mezi dolními stranami výrazně menší než mezi horními stranami. Je to způsobeno tím, že relativně stejný úsek souboru - např. 25% hodnot mezi dolním kvartilem a mediánem - je koncentrován do menšího rozpětí hodnot na ose Y. U pravostranného rozdělení je situace opačná - menší vzdálenosti jsou mezi horními stranami obdélníků.

Modus (nejčastěji se vyskytující hodnota v souboru) se pozná podle toho, že na kvantilové funkci je vytvořen „schod“ - úsek rovnoběžný s osou X. Je to způsobeno tím, že je zde koncentrováno více stejných hodnot. Vícemodální rozdělení mají takových stejných „schodů“ několik (nejpočetnější výskyt v souboru má více hodnot).

Odlehlé hodnoty identifikujeme tak, že na kvantilové funkci se projeví na pravém konci náhlý vzrůst (nebo pokles na levém konci). Grafické schéma je na obrázku 8.3 .



Obrázek 8.3 - Obecné schéma grafu rozptýlení s kvantily a jeho srovnání s grafem rozptýlení a krabicovým grafem. Vysvětlení symbolů viz v textu (podle MELOUN - MILITKÝ 1994).

8.1.5 Kvantil – kvantilový graf (Q-Q graf), normální pravděpodobnostní graf

Tento typ grafu **porovnává kvantily experimentálního a vybraného teoretického rozdělení** (tedy vlastně vzestupně uspořádané naměřené hodnoty a odpovídající hodnoty stanovené pomocí příslušné pravděpodobnostní funkce daného rozdělení). Jsou konstruovány tak, že pokud experimentální rozdělení plně odpovídá teoretickému, potom je grafem přímka. Jakékoli odchylky od tohoto „ideálního“ tvaru indikují odchylky od předpokládaného teoretického rozdělení. Q-Q graf lze sestrojit pro různá rozdělení, pouze se jinak stanovují příslušné hodnoty na osách X a Y. Podrobněji ke konstrukci Q-Q grafů pro vybraná známá rozdělení viz např. MELOUN - MILITKÝ 1994.

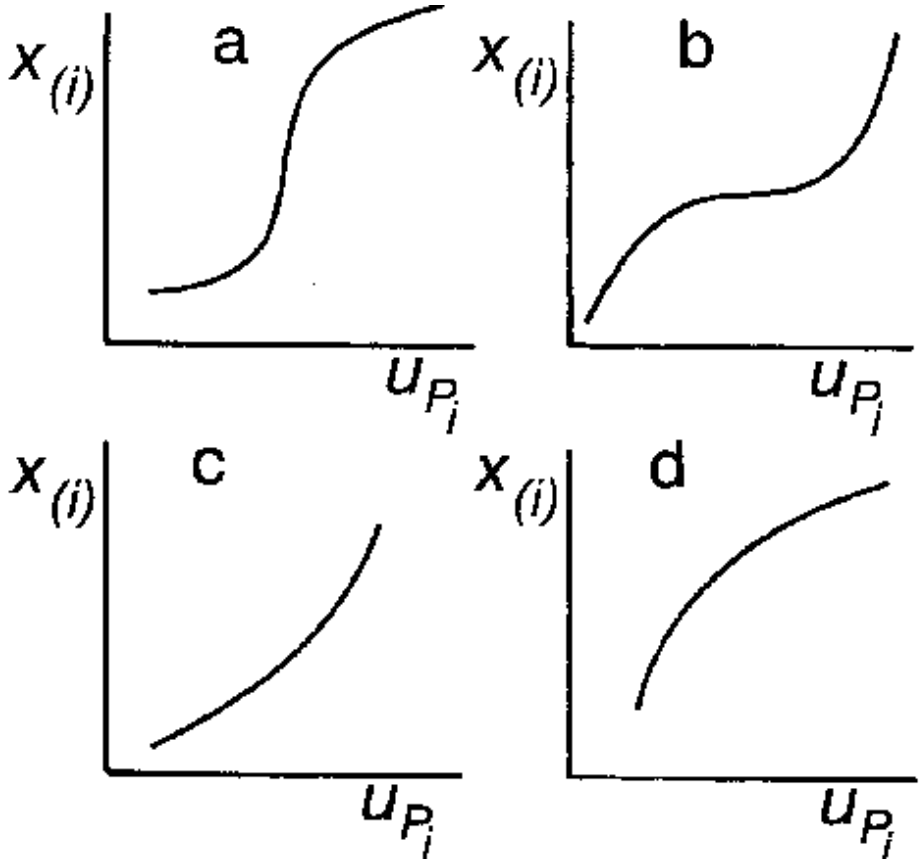
Speciálním případem Q-Q grafu pro normální rozdělení je **rankitový graf**.

Rankitový graf je konstruován tak, že na jedné ose jsou vynášeny kvantily normovaného normálního rozdělení u_{Pi} (to jsou tabelované hodnoty nebo je možné je získat např. v Excelu pomocí funkce NORMSINV) a na druhé ose pořádkové statistiky $x_{(i)}$. Pokud zkoumané rozdělení skutečně odpovídá normálnímu, potom je grafem přímka. Ve statistických programech je obvykle pro srovnání vykreslena srovnávací přímka, na které by ležely všechny body v případě ideální shody s normálním rozdělením. Na základě typických tvarů sestrojeného grafu, které jsou schématicky uvedeny na obrázku 8.4 , je možné soudit na hlavní příčiny odchylky od normality. Kromě těchto základních vzorů je možné také detekovat i jiné případy, např. silně odlehlá měření (odlehlý bod je daleko od ostatních, zpravidla mimo srovnávací přímku).

8.1.6 Graf hustoty pravděpodobnosti

Pojem hustoty pravděpodobnosti známe již z I. dílu, z kapitoly o 5.3 o funkcích náhodných proměnných. Víme tedy, že pro teoretická rozdělení je možné konstruovat tzv. frekvenční funkci, která se také nazývá (v případě spojitých veličin) hustota pravděpodobnosti. Tato funkce je velmi užitečná pro posouzení rozložení dat, pro detekci nehomogenity (výskyt více oblastí s vyšší koncentrací dat nebo odlehlých hodnot) ne-

bo sešikmení (nesouměrnost) rozdělení. Z toho vyplývá, že kdybychom byli schopni sestrojit graf hustoty pravděpodobnosti pro empirická data, porovnat jej s příslušným teoretickým (obvykle normálním) rozdělením, získali bychom velmi dobrý prostředek pro posouzení odchylek od příslušného teoretického rozdělení. Sestrojit frekvenční funkci teoretického rozdělení je možné jako derivaci distribuční funkce. Jak ale tuto funkci sestrojit pro empirická data, u nichž žádnou teoretickou funkci neznáme? Řešení nabízí technika nazývaná **jádrový odhad hustoty**.



Obrázek 8.4 – Základní tvary odchylek od normálního rozdělení v rankitovém grafu – rozdělení ploché (a), špičaté (b), levostranně nesouměrné (c) a pravostranně nesouměrné (d). POZOR! Tato interpretace platí pro uspořádání os, které je uvedeno na obrázku. Pokud jsou osy přehozeny (tj. na ose X jsou měřené hodnoty a na ose Y jsou očekávané kvantily normálního rozdělení) je interpretace opačná!!

Princip metody je poměrně jednoduchý, matematické provedení ale dost komplikované a její rutinní užití je možné pouze s využitím specializovaných statistických programů.

Vycházíme z následující myšlenky: pro každou z N empirických hodnot se sestrojí elementární křivka hustoty pravděpodobnosti s plochou pod křivkou 1/N, která se nazývá **jádro**. Toto jádro může mít teoreticky jakýkoli tvar, obvykle se používá frekvenční funkce normálního rozdělení (Gaussova křivka). Tyto elementární křivky se sečtou a výsledkem je křivka, která určitým způsobem modeluje rozložení empirických hodnot. Princip konstrukce je schématicky znázorněn na obrázku 8.5 . Je nutné zdůraznit, že se jedná o odhad rozložení hodnot, není to jednoznačně determinovaná

křivka, kterou by bylo možné vyjádřit nějakým jednoduchým vzorcem. Výsledný tvar závisí především na dvou faktorech:

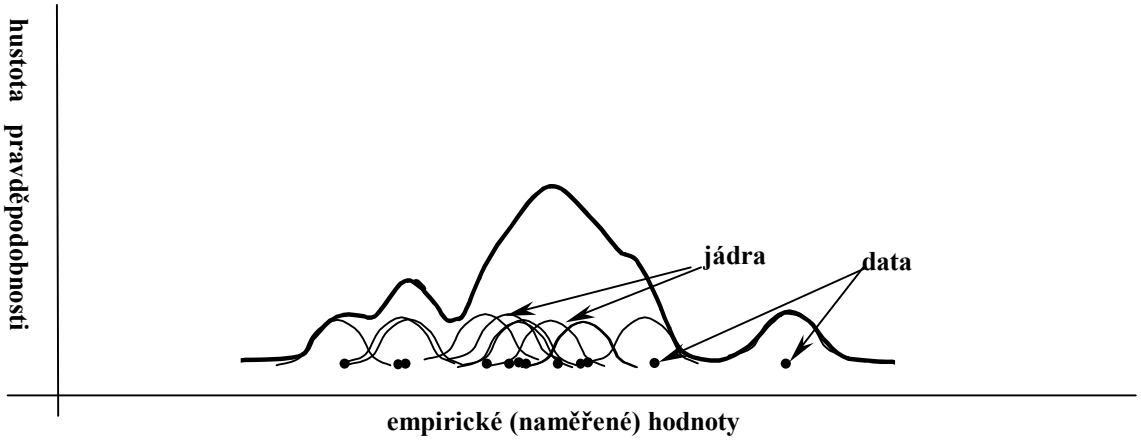
- tvaru jádra,
- šířce jádra.

Tvar jádra může být v podstatě libovolný, obvykle se používá normální rozdělení. Velmi důležitá je šířka jádra (tj. šířka elementárních funkcí sestrojených kolem datových bodů). Pokud je šířka malá, vypadá výsledná křivka jako pohoří s mnoha štíty a neposkytuje informaci o podstatných vlastnostech daného rozdělení. Naopak velká šířka způsobí, že křivka je velmi hladká a výsledek z hlediska interpretace je stejný nebo ještě horší než v případě malého (úzkého) jádra. Správný odhad šířky jádra vyžaduje určitou zkušenost, a v případě, že máme možnost šířku jádra volit, tak i experimentování. Některé programy umožňují tuto volbu, jiné se snaží o optimální odhad jádra na základě vestavěných (zpravidla iteračních) algoritmů, ale v obou případech si musíme uvědomit, že se jedná o odhad a výsledek není zcela objektivní. I přes uvedené nedostatky je graf hustoty pravděpodobnosti velmi oblíbeným diagnostickým nástrojem, především pro možnost rychlého a názorného porovnání empirických hodnot s teoretickým rozdělením. Uvádí se empirické pravidlo (KUPKA 1997), že při dostatečné velikosti výběru (N > 50) dvě výrazná maxima na grafu hustoty pravděpodobnosti svědčí o pravděpodobné nehomogenitě výběru a lze uvažovat o jeho rozdělení na dvě části. Výskyt velkého množství lokálních maxim svědčí obvykle o příliš úzkém jádru.

Naproti tomu použití tohoto grafu má také svá omezení. Nelze jej použít k odhadu kvantilů nebo ke konstrukci distribuční funkce.

Statistické programy, pokud tento graf mají ve své výbavě, obvykle jej vykreslují ve srovnání s normálním rozdělením.

Zájemci o matematickou formulaci konstrukce grafu, o postupy k vedoucí k určení šířky jádra najdou nejpoužívanější techniky např. v MELOUN-MILITKÝ 1994.



Obrázek 8.5 – Schéma konstrukce grafu hustoty pravděpodobnosti. Výsledná součtová křivka je znázorněna tučně.

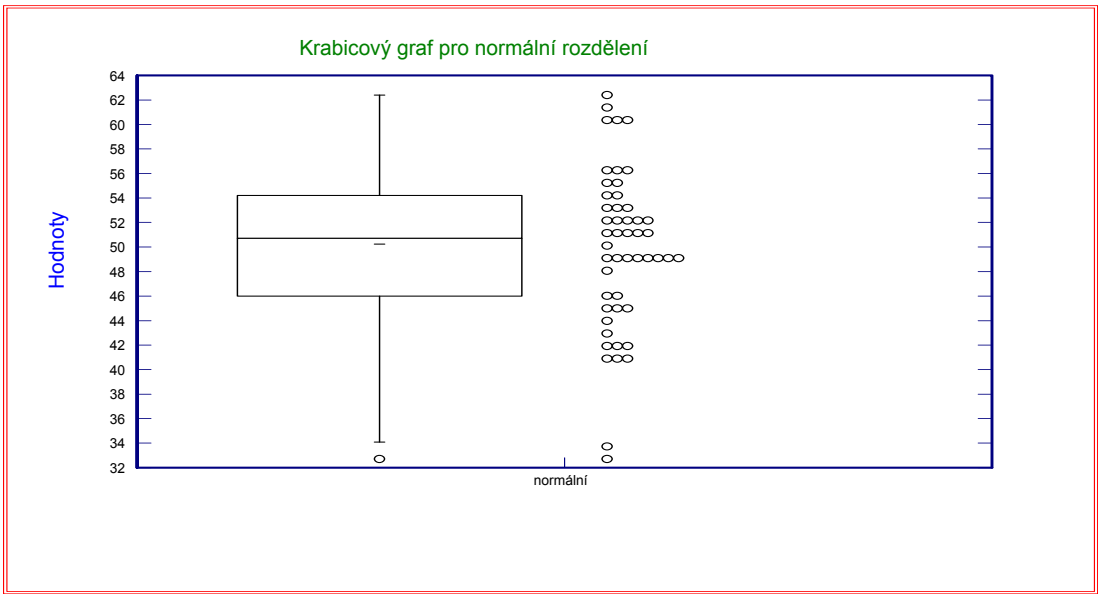
Příklad 8.2

Proveďte průzkumovou analýzu dat pro zadané soubory pomocí grafických metod.

Pro ilustraci provedení a interpretace průzkumové analýzy dat pomocí základních grafických metod byly generovány 3 výběry - podle rovnoměrného, normálního a exponenciálního rozdělení. Rozdělení byla vybrána tak, že kromě základního statistického rozdělení (normálního) se zde vyskytuje i rozdělení výrazně nesymetrické (exponenciální) a naopak rozdělení s velmi pravidelným rozložením hodnot v daném intervalu (rovnoměrné). Základní zadání je v tabulce 8.4.

Pro aplikaci průzkumové analýzy dat je nutné z prvotního zápisu udělat pořádkovou statistiku, tj. vzestupně uspořádaný výběr. Poté můžeme aplikovat výše popsané základní grafické metody.

Výsledek pro normální rozdělení je na obrázcích 8.6 , 8.7 , 8.8 a 8.9 . Z grafu rozptýlení (tečkového grafu) na obrázku 8.6 vidíme, že daný výběr vykazuje určité lokální koncentrace dat (skupiny nahloučených bodů). V oblasti dolních hodnot jsou dvě poměrně izolované hodnoty, ale z krabicového grafu je zřejmé, že se zřejmě nejedná o vybočující (extrémní) hodnoty, neboť pouze jedna vybočuje z vnitřních hradeb souboru, a to velmi těsně. Srovnání polohy mediánu a aritmetického průměru indikuje velmi dobrou shodu, což je typické právě pro normální rozdělení nebo symetrická rozdělení blízká normálnímu. Analýza kvartilů („krabičky“) naznačuje, že daný výběr bude zřejmě velmi mírně pravostranný, neboť dolní část „krabičky“ je o něco delší než horní, což znamená, že v úseku mezi mediánem a horním kvartilem (horní část krabičky) jsou data více koncentrována než v dolní části (tj. mezi mediánem a dolním kvartilem).



Obrázek 8.6 – Krabicový graf a graf rozptýlení pro generovaná data normálního rozdělení. Popis jednotlivých prvků grafu je v textu. Krátká čárka označuje polohu aritmetického průměru.

K podobným závěrům můžeme dojít pomocí grafu rozptýlení s kvantily. Jednotlivé kvantilové obdélníky jsou v podstatě symetrické, což indikuje prakticky symetrické rozložení bodů mezi jednotlivými významnými kvantily. Čára spojující jednotlivé hodnoty vykazuje určitou „stupňovitost“ danou právě lokálními koncentracemi dat.

Další dva grafy na obrázcích 8.8 a 8.9 umožňují kvalitně posoudit shodu s normálním rozdělením.

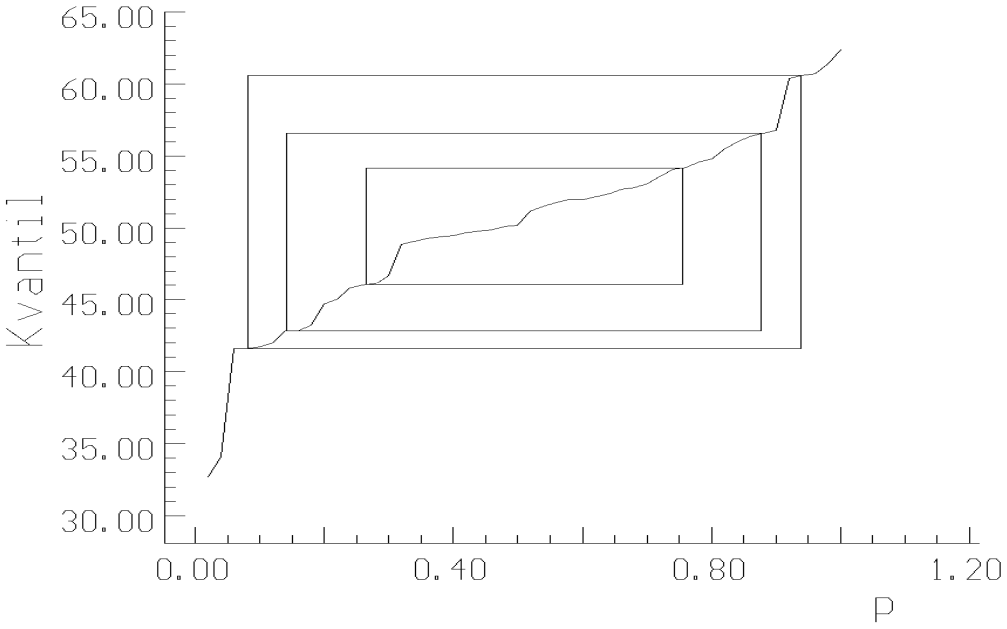
Kvantil-kvantilový graf vykazuje dobrou shodu, která je indikována tím, že jednotlivé body (kvantily) leží velmi těsně kolem srovnávací linie. Je nutné si uvědomit, že ideální shodu s přímkou nedosáhneme prakticky nikdy, jde v podstatě o míru těsnosti, s jakou se měřené (nebo v tomto případě generované) hodnoty přimykají srovnávací linii. Větší odchylku vykazují pouze dvě nejnižší hodnoty, ale vzhledem k tomu, že výběr je dostatečně velký (50 hodnot), zřejmě tato odchylka nebude mít větší vliv.

Tento závěr potvrzuje i graf hustoty pravděpodobnosti, kdy jádrový odhad hustoty empirické křivky (čárkovaně) se téměř shoduje s teoretickým průběhem normálního rozdělení vypočítaného pomocí aritmetického průměru a směrodatné odchylky výběru. Potvrzuje předpoklad velmi mírné špičatosti (empirická křivka je vyšší než teoretická, což indikuje vyšší koncentraci hodnot v oblasti tohoto vrcholu) a pravostranné nesouměrnosti (vrchol empirické křivky je mírně vpravo od teoretické křivky).

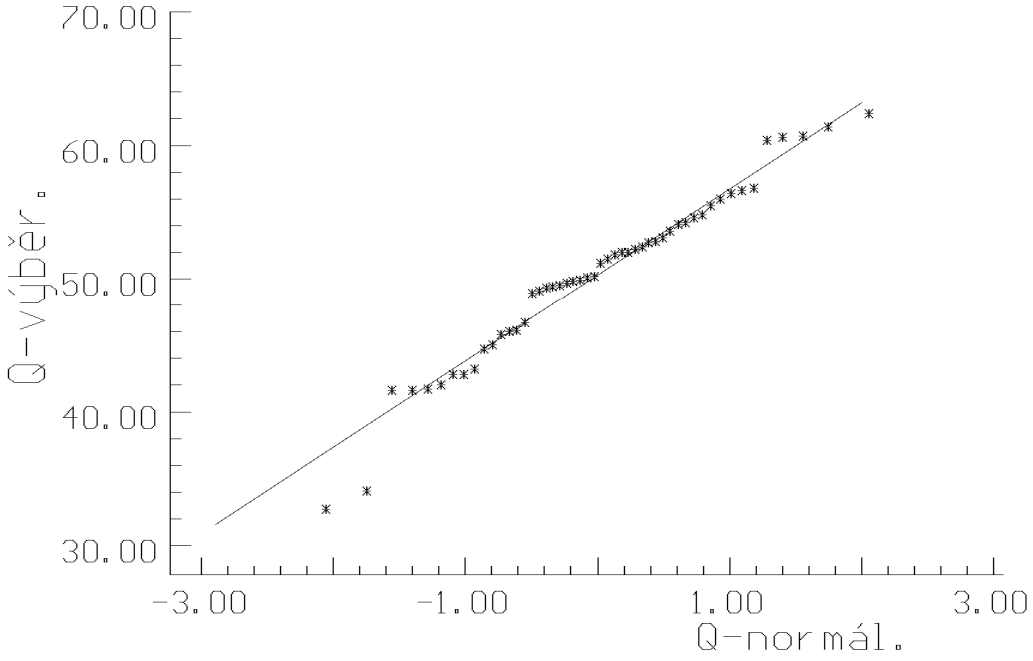
Stejné výstupy byly vytvořeny pro rovnoměrné rozdělení na obrázcích 8.10 , 8.11 , 8.12 a 8.13 . Pro rovnoměrné rozdělení je typické to, jak již název napovídá, že data jsou v podstatě stejnoměrně rozdělena v daném intervalu (je to také symetrické rozdělení, od normálního se liší tím, že v oblasti kolem střední hodnoty nedochází k vyšší koncentraci dat než na „okrajích“ rozdělení, jejich hustota je stále stejná).

Původní hodnoty				Pořádkové statistiky					
Rozdělení				Nornální		Exponenciální		Rovnoměrné	
Číslo prvku	normální	exponenciální	rovnoměrné	Číslo prvku	Hodnota	Číslo prvku	Hodnota	Číslo prvku	Hodnota
1	50.1	8.7	68.5	25	32.7	14	2.5	11	10.5
2	60.4	68.9	60.7	44	34.1	43	4.1	39	13.9
3	54.1	24.2	17.5	11	41.6	34	4.2	49	16.2
4	49.5	7.1	36.4	24	41.6	28	4.7	33	16.7
5	53.6	48.8	53.9	42	41.7	7	5.9	3	17.5
6	60.6	23.7	26.9	45	42.0	4	7.1	41	17.6
7	46.0	5.9	66.3	9	42.8	45	7.7	18	19.4
8	56.4	26.7	35.5	26	42.8	17	7.8	30	19.7
9	42.8	93.0	52.1	14	43.2	1	8.7	46	19.8
10	62.4	54.5	56.1	19	44.7	25	10.1	27	20.5
11	41.6	80.0	10.5	32	45.0	23	15.0	12	20.7
12	53.1	179.6	20.7	38	45.8	15	15.1	31	21.0
13	52.2	151.0	59.0	7	46.0	32	15.1	20	23.4
14	43.2	2.5	44.2	30	46.1	36	19.7	42	26.6
15	52.0	15.1	66.7	47	46.7	20	20.6	6	26.9
16	48.9	115.0	40.4	16	48.9	26	22.8	22	27.2
17	51.5	7.8	59.2	34	49.1	30	22.9	45	30.1
18	52.0	65.4	19.4	50	49.3	6	23.7	29	31.5
19	44.7	72.3	33.1	31	49.4	3	24.2	19	33.1
20	54.6	20.6	23.4	4	49.5	33	25.7	47	35.2
21	56.8	146.3	56.1	49	49.7	8	26.7	8	35.5
22	56.6	31.8	27.2	33	49.8	50	28.1	4	36.4
23	56.0	15.0	48.6	39	49.9	22	31.8	26	36.6
24	41.6	67.9	57.0	1	50.1	44	35.6	34	38.9
25	32.7	10.1	57.5	28	50.2	40	36.9	16	40.4
26	42.8	22.8	36.6	35	51.2	49	43.2	40	42.8
27	54.2	45.8	20.5	17	51.5	27	45.8	14	44.2
28	50.2	4.7	60.3	36	51.8	42	46.9	36	48.5
29	60.7	175.3	31.5	15	52.0	5	48.8	23	48.6
30	46.1	22.9	19.7	18	52.0	10	54.5	37	49.6
31	49.4	55.2	21.0	13	52.2	31	55.2	38	52.0
32	45.0	15.1	60.1	48	52.4	38	61.0	9	52.1
33	49.8	25.7	16.7	46	52.7	18	65.4	44	53.3
34	49.1	4.2	38.9	40	52.8	24	67.9	5	53.9
35	51.2	72.3	64.8	12	53.1	2	68.9	10	56.1
36	51.8	19.7	48.5	5	53.6	37	72.1	21	56.1
37	55.5	72.1	49.6	3	54.1	19	72.3	24	57.0
38	45.8	61.0	52.0	27	54.2	35	72.3	25	57.5
39	49.9	80.3	13.9	20	54.6	47	79.0	13	59.0
40	52.8	36.9	42.8	41	54.8	11	80.0	17	59.2
41	54.8	130.3	17.6	37	55.5	39	80.3	32	60.1
42	41.7	46.9	26.6	23	56.0	48	85.5	28	60.3
43	61.4	4.1	67.2	8	56.4	9	93.0	2	60.7
44	34.1	35.6	53.3	22	56.6	16	115.0	50	62.6
45	42.0	7.7	30.1	21	56.8	41	130.3	35	64.8
46	52.7	139.4	19.8	2	60.4	46	139.4	7	66.3
47	46.7	79.0	35.2	6	60.6	21	146.3	48	66.4
48	52.4	85.5	66.4	29	60.7	13	151.0	15	66.7
49	49.7	43.2	16.2	43	61.4	29	175.3	43	67.2
50	49.3	28.1	62.6	10	62.4	12	179.6	1	68.5

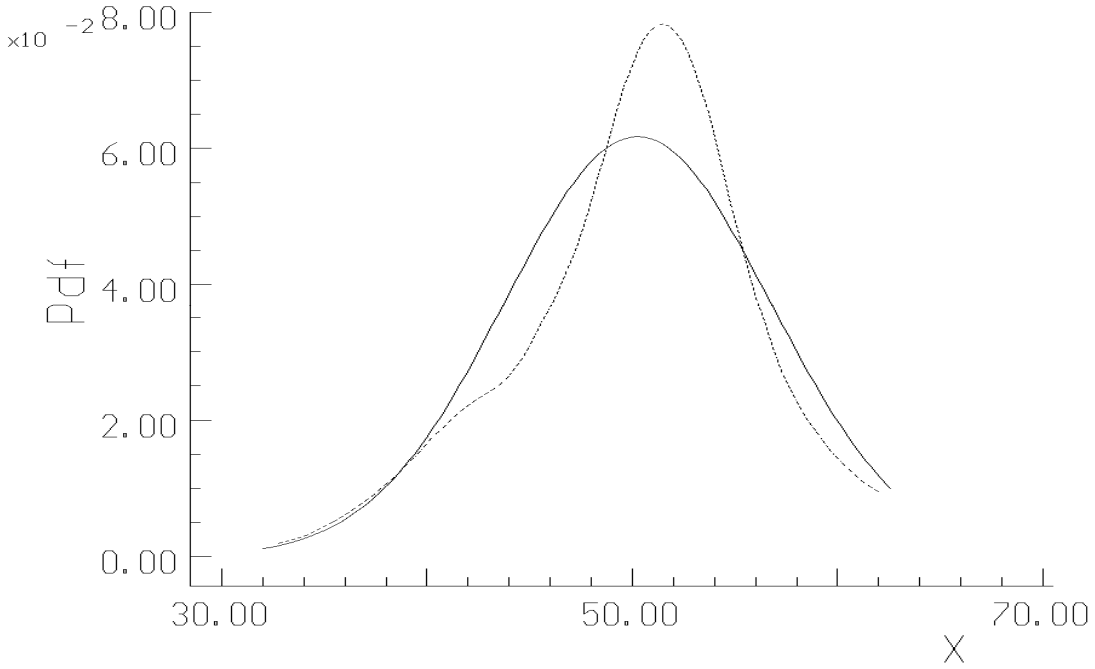
Tabulka 8.4- Generovaná rozdělení pro ilustraci použití grafických metod průzkumové analýzy dat



Obrázek 8.7 – Graf rozptýlení s kvantily pro normální rozdělení



Obrázek 8.8 – Kvantil-kvantilový graf pro normální rozdělení

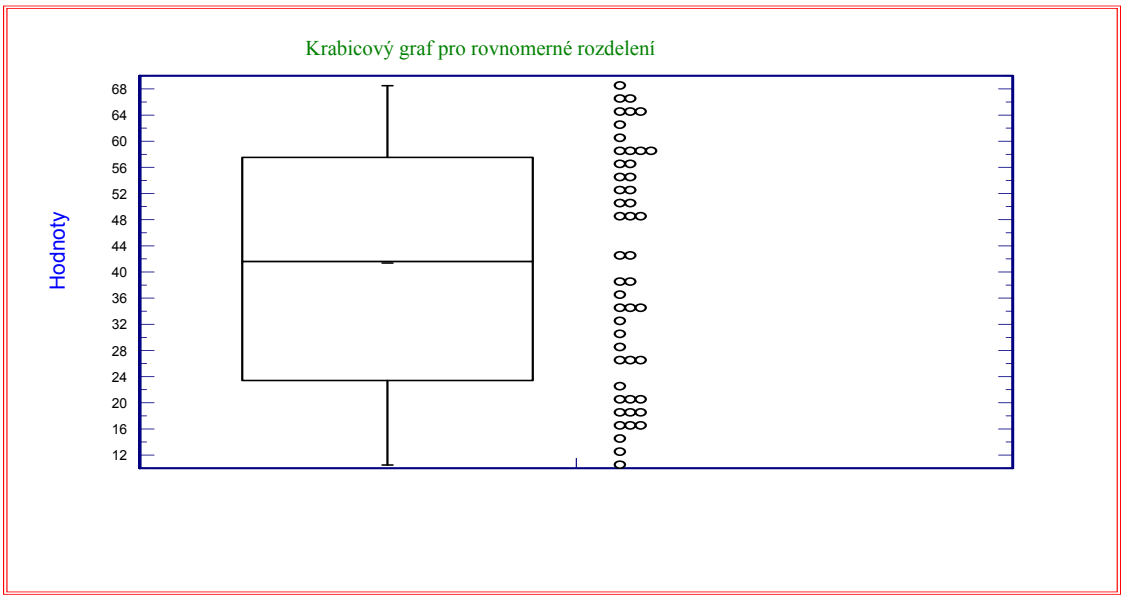


Obrázek 8.9 – Graf hustoty pravděpodobnosti pro normální rozdělení. Čárkovaná čára je jádrový odhad hustoty empirických hodnot, plná čára je frekvenční funkce normálního rozdělení.

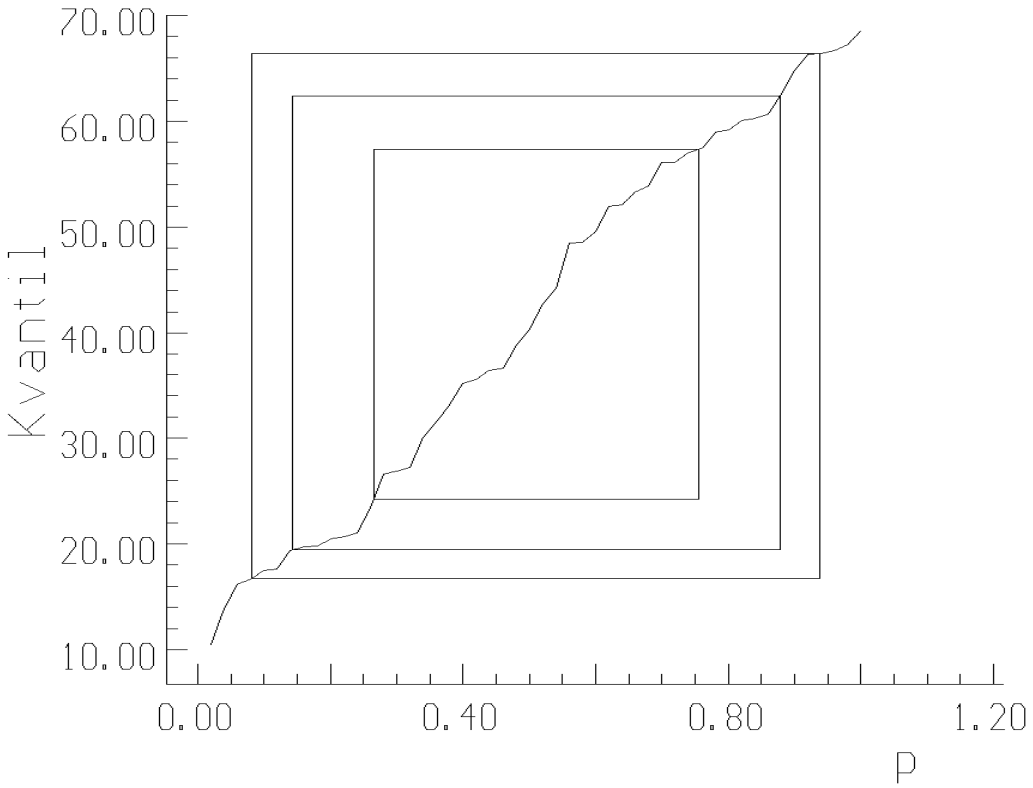
Tyto vlastnosti jsou potvrzeny také příslušnými grafy. Na grafu rozptýlení (tečkovém) a krabicovém vidíme, že „krabička“ je ve srovnání s normálním rozdělením poměrně dlouhá (to je právě indikace skutečnosti, že kolem střední hodnoty nedochází k větší koncentraci dat, to potvrzuje i tečkový graf vedle). Také aritmetický průměr se velmi dobře shoduje s mediánem (hodnoty prakticky splývají). Vzhledem ke značnému interkvartilovému rozpětí žádná hodnota neleží mimo vnitřní hradby souboru. Tento typ krabicového grafu je typický pro „plochá“ rozdělení, tj. pro taková, která nevykazují významnější koncentrace hodnot.

Výše uvedené typické vlastnosti rovnoměrného rozdělení se na grafu rozptýlení s kvantily projeví tím, že kvantilové obdélníky na jsou skoro čtvercového tvaru a spojnice empirických hodnot je téměř přímka (oproti esovitému tvaru u normálního rozdělení).

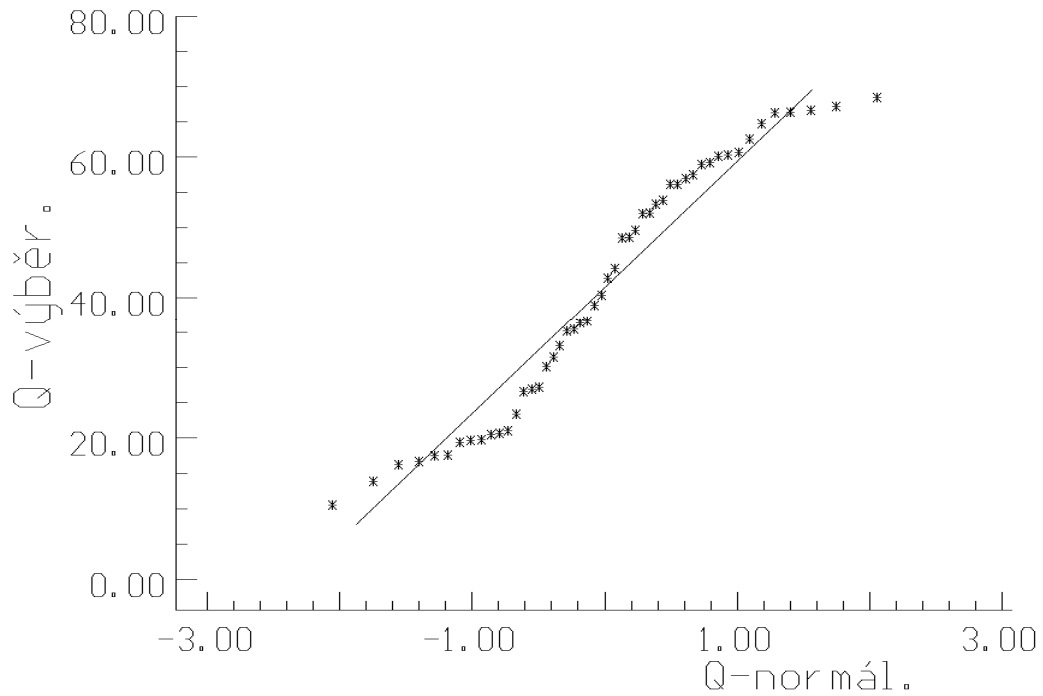
Kvantil-kvantilový graf a graf hustoty pravděpodobnosti také potvrzují typické vlastnosti rovnoměrného rozdělení – na Q-Q grafu (obrázek 8.12) je patrný typický tvar pro ploché rozdělení (viz schématická znázornění na obrázku 8.4). Také empirická křivka grafu hustoty pravděpodobnosti ukazuje na ploché a víceméně souměrné rozdělení (křivka je plošší – nižší – a širší, tj. má vyšší variabilitu, než křivka normálního rozdělení). Z obou obrázků je zřejmé, že rozdíly mezi rovnoměrným a normálním rozdělením nejsou velké a že modelování takového rozdělení pomocí obvyklého normálního rozdělení ve většině případů vyhoví.



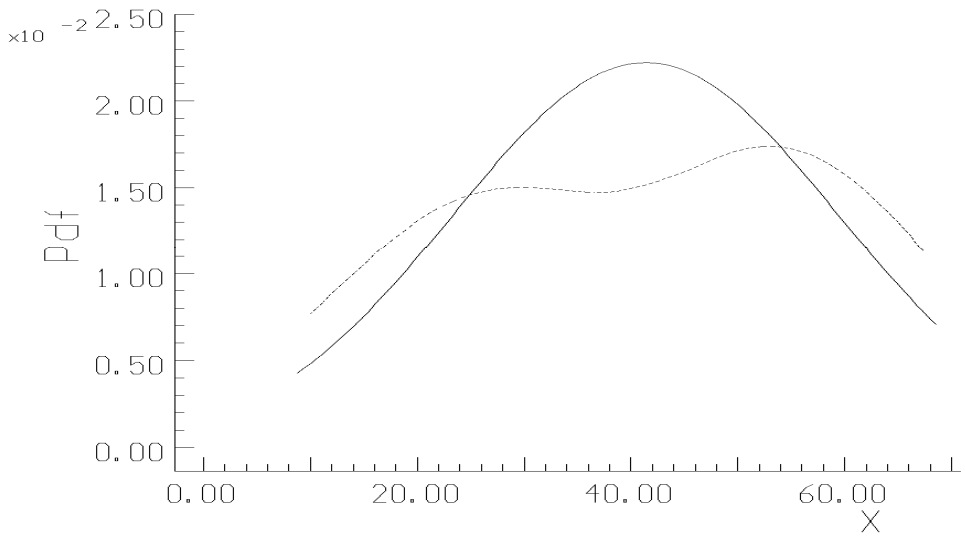
Obrázek 8.10 – Krabicový graf a diagram rozptýlení pro generované rovnoměrné rozdělení



Obrázek 8.11 - Graf rozptýlení s kvantily pro generované rovnoměrné rozdělení



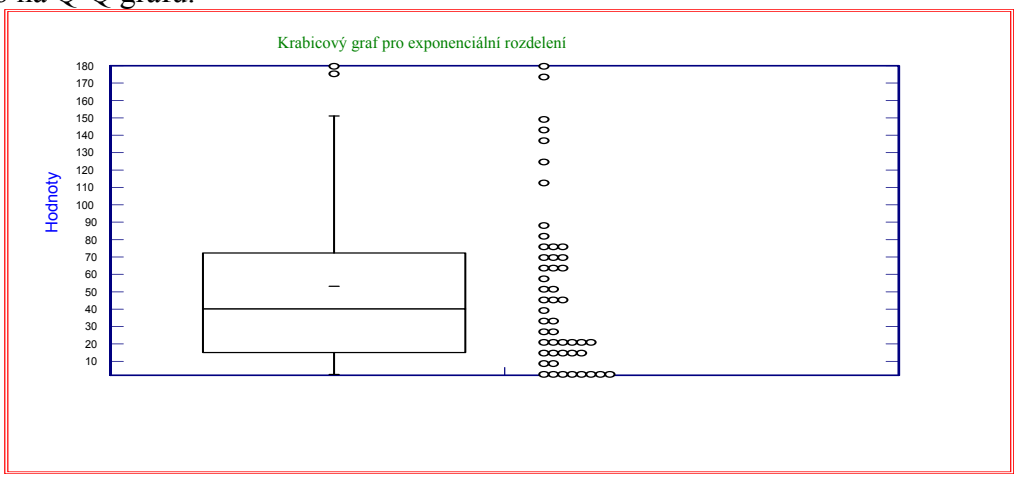
Obrázek 8.12 – Kvantil-kvantilový graf pro rovnoměrné rozdělení



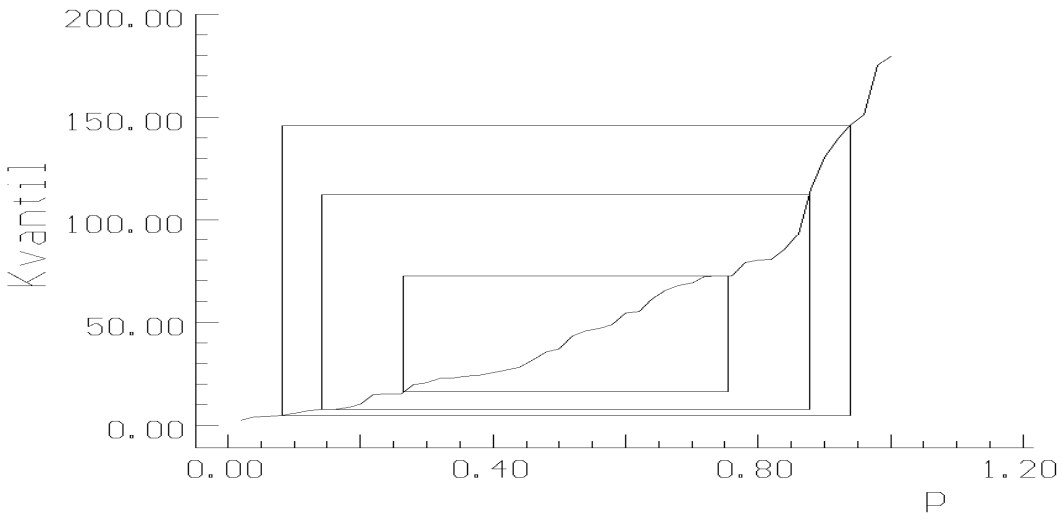
Obrázek 8.13 – Graf hustoty pravděpodobnosti pro rovnoměrné rozdělení

Posledním příkladem je exponenciální rozdělení. Jeho grafické interpretace jsou na obrázcích 8.14 , 8.15 , 8.16 a 8.17 . Je to typicky výrazně nesouměrné rozdělení,

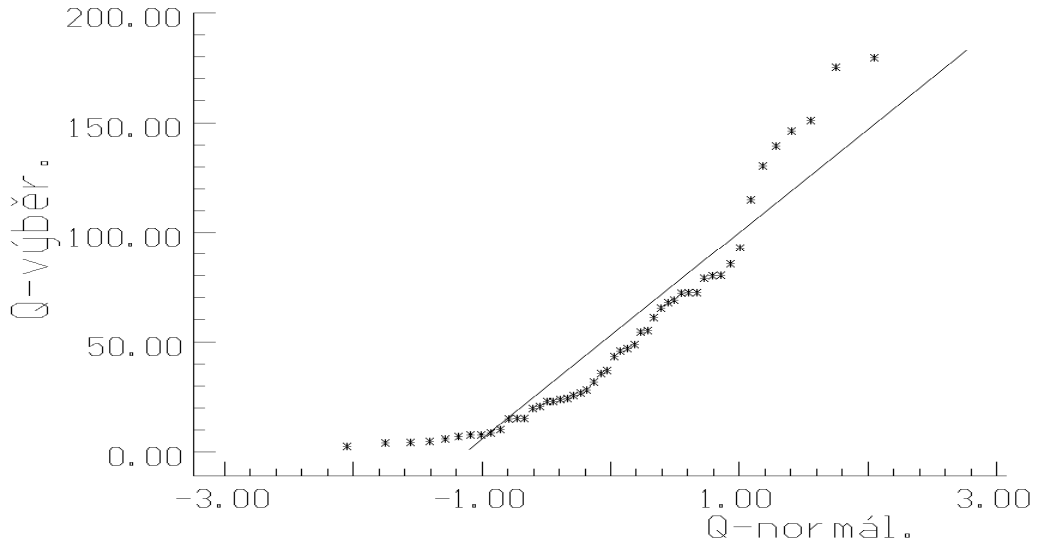
což je ihned názorně vidět z grafického zobrazení na obrázcích 8.14 a 8.15 . Na diagramu rozptýlení (tečkový graf) vidíme, že většina hodnot je koncentrována v dolní části (oblast nižších hodnot), jedná se tedy o výrazně levostranně nesouměrné rozdělení. O této skutečnosti také svědčí výrazný rozdíl mezi mediánem a aritmetickým průměrem (krátká čárka). Na horní straně (vyšší hodnoty) vidíme několik hodnot výrazně přesahujících vnitřní hradby souboru, přičemž by tyto hodnoty musely být v případě konkrétních měření velmi pozorně posuzovány z hlediska jejich správnosti a vypovídací schopnosti. Na grafu rozptýlení s kvantily je levostranné sešikmení vidět velmi názorně: vzdálenosti mezi dolními a horními stranami kvantilových obdélníků jsou značně odlišné - velká koncentrace nízkých hodnot způsobuje, že dolní strany jsou u sebe velmi blízko, což je typické právě pro levostrannou nesouměrnost. Také spojnice empirických hodnot vykazuje tvar typický pro levostranné rozdělení – stejný jako na Q-Q grafu.



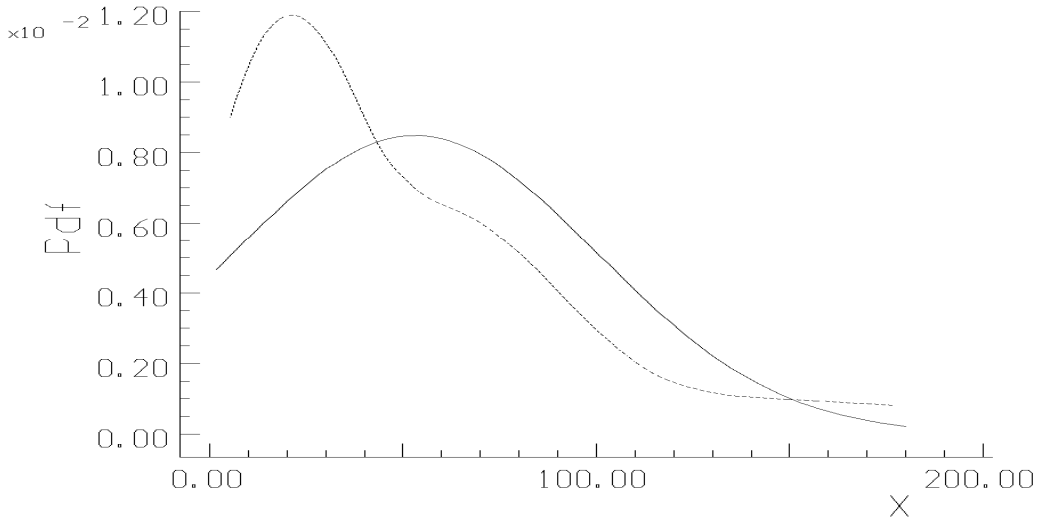
Obrázek 8.14 – Krabicový graf exponenciálního rozdělení



Obrázek 8.15 – Graf rozptýlení s kvantily pro exponenciální rozdělení



Obrázek 8.16 – Kvantil-kvantilový graf exponenciálního rozdělení



Obrázek 8.17 – Graf hustoty pravděpodobnosti exponenciálního rozdělení

Grafy shody s normálním rozdělením potvrzují výraznou odchylku od normálního rozdělení. Na kvantil-kvantilovém grafu snadno rozeznáme výrazné levostranné rozdělení (podle typického tvaru z obrázku 8.4 c). Stejný závěr potvrzuje obrázek 8.17 , kde můžeme potvrdit levostrannost a špičatost rozdělení.

Tabulka 8.5 uvádí pro srovnání základní statistické charakteristiky všech tří výběrů. Vidíme, že statistické charakteristiky dobře odpovídají předběžným závěrům, které jsme učinili na základě rozboru průzkumových grafů (normální rozdělení je mír-

ně pravostranné, rovnoměrné má vyšší variabilitu a je souměrné, exponenciální je silně levostranné s nejvyšší variabilitou danou odlehlými hodnotami). Je to potvrzení faktu, že z těchto relativně jednoduchých exploratorních grafů můžeme poměrně rychle a spolehlivě analyzovat základní vlastnosti posuzovaných výběrů.

Charakteristika (bodové odhady základního souboru)	Rozdělení		
	normální	rovnoměrné	exponenciální
aritmetický průměr	50.25	41.38	53.11
medián	50.70	41.60	40.05
rozptyl	41.77	322.48	2210.70
směrodatná odchylka	6.46	17.96	47.02
koeficient nesouměrnosti	- 0.45	- 0.08	1.12
koeficient špičatosti	3.29	1.59	3.49

Tabulka 8.5 – Statistické charakteristiky tří generovaných rozdělení (koeficient špičatosti pro normální rozdělení je roven 3, koeficient nesouměrnosti nule)

8.2 Ověření předpokladů o datech

Při použití obvyklých metod matematické statistiky (tedy pokud pracujeme s výběry) se zpravidla předpokládá, že se jedná o nezávislé náhodné veličiny pocházející z normálního rozdělení a že výběr má dostatečný rozsah pro spolehlivý odhad parametrů a testování hypotéz. Před provedením vlastní statistické analýzy bychom tedy měli ověřit následující vlastnosti:

- dostatečný rozsah výběru,
- nezávislost prvků výběru,
- normalitu výběru,
- homogenitu výběru.

8.2.1 Určení minimální velikosti výběru

Základní postupy týkající se potřebné velikosti výběru byly uvedeny v I. dílu, v kapitole 5.5.3 na str. 88.

8.2.2 Ověření normality výběru

Normalita výběrového rozdělení je jedním z nejdůležitějších předpokladů analýzy dat, je na něm založena většina obvykle používaných statistických metod, např.

metody korelační a regresní analýzy, mnohé testy apod. Pokud není normalita výběru prokázána, je nutno hlouběji analyzovat data a pokusit se zjistit příčiny. Data, u kterých se normalita neprokázala, je možné také analyzovat (zpravidla speciálními nebo modifikovanými metodami) nebo je možné data přiblížit normalitě pomocí tzv. **transformace**.

Grafické metody posouzení normality jsme probrali v předchozí kapitole (je to především kvantil-kvantilový, resp. rankitový graf a dále graf hustoty pravděpodobnosti). Kromě toho existuje ještě celá řada testů normality. Jeden z nich je uveden v 1. dílu na straně 115 (kapitola 7.4.1.5). Kromě něho se často používají např. Shapiro-Wilkův, D'Agostinův omnibus test, dále Anderson – Darlingův, Jarque – Berův, Kolmogorov- Smirnovův test a další.

Uvedeme ještě dva testy, které jsou často používány ve statistických programech, a to D'Agostinův omnibus test a Shapiro-Wilkův test.

.

D'Agostinův omnibus test (test kombinace výběrové šikmosti a špičatosti) (MELOUN - MILITKÝ 1994)

Pro reálné velikosti výběrů se používá testovací statistika

$$C = Z^2(g_1) + Z^2(g_2)$$

kde hodnoty $Z(g_1)$ a $Z(g_2)$ jsou normální aproximace výběrové šikmosti, resp. špičatosti. Pro výpočet $Z(g_1)$ potřebujeme vypočítat následující pomocné veličiny:

$$Y = g_1\sqrt{\frac{(n+1)(n+3)}{6(n-2)}}$$

$$G = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$

$$W = -1+\sqrt{2G-1}$$

$$A = \sqrt{\frac{2}{W^2-1}}$$

Z těchto pomocných veličin se určí aproximace

$$Z(g_1) = \frac{1}{\sqrt{\ln W}}\ln\left(\frac{Y}{A}+\sqrt{\left(\frac{Y}{A}\right)^2+1}\right)$$

Pro výpočet normální aproximace špičatosti vypočítáme veličinu S pomocí vztahu

$$S = \frac{g_2 - E(g_2)}{\sqrt{D(g_2)}}$$

kde je

g_2 vypočítaná výběrová špičatost

$E(g_2)$ střední hodnota výběrové špičatosti, která se pro normální rozdělení vypočítá podle vzorce

$$E(g_2) = 3 - \frac{6}{n+1}$$

$D(g_2)$ je rozptyl výběrové špičatosti vypočítaný podle vzorce

$$D(g_2) \approx \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}.$$

Dále se vypočítá šikmost veličiny S

$$g_1(S) = \frac{6(n^2-5n+2)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}$$

a pomocná hodnota

$$A = 6 + \frac{8}{g_1(S)}\left[\frac{2}{g_1(S)} + \sqrt{1+\frac{4}{g_1^2(S)}}\right],$$

Aproximace špičatosti se vypočítá

$$Z(g_2) = \frac{\left(1-\frac{2}{9A} - \sqrt[3]{\frac{1-\frac{2}{A}}{1+S\sqrt{\frac{2}{A-4}}}}\right)}{\sqrt{\frac{2}{9A}}}$$

Pokud zkoumaný výběr pochází z normálního rozdělení, potom statistika C má rozdělení χ^2 se dvěma stupni volnosti. Tento test je považován za velmi silný. Má výhodu v tom, že pomocí něho lze odděleně testovat samostatné hypotézy o vlivu šikmosti nebo špičatosti na normalitu, resp. nenormalitu výběru. Aproximace $Z(g_1)$ a $Z(g_2)$ tedy lze použít jako samostatné testovací statistiky. V těchto případech mají obě aproximace normované normální rozdělení N (0,1).

Vzhledem k relativně zdlouhavému výpočtu se doporučuje pro použití tohoto testu vypracovat jednoduchý program, který vypočítá hodnotu C i hodnoty obou aproximací. Pokud alespoň jedna z aproximací nevyhovuje normalitě, je celé rozdělení považováno za statisticky významně odlišné od normálního.

Shapiro – Wilkův test

Tento test byl odvozen pro menší výběry (doporučený rozsah výběru 3 – 50 prvků). Testové kritérium je

$$W = \frac{\left[\sum_{i=1}^{N} a_i x_{(i)}\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

kde koeficienty a_i jsou tabelovány ve speciálních tabulkách. Nulová hypotéza o normalitě se zamítá, pokud kritérium W je menší než tabelovaná kritická hodnota $W_{1;\alpha}$.

8.2.3 Ověření předpokladu nezávislosti prvků výběru

Základní test autokorelace I. řádu (von Neumanův test) je uveden v I. dílu na str. 116 (kapitola 7.1.4.6).

8.2.4 Ověření homogenity výběru

Problematika nehomogenních výběrů je velmi složitá, neboť jejích příčin může být mnoho (změna podmínek experimentu, nestejnoměrnost měřených vlastností apod.). Zde se omezíme na případ tzv. odlehlých (vybočujících) měření. Jsou to hodnoty, které se svou velikostí velmi výrazně liší od ostatních, jsou „podezřelá“, že nepatří do zkoumaného výběru.

Při komplexním statistickém ověřování „odlehlosti“ hodnot se používají komplikované procedury zahrnující sestavení modelu jejich chování, je nutno znát jejich rozdělení apod.

Existují ovšem relativně jednodušší metody, zpravidla založené pouze na předpokladu, že „správná“ data mají normální rozdělení. Mezi tyto metody patří **modifikace vnitřních hradeb.** S pojmem vnitřních hradeb souboru jsme se již setkali u krabicových grafů. Vypočítaly se jako dolní (resp. horní) kvartil $\pm$ 1.5-násobek interkvartilového rozpětí. Jejich modifikace spočívá v tom, že místo konstantní hodnoty 1.5 se používá parametr K, který je volen tak, aby pravděpodobnost P(n,K), že z výběru velikosti n pocházejícího z normálního rozdělení nebude žádný prvek mimo vnitřní hradby byla dostatečně vysoká (např. 0.95).Pro výběry v rozmezí $8 \le n \le 100$ se používá aproximace (MELOUN - MILITKÝ 1994)

$$K \approx 2.25 - \frac{3.6}{n}$$

Potom se horní (B_H^*) a dolní (B_D^*) modifikovaná hradba vypočítá

$$B_D^* = F_D - K\,R_F \qquad\qquad B_H = F_H + K\,R_F$$

Prvky, které leží mimo tyto modifikované hradby, považujeme za „podezřelé“ a podrobíme je další analýze.

Tyto hodnoty mohou totiž velmi výrazně ovlivnit především aritmetický průměr a rozptyl (a všechny na nich založené charakteristiky), a proto si vyžadují speciální pozornost. Zásadně nelze tyto hodnoty ihned z další analýzy vyloučit! Musíme velmi pozorně analyzovat příčiny, které vedly k výskytu takto odlehlých hodnot. Na jedné straně to mohou být opravdu příčiny opravňující vyloučení těchto hodnot z další analýzy, např. hrubá chyba měření, špatný zápis dat apod., ale na druhé straně musíme velmi pečlivě zvažovat „přirozené“ příčiny jejich výskytu. Jednou z možností je např. to, že měřená veličina může být charakterizována sešikmeným rozdělením, kde takováto - zdánlivě vybočující - hodnota může být přijatelná. Pokud opravdu zjistíme, že se jedná o vybočující hodnotu, potom můžeme použít k další analýze tzv. robustních metod, což jsou metody založené zpravidla na kvantilech, u nichž je vliv vybočujících hodnot výrazně oslaben. Vždy bychom měli mít na paměti, že vyloučení hodnoty z

analýzy je poslední a krajní možností a měli bychom ji užívat jen v případech, kdy jsme zcela přesvědčeni o nepřijatelnosti dané hodnoty.

Jiné často používané testy homogenity, např. Grubbsův nebo Dixonův – jsou uvedeny v I. dílu na stranách 114 a 132.

V následující části si ukážeme vyšetření základních předpokladů výběrů na praktickém příkladu.

Příklad 8.3:

Při výzkumu týkající se mechanických a fyzikálních vlastností dřeva byla kromě jiných údajů měřena hustota dřeva na zkušebních tělíscích. Dále byla na každém tělísku zjištěna průměrná šířka letokruhů (průměr z 10 letokruhů). Stanovte odhady základních parametrů obou veličin a pomocí průzkumové analýzy dat zjistěte, zda byly splněny základní předpoklady pro použití momentových odhadů. Měřená data jsou v tabulce 8.6

Na obrázcích 8.18 , 8.19 , 8.20 a 8.21 jsou postupně znázorněny krabicové grafy, kvantil-kvantilové grafy, grafy hustoty pravděpodobnosti a grafy rozptýlení s kvantily.

Jaké hodnocení výběru „hustota dřeva“ přináší průzkumové grafy?

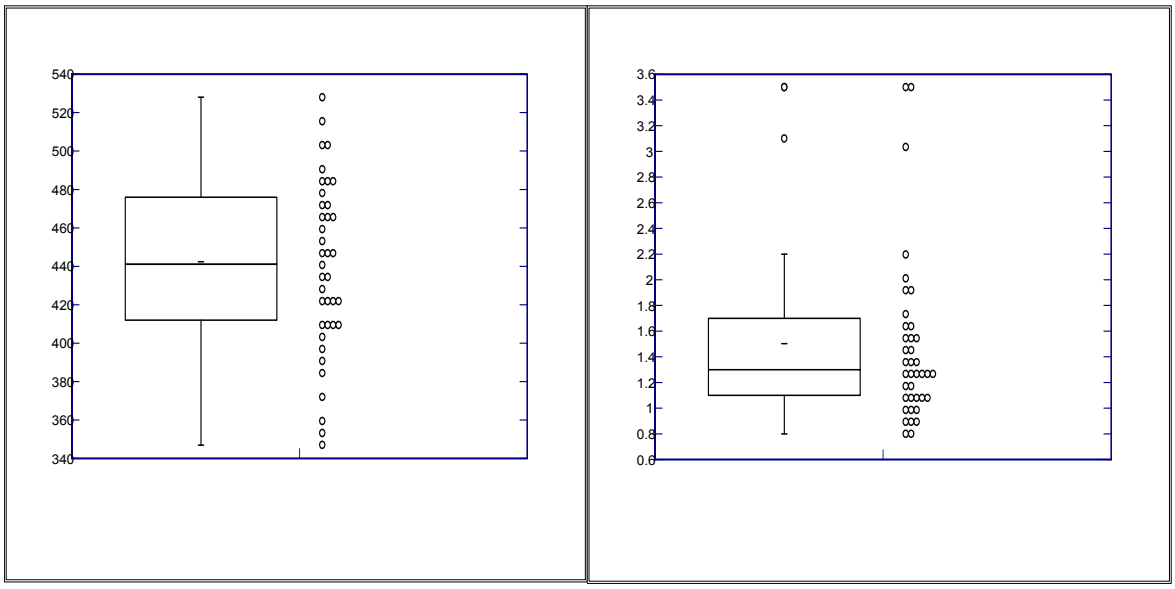
V případě krabicového grafu vidíme souměrnou „krabičku“, což svědčí o symetrii dat v okolí mediánu. Aritmetický průměr se téměř kryje s mediánem, graf neindikuje žádné odlehlé body. Velmi dobrou symetrii výběru hustota dřeva potvrzuje také graf rozptýlení s kvantily (obrázek 8.21) – všechny kvantilové obdélníky jsou vzájemně symetrické a nejsou indikovány žádné vybočující body. Kvantil-kvantilový graf i graf hustoty pravděpodobnosti potvrzují vynikající shodu s normálním rozdělením – v případě Q-Q grafu téměř všechny body leží na srovnávací přímce, v grafu hustoty pravděpodobnosti se jádrový odhad hustoty (čárkovaná čára) téměř kryje s křivkou normálního rozdělení (plná čára).

Zcela jinak vypadá situace u výběru „průměrná šířka letokruhů“. Všechny grafy signalizují výraznou nesouměrnost a přítomnost vybočujících bodů. Krabicový graf ukazuje silnou koncentraci hodnot mezi dolním kvartilem a mediánem (velmi úzká spodní část krabičky), značný rozdíl mezi hodnotou mediánu a aritmetického průměru a tři nejvyšší hodnoty jsou silně „podezřelé“. Koncentraci dat jemněji analyzuje graf rozptýlení s kvantily, kde je zřetelná nejvyšší míra koncentrace mezi sedecilem a oktilem a také mezi oktilem a dolním kvartilem. Také grafy popisující shodu s normálním rozdělením zcela jasně indikují silně levostranné rozdělení (typický „prohnutý“ tvar Q-Q grafu a velmi výmluvný tvar jádrového odhadu hustoty oproti normálnímu rozdělení v grafu hustoty pravděpodobnosti svědčící o levostranném a špičatém rozdělení).

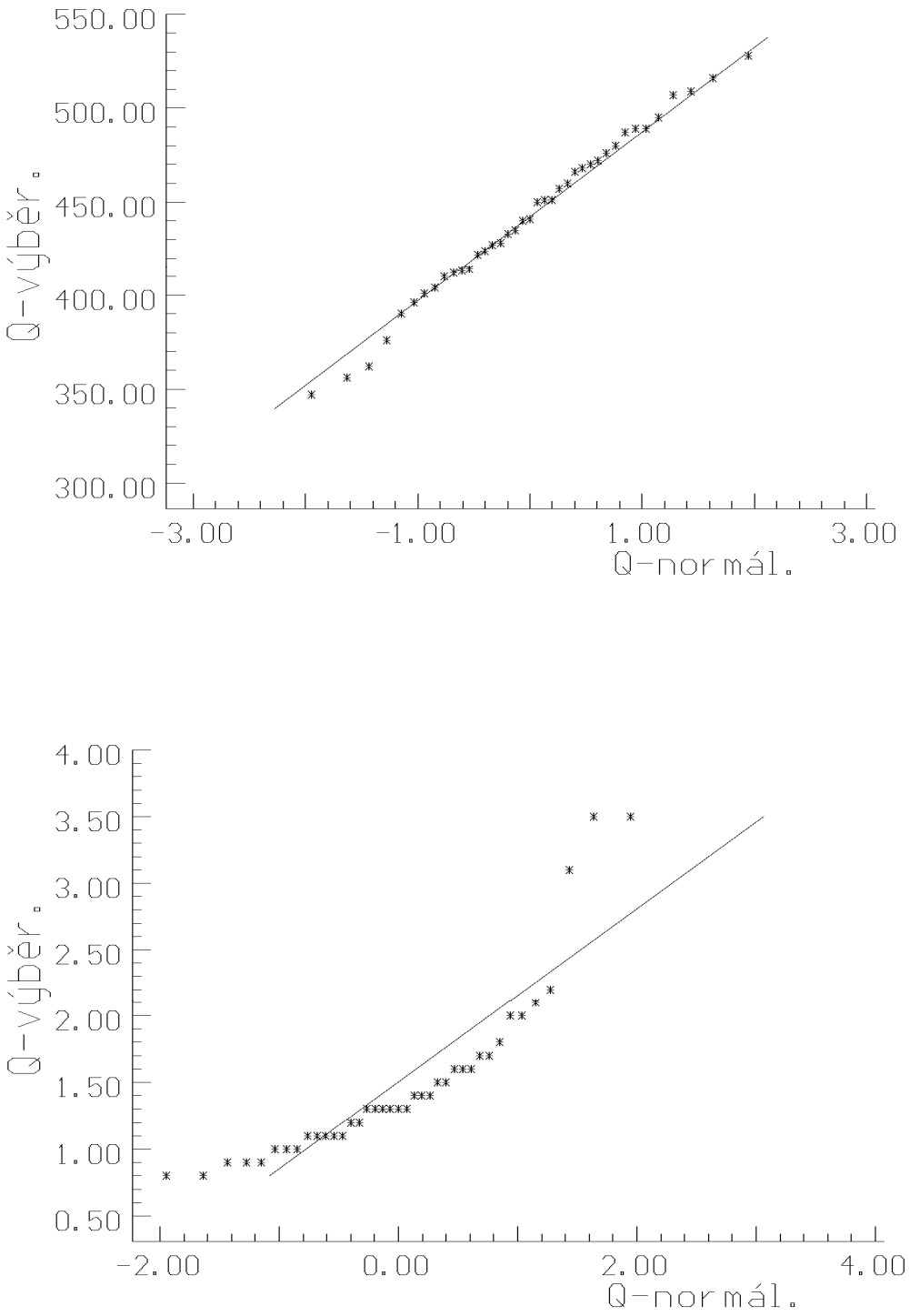
Interpretace všech grafů průzkumové analýzy dat je tedy velmi zřetelná. Výběr „hustota dřeva“ pochází zřejmě z normálního rozdělení s téměř ideální souměrností a vykazuje jen velmi mírné zploštění rozdělení, zatímco výběr „průměrná šířka letokruhů“ vykazuje jasné znaky silně sešikmeného levostranného rozdělení se třemi silně vybočující body.

Číslo vzorku	Hustota dřeva (kg/m^3)	Průměrná šířka letokruhu (mm)	Číslo vzorku	Hustota dřeva (kg/m^3)	Průměrná šířka letokruhu (mm)
1	516	1.3	21	413	1.0
2	528	1.5	22	362	1.3
3	396	2.0	23	489	1.0
4	487	1.1	24	466	1.1
5	356	3.5	25	472	1.0
6	507	0.8	26	404	3.1
7	390	1.4	27	401	1.5
8	427	1.3	28	450	0.9
9	347	1.2	29	433	0.9
10	376	2.1	30	470	1.1
11	457	1.4	31	451	1.1
12	451	0.9	32	412	1.6
13	509	1.7	33	428	1.3
14	435	3.5	34	468	1.3
15	424	1.2	35	422	1.7
16	410	1.1	36	414	1.6
17	480	1.6	37	476	2.0
18	441	1.3	38	440	1.4
19	460	0.8	39	489	1.8
20	495	2.2			

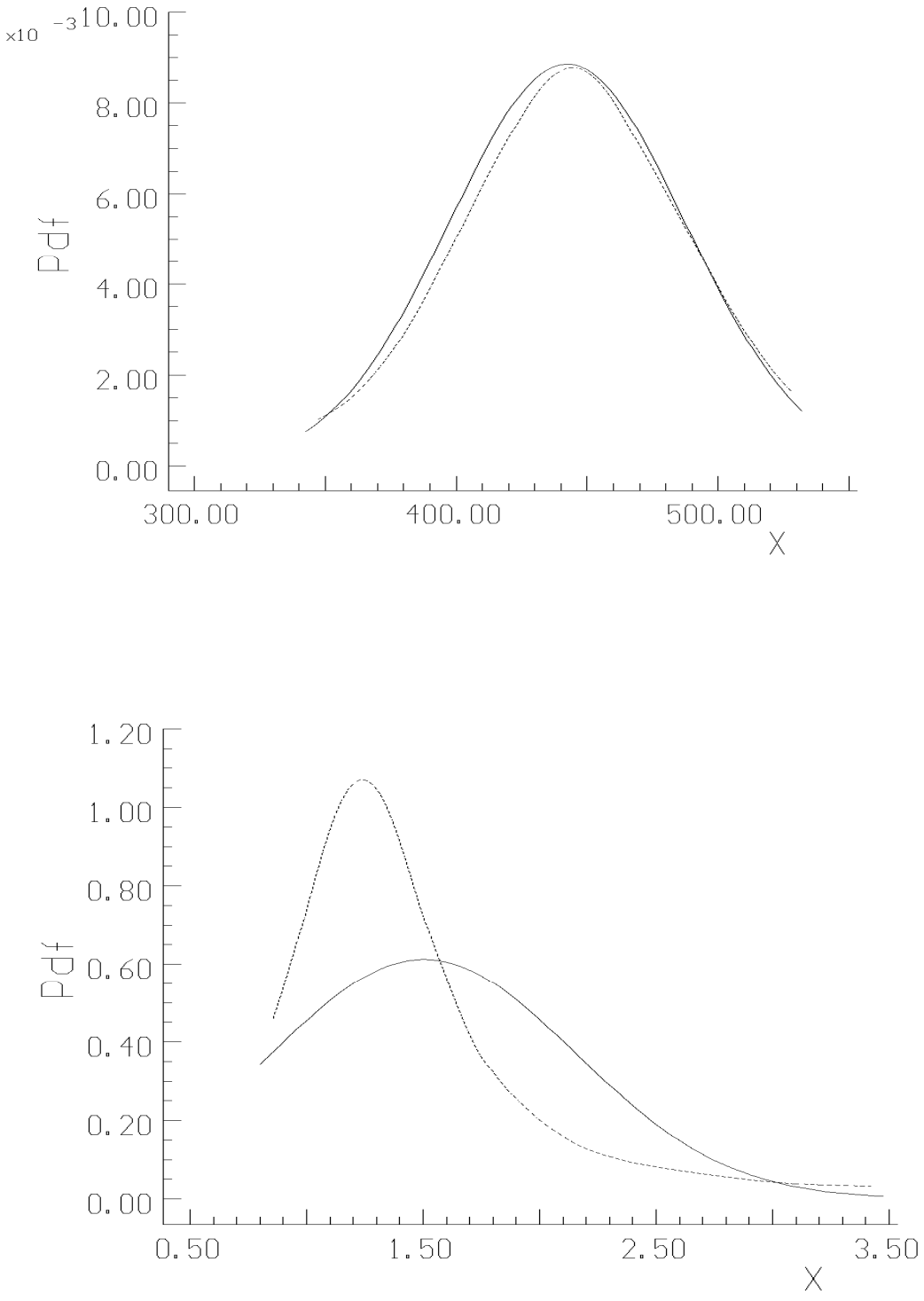
Tabulka 8.6 – Hodnoty výběrů „hustota dřeva“ a „průměrná šířka letokruhů“



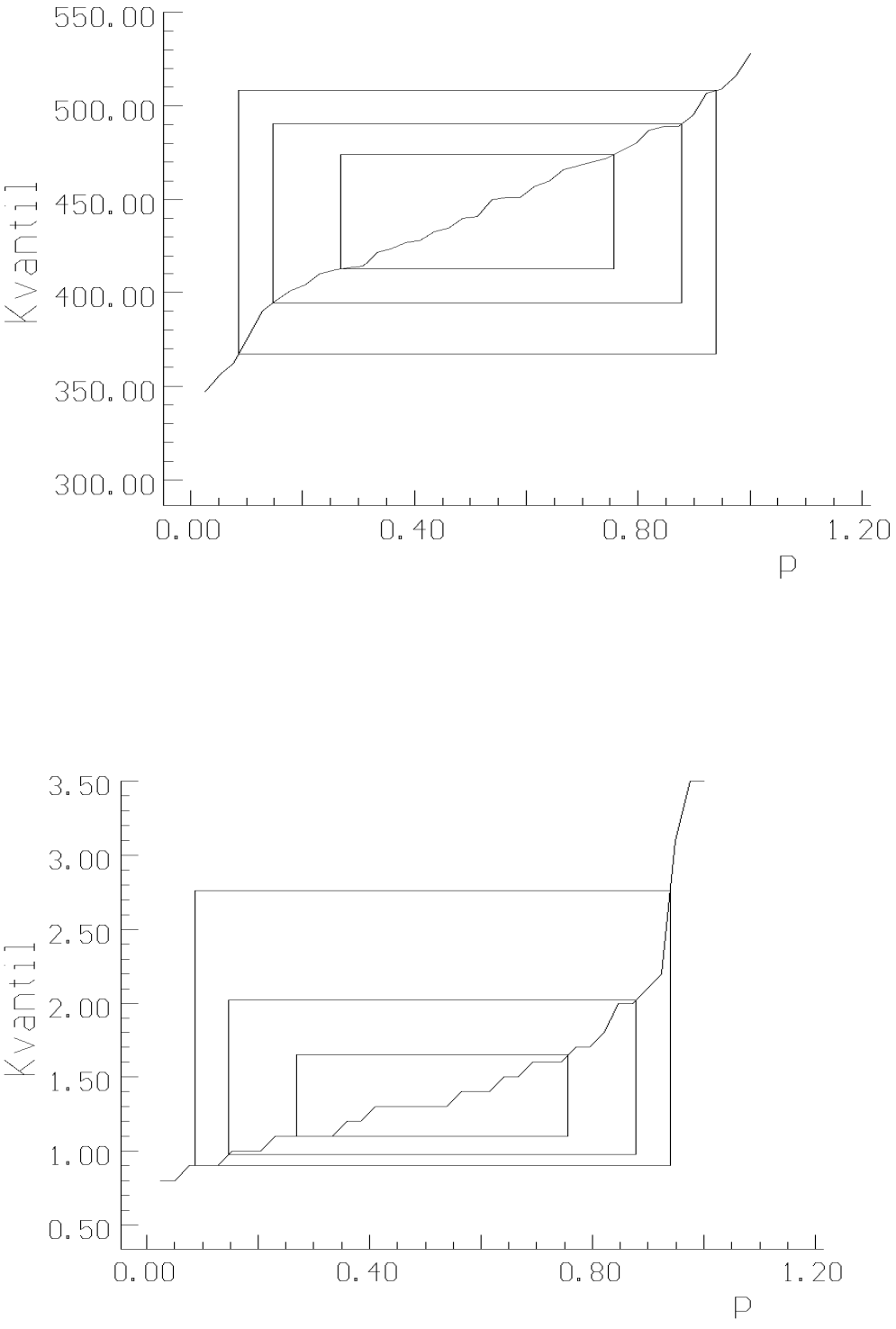
Obrázek 8.18 – Krabicové grafy pro výběry „hustota dřeva“ (vlevo) a „průměrná šířka letokruhu“ (vpravo)



Obrázek 8.19 – Kvantil – kvantilový graf pro výběr „hustota dřeva“ (nahoře) a „průměrná šířka letokruhu“ (dole)



Obrázek 8.20 – Grafy hustoty pravděpodobnosti pro výběry „hustota dřeva“ (nahoře) a „průměrná šířka letokruhu“ (dole)



Obrázek 8.21 – Grafy rozptýlení s kvantily pro výběry „hustota dřeva“ (nahoře) a „průměrná šířka letokruhu“ (dole)

Parametr	Výběr	
	hustota dřeva	průměrná šířka letokruhu
aritmetický průměr	442.36	1.50
medián	441.00	1.30
rozptyl	2028.80	0.43
směrodatná odchylka	45.04	0.65
šikmost	- 0.18	1.80
špičatost	2.41	5.98

Tabulka 8.7 – Odhady parametrů obou výběrů

Tyto předběžné závěry potvrzují odhady parametrů výběru, uvedené v tabulce 8.7 a také výsledky testů předpokladů výběru, které jsou uvedeny v tabulce 8.8 .

Je zřejmé, že hodnoty parametrů z tabulky 8.7 potvrzují, že výběr „průměrná šířka letokruhů“ je silně levostranný (hodnota koeficientu šikmosti 1,80) a špičatý (hodnota 5,98). O přítomnosti vychýlených hodnot svědčí také značná odchylka aritmetického průměru a mediánu (odchylku musíme porovnávat vzhledem k hodnotám, ze kterých je počítána – rozdíl mezi mediánem a aritmetickým průměrem v případě šířek letokruhů (absolutně 0,2 mm) činí asi 13 % z hodnoty průměru, zatímco rozdíl v případě hustoty dřeva (absolutně 1,36 kg/m^3) činí asi 0,3 % z hodnoty průměru). Naproti tomu hodnoty koeficientů šikmosti a špičatosti u výběru „hustota dřeva“ potvrzují, že výběr pochází z rozdělení, které je velmi blízké normálnímu.

Test	Výběr	Testové kritérium	Kritická hodnota	Výsledek testu
D'Agostinův test normality	hustota	0.701	5.9915	normalita nezamítnuta
	letokruhy	47.599	5.9915	normalita zamítnuta
von Neumannův test nezávislosti	hustota	0.240	2.0211	nezávislost prokázána
	letokruhy	1.128	2.0211	nezávislost prokázána
Homogenita (modifikované vnitřní hradby)		Hranice		Odlehlé hodnoty
		dolní	horní	
	hustota	279.80	606.70	žádné
	letokruhy	- 0.09	2.84	hodnoty 5(3,5); 14(3,5); 26(3,1)

Tabulka 8.8 – Výsledků testů předpokladů výběru

Také tabulka 8.8 potvrzuje výsledky průzkumové analýzy, kdy normalita byla zamítnuta pro výběr průměrná šířka letokruhu a byly zde potvrzeny tři odlehlé body.

Je nutné si uvědomit, že výše popisovaný příklad byl volen úmyslně tak, aby jednotlivé postupy průzkumové analýzy dat byly jasně vidět a že v mnoha případech je rozhodování o vlastnostech výběrů a jejich příčinách daleko obtížnější a vyžaduje značné znalosti a zkušenosti. Také výběr metod v této kapitole byl zúžen jen na ty nejjednodušší a nejpoužívanější. Další příklady a rozbor dalších používaných metod je možné nalézt např. v publikacích MELOUN - MILITKÝ 1994, TUKEY 1977, CHAMBERS ET ALL. 1983 a dalších.

8.3 Transformace dat

Pokud průzkumová analýza dat odhalí, že rozdělení výběru dat se příliš liší od normálního rozdělení, nastává problém s volbou statisticky správného způsobu vyhodnocení dat a získání co nejvěrohodnějších odhadů parametrů. Nenormalita dat totiž znemožňuje použít např. aritmetický průměr jako odhad střední hodnoty (a samozřejmě všechny charakteristiky na něj výpočetně vázané), použít obvyklé postupy pro intervalový odhad, pro stanovení důležitých kvantilů, apod. Odmítnutí normality je většinou způsobeno asymetrií dat, proto většina způsobů odstranění nenormality se snaží asymetrii odstranit. O některých způsobech transformace jsme se zmínili v I. dílu, v kapitole 5.4.2.2.3 na straně 78.

Jedou z nejúčinnějších metod pro odstranění asymetrie dat je **nelineární transformace**. Její princip si popíšeme podle obrázku 8.22 :

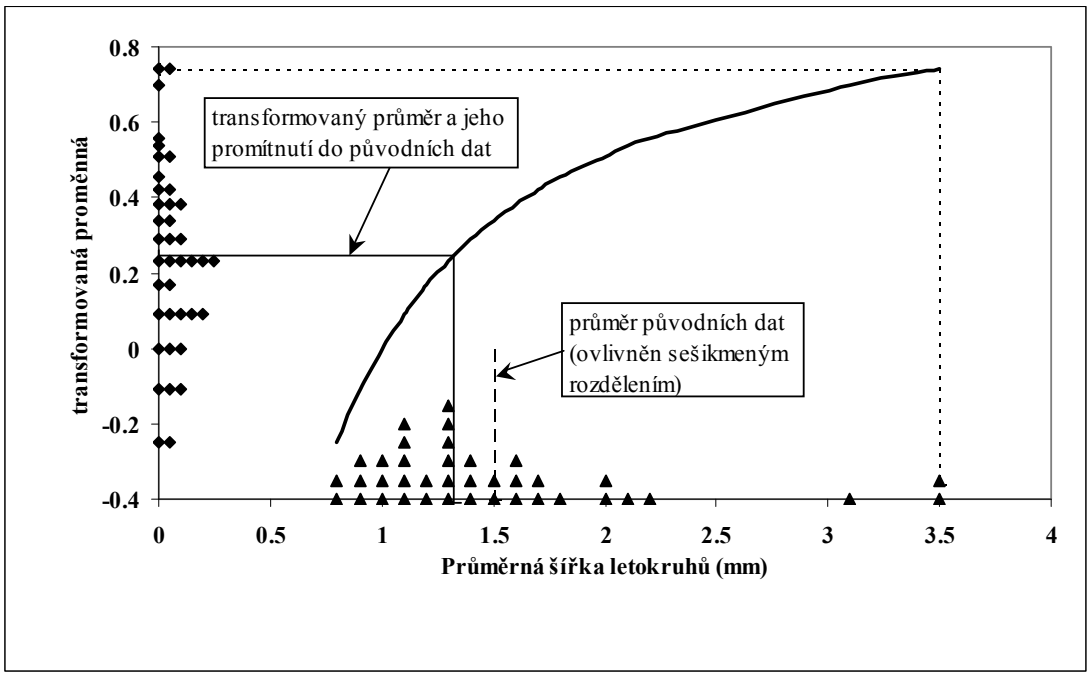
- Máme výběr, který se vyznačuje silnou asymetrií (data vyznačena černými trojúhelníčky – jsou to hodnoty výběru „průměrná šířka letokruhu“ z předchozího příkladu). Data se vyznačují jednak silnou koncentrací mezi hodnotami 1 a 1,5, jednak odlehlými hodnotami. Musíme nalézt vhodný tvar transformační funkce (na obrázku 8.22 vyznačena tučnou čarou);
- pomocí vhodné funkce transformujeme původní data tak, aby nová data (na obrázku 8.22 jsou jejich hodnoty vyznačeny černými kosočtverci) byla symetrická (je vidět, že transformace odstranila hlavní odlehlé hodnoty a že „nová data“ vykazují podstatně vyšší míru symetrie než původní – transformace pro nejvychýlenější původní hodnoty - 3,5 - je vyznačena pomocí krátce čárkované čáry);
- v souboru „nových dat“ již můžeme vypočítat aritmetický průměr běžným způsobem (tato data jsou normální), stejně jako interval spolehlivosti, apod.;
- odhady parametrů vypočítané pro transformované hodnoty promítneme (retransformujeme) do původních souřadnic pomocí inverzní funkce. Tím získáme spolehlivější odhady parametrů a intervaly spolehlivosti než z původních dat.

Hlavním problémem je najít vhodnou funkci, která by měla splnit tato kritéria:

- musí být nelineární (lineární funkce by pouze změnila měřítko a posunula data);
- musí být monotónní (aby zůstalo zachováno pořadí dat – tj. vyšší původní hodnoty budou i vyšší transformované);
- měla by zajistit maximální symetrii nebo (lépe) maximální normalitu dat.

Velmi vhodnou funkcí je Box-Coxova transformace, což je funkce patřící mezi mocninné transformační funkce. Její tvar je následující:

$$\Psi(x) = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases}$$



Obrázek 8.22 – Princip nelineární transformace dat

Tato transformace účinně přibližuje výběr normalitě jak z hlediska šikmosti, tak i z hlediska extrémních hodnot. Určení hodnoty λ, samotná transformace a především následná retransformace parametrů jsou teoreticky i výpočetně velmi náročné postupy a pro jejich realizaci je třeba výkonný statistický program (z těch dostupnějších tuto transformaci provádí např. ADSTAT). Teorii Box – Coxovy transformace viz např. MELOUN - MILITKÝ 1994 nebo KUPKA 1997.

Její účinnost si prokážeme na datech z předchozího příkladu (na výběru „průměrná šířka letokruhů“).

Příklad 8.4:

Pomocí Box-Coxovy transformace stanovte odhady parametrů výběru „průměrná šířka letokruhů“, jehož data jsou uvedena v tabulce 8.6 .

Praktické provedení výpočetně náročné nelineární transformace je možné jen s použitím specializovaného statistického programu.

Nejdříve je nutné posoudit oprávněnost transformace (tedy zda transformace bude mít „statistický přínos“, tj. podstatně zlepší odhady parametrů).

Pomocí programu ADSTAT byla nalezena optimální hodnota $\lambda = -0{,}93$ (tato hodnota zajišťuje maximální symetrii i normalitu transformovaných dat, která je charakterizována koeficientem špičatosti = 2,58 a koeficientem šikmosti –0,0085, což je velmi podstatná změna oproti hodnotám těchto koeficientů pro netransformovaná data, které jsou uvedeny v tabulce 8.7).

Pomocí hodnoty λ byly vypočítány transformované hodnoty (x') s normálním rozdělením, pro které již byl vypočítán klasický průměr ($\bar{x}' = 0{,}247$) a hodnoty rozptylu ($S'^2 = 0{,}0625$) a směrodatné odchylky (S' = 0,25).

Zlepšení normality je možné posoudit také na Q-Q grafech původních a transformovaných dat (viz obrázek 8.24). Je zřetelné, že transformace data „znormalizovala", neboť téměř všechny body leží perfektně na přímce, zatímco horní Q-Q graf (před transformací) vykazuje průběh typický pro levostranné rozdělení.

Poté je nutné posoudit oprávněnost transformace (tedy zda transformace bude mít „statistický přínos", tj. podstatně zlepší odhady parametrů) pro stanovenou hodnotu parametru λ. To se provádí pomocí grafu logaritmu věrohodnostní funkce (viz obrázek 8.23). V tomto grafu se na ose X vynáší hodnoty λ a na ose Y hodnoty logaritmu věrohodnostní funkce stanovené podle vztahu

$$\ln L(\lambda) = -\frac{n}{2}\ln s^2(x') + (\lambda - 1)\sum_{i=1}^{n}\ln x_i$$

kde je $s^2(x')$ výběrový rozptyl transformovaných dat.

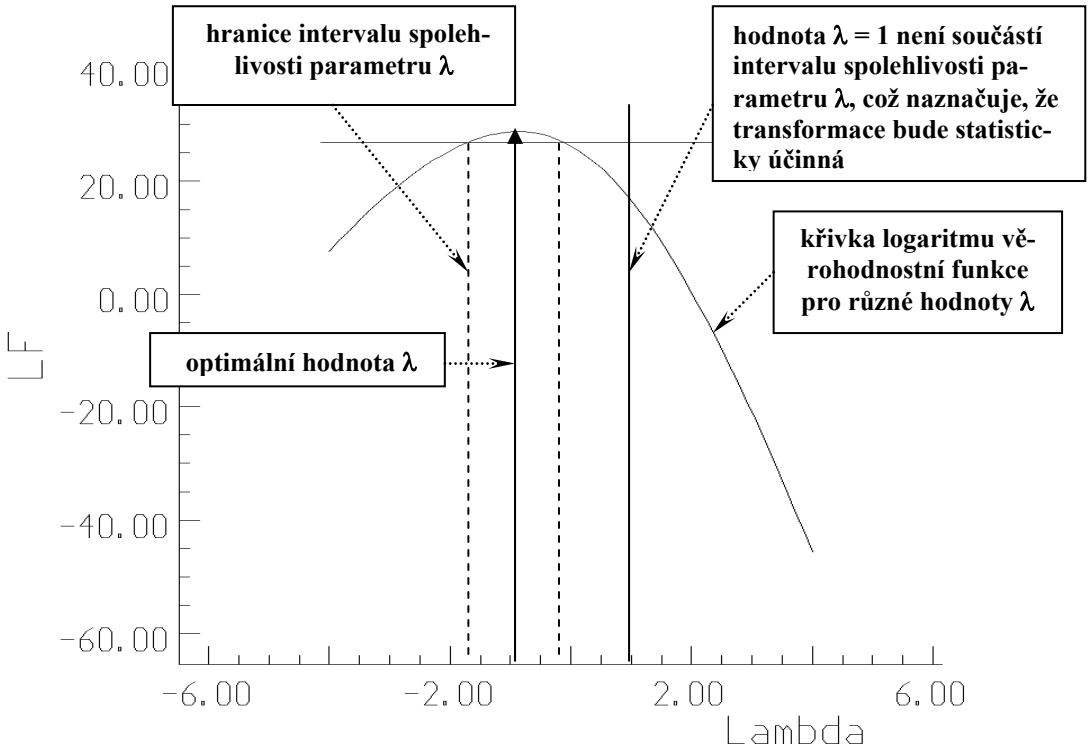
V grafu je také nakreslen interval spolehlivosti optimální hodnoty λ. Pokud tento interval obsahuje hodnotu λ = 1, potom Box-Coxova transformace není ze statistického hlediska přínosem. Je nutno upozornit, že v některých případech tento graf nedává výsledky, které by byly jednoznačně interpretovatelné (transformace ano – ne), v tom případě je nutno se řídit porovnáním původních a transformovaných odhadů parametrů a provést rozbor statistického přínosu transformace.

Transformované hodnoty statistických charakteristik byly pomocí Taylorova rozvoje v okolí transformovaného průměru promítnuty (retransformovány) do původního měřítka, čímž byly získány hodnoty uvedené v pravém sloupci tabulky 8.9 .
Je zřejmé, že transformací došlo k výraznému posunu střední hodnoty, která již není tolik zatížena odlehlými měřeními. Významná změna nastala také u konfidenčního intervalu, který je ve shodě s nesouměrným rozdělením původních dat také nesouměrný (dolní část má rozsah 1,32-1,20 = 0,12, zatímco horní 1,48-1,32 = 0,16), což lépe odpovídá realitě (nesouměrnému rozdělení s delším „horním" koncem), než souměrný interval vypočítaný z původních hodnot. Z retransformovaného rozdělení je také možné vypočítat jakékoli kvantily, jež budou také nesymetrické. Dalším přínosem transformace bylo snížení variability (menší hodnoty rozptylu a směrodatné odchylky).

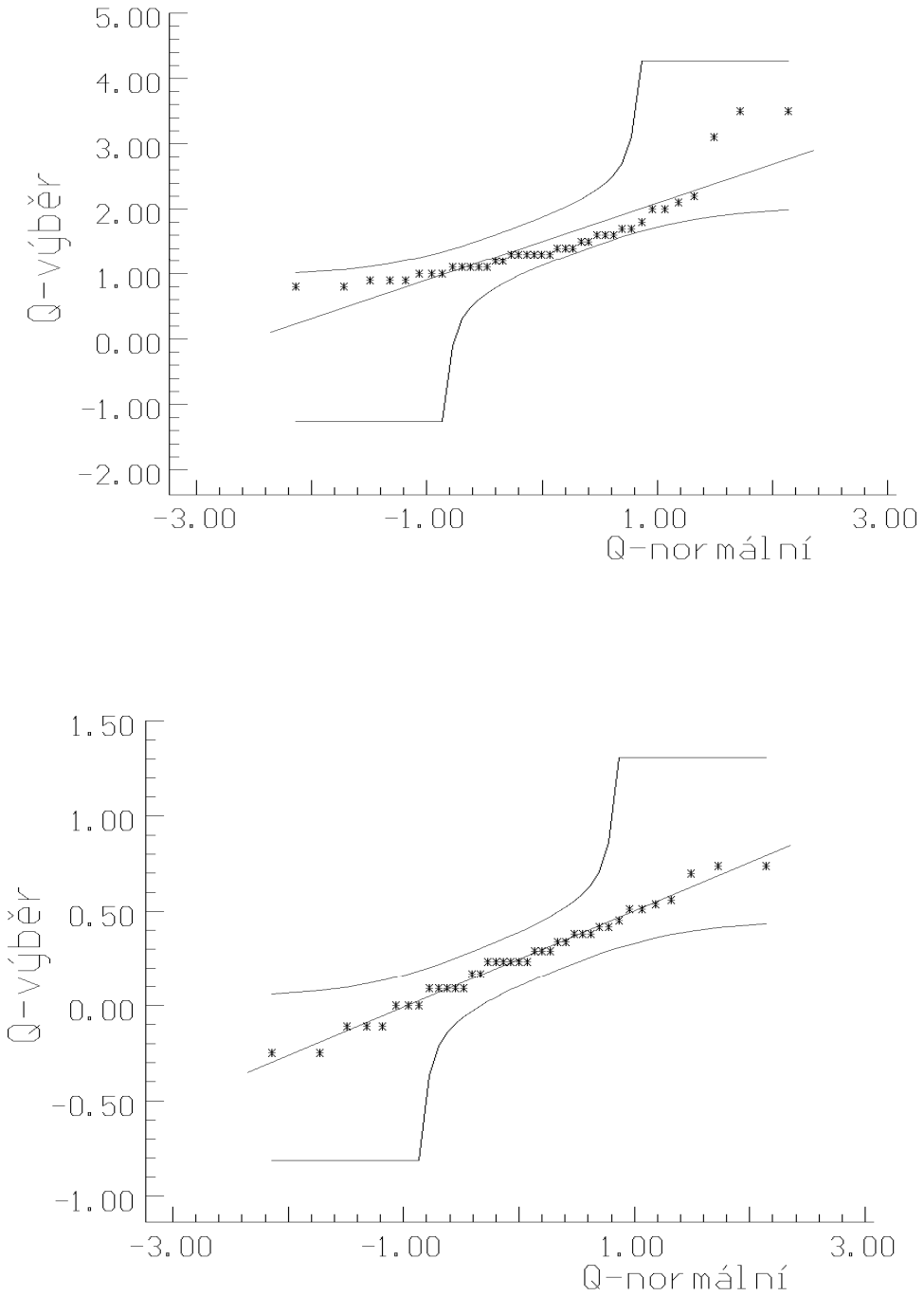
Závěrem je nutno zdůraznit, že ačkoli je Box-Coxova transformace jedna z nejlepších a neúčinnějších, v některých případech její postup selže (zvláště, jestliže je optimální hodnota parametru λ určována automaticky programem) a získané hodnoty nejsou použitelné (např. nepřijatelně malé nebo naopak velké hodnoty charakteristik variability, nevěrohodné hranice konfidenčních intervalů apod.). V těchto případech musíme přistoupit buď k „ručnímu" hledání optimální hodnoty λ, což je postup velmi náročný na znalosti analytika nebo zkusit použít jinou transformaci, která vede k normálnímu rozdělení. Další možností je použít pro původní data kvantilových (robustních) odhadů parametrů.

Parametr	před transformací (původní parametry)	po retransformaci („opravený parametr“)
aritmetický průměr	1,50	1,32
dolní hranice konfidenčního intervalu průměru	1,29	1,20
horní hranice konfidenčního intervalu průměru	1,71	1,48
rozptyl	0,43	0,18
směrodatná odchylka	0,65	0,43

Tabulka 8.9 – Odhady parametrů pro původní a retransformované hodnoty



Obrázek 8.23 – Graf logaritmu věrohodnostní funkce pro posouzení oprávněnosti Box-Coxovy transformace



Obrázek 8.24 – Porovnání Q-Q grafů původních dat (nahoře) a dat po transformaci (dole)

9 Analýza rozptylu (ANOVA)

S pojmem „analýza rozptylu“ (zkratkou se označuje jako ANOVA – z anglického názvu **AN**alysis **O**f **VA**riance – což je mezinárodně užívané a srozumitelné označení) jsme se již setkali v I. dílu tohoto textu – v kapitole 7.4.3.2 na straně 125, tedy v kapitole o statistických testech pro více výběrů. Zde byla pouze stručná zmínka o existenci této metody, s tím, že podrobněji bude rozebrána v samostatné kapitole. Důvodem je hlavně to, že se jedná, ve srovnání s ostatními testy, o relativně složitou a rozsáhlou metodiku s mnoha variantami.

ANOVA je vlastně statistický test, který testuje nulovou hypotézu o shodě středních hodnot pro více výběrů. Pojmem více výběrů rozumíme 3 a více (testy pro 1 a 2 výběry viz I. díl v kapitolách 7.4.1, 7.4.2 a 7.5, tedy „klasické“ F-testy, t-testy, event. neparametrické testy). Na tomto místě je nutné připomenout fakt, který byl již zdůvodněn v kapitole 7.4.3, že je **nepřípustné používat pro simultánní hypotézu (tj. pro více než 2 výběry) o rovnosti průměrů opakované t-testy** (souvisí to se zvyšováním hodnoty chyby I. druhu nad nastavenou mez). Naopak to ovšem možné je – tedy analýzu rozptylu můžeme použít pro srovnání dvou výběrů, dosažená hladina významnosti bude shodná s t-testem.

Uveďme si několik příkladů, kdy je vhodné tuto metodu použít.

1) Při ověřování účinnosti nového typu hnojiva, o kterém se předpokládá, že bude vhodné do lesních školek, je nutné stanovit vliv různých dávek hnojiva na růst semenáčků.

2) Je potřeba prokázat vliv různých druhů hnojiv na růst.

3) Byl vyvinut nový počítačový sortimentační program, který na základě změřených biometrických veličin porostu je schopen vypočítat výtěžnost jednotlivých sortimentů. Je nutné porovnat jeho výsledky s jinými, dosud používanými metodami (např. metodou kvalifikovaného odhadu a stávajícími sortimentačními tabulkami).

4) Byly odebrány vzorky dřeva určité dřeviny v různých lokalitách. Úkolem analýzy je vyšetřit, zdali se mechanické a fyzikální vlastnosti dřeva liší podle lokalit.

Všechny uvedené příklady spojuje společná myšlenka – postihnout vliv jednotlivých úrovní určitého faktoru (např. druhu nebo dávky hnojiva, různých metod, různé lokality) na nějakou měřenou veličinu (např. výšku semenáčků, hustotu dřeva, zpeněžení sortimentů, apod.). Jak je možné tuto velmi častou úlohu vyřešit?

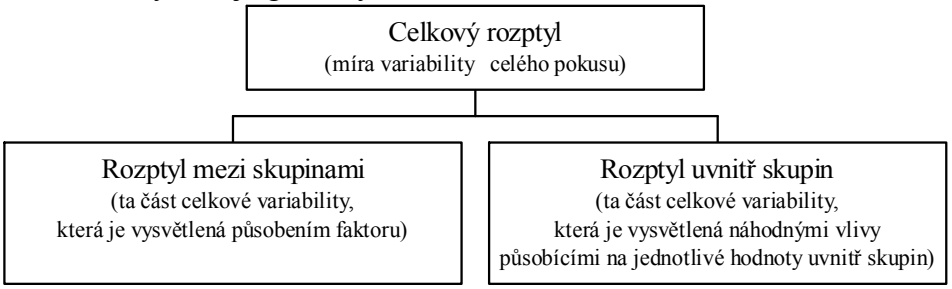
Vycházíme z následující úvahy: pokud by zkoumané faktory neměly na příslušnou měřenou veličinu žádný vliv, potom se jejich působení neprojeví na statistických charakteristikách této veličiny. Pokud by např. různé dávky hnojiva neměly vliv na růst semenáčků, podle měřitelných faktorů (např. výšky, tloušťky kořenového krčku, apod.) nijak nepoznáme, na které semenáčky bylo hnojivo aplikováno. Naopak, pokud bude vliv daného faktoru (resp. určité jeho úrovně, např. určité dávky hnojiva) významný, potom se to zřejmě projeví na příslušných statistických charakteristikách měřené veličiny, především na míře variability – rozptylu - a hlavní míře polohy – aritmetickém průměru. **Odlišnost rozptylů a aritmetických průměrů jednotlivých po-**

rovnávaných výběrů se tedy považuje za míru intenzity působení posuzovaných faktorů (jejich úrovní).

Princip analýzy rozptylu můžeme vysvětlit s určitým zjednodušením vysvětlit takto:

- testujeme nulovou hypotézu, že střední hodnoty jednotlivých výběrů (skupin) se neliší;
- tento předpoklad si můžeme představit tak, že každá skupina (výběr) je výběrem ze stejného základního souboru;
- jestliže tento předpoklad platí, potom ve všech skupinách bude stejná úroveň rozptylu, z čehož vyplývá, že rozptyl základního souboru můžeme odhadnout pomocí rozptylu uvnitř skupin;
- na základě tohoto odhadu celkového rozptylu můžeme odhadnout i předpokládaný rozptyl mezi skupinami
- tento předpokládaný rozptyl porovnáme se skutečným rozptylem mezi skupinami;
- pokud je skutečný rozptyl mezi skupinami nepravděpodobně velký, což otestujeme F-testem, pak nulovou hypotézu o rovnosti průměrů skupin zamítneme.

Znamená to, že základem metodiky analýzy rozptylu je rozklad celkového rozptylu na dvě složky, což je graficky znázorněno na obrázku 9.1 .



Obrázek 9.1 – Schéma rozkladu celkového rozptylu na dvě složky

Jak již bylo uvedeno, ANOVA má široké možnosti použití a tomu odpovídající množství variant. V následujícím textu se zmíníme pouze o nejběžnějších z nich.

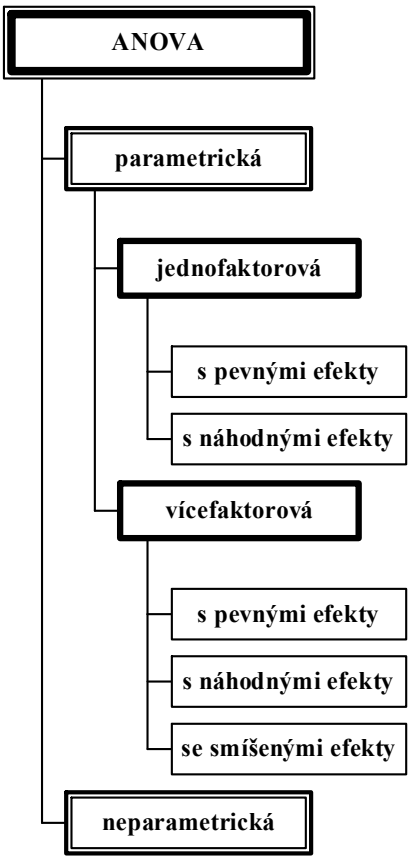
Základní varianta analýzy rozptylu, na které si vysvětlíme základní principy, se nazývá **jednofaktorová parametrická ANOVA**. Vychází z předpokladu, že jsou splněny následující podmínky:

- jednotlivé posuzované **výběry jsou navzájem** zcela **nezávislé**,
- všechny **výběry pocházejí z normálního rozdělení,**
- všechny **výběry mají homogenní rozptyl** (tj. všechny výběry pochází ze základních souborů se stejným rozptylem).

Pokud jsou splněny tyto podmínky, můžeme porovnávat průměry (tedy parametry) jednotlivých výběrů. Pokud tyto podmínky splněny nejsou (hlavně normalita), potom musíme použít neparametrickou obdobu analýzy rozptylu, která se ve své jednofaktorové podobě nazývá Kruskal – Wallisův test. V této souvislosti je nutné dodat, že vůči mírnému porušení předpokladů je ANOVA poměrně robustní (tedy její výsledky

a interpretace není zásadně ovlivněna mírným nesplněním předpokladů). Platí, že čím jsou větší výběry, tím je možné očekávat vyšší robustnost vůči nesplnění podmínek. Pro odolnost vůči nesplnění podmínky homogenity rozptylů je důležité, aby jednotlivé výběry měly stejnou velikost. Čím jsou výběry menší a čím jsou větší rozdíly v jejich četnostech, tím je použití neparametrické analýzy rozptylu oprávněnější.

Obrázek 9.2 ukazuje základní typy analýzy rozptylu. Podrobnější dělení je provedeno pro používanější – parametrickou – analýzu rozptylu, podobné členění je možné udělat i pro neparametrickou část, ale kromě jednofaktorové neparametrické analýzy rozptylu tyto metody nejsou moc používané.



Obrázek 9.2 - Rozdělení základních typů analýzy rozptylu

.

Kromě jednofaktorové analýzy rozptylu je možné posuzovat i vliv více faktorů. Poměrně běžně se používá dvoufaktorová ANOVA (ta bude podrobněji rozebrána v kapitole 9.2), troj- a vícefaktorové analýzy rozptylu jsou již poměrně vzácné, protože v těchto případech je značně obtížné sestavení vhodného modelu a interpretace výsledků, v neposlední řadě je také obtížené založení pokusu (nutnost mnoha pokusných skupin). Dříve od těchto složitějších variant odrazovala také technická náročnost výpočtu, ale v dnešní době, kdy je možné využít výkonné statistické programy, toto již není překážkou.

Modely analýzy rozptylu se také dělí podle typu úrovní posuzovaného faktoru (tyto úrovně se nazývají **efekty** nebo **hladiny**). Pokud jsou efekty pevně nastavované experimentátorem (např. pevně stanovené dávky hnojiv), potom hovoříme o **pevných** faktorech. Pokud jsou efekty výsledkem měření (je to tedy náhodná veličina), jedná se o **náhodné** efekty. Ve vícefaktorových modelech je možné se setkat i se smíšenými efekty, kdy část faktorů je pevných a část náhodných. V literatuře se setkáváme také s označením Model I (pro modely s pevnými efekty) a Model II (pro náhodné efekty), event. Model III (pro smíšené efekty).

9.1 Jednofaktorová analýza rozptylu

9.1.1 Základní model a výpočet tabulky analýzy rozptylu

Jak již bylo uvedeno, jednofaktorová ANOVA testuje následující hypotézu:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$$

(tj. střední hodnoty *k* skupin jsou shodné)

oproti hypotéze

H_1: alespoň mezi dvěma skupinami je statisticky významný rozdíl středních hodnot.

Základní model analýzy rozptylu je možné zapsat takto:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{9.1}$$

kde je

y_{ij} *j-tá* měřená hodnota (pozorování) v *i-té* skupině

μ konstanta společná pro všechny pozorování, tj. průměrná teoretická hodnota měřené veličiny za předpokladu, že by nepůsobily žádné faktory (za předpokladu zanedbání náhodné chyby)

α_i efekt - hodnota vyjadřující účinek úrovně A_i působícího faktoru A

ε_{ij} náhodná chyba s $N(0,\sigma^2)$, tj. ta část hodnoty y_{ij}, kterou není možné vysvětlit ani konstantní úrovní (μ) ani působením faktoru

Uspořádání dat pro jednofaktorovou analýzu rozptylu je v tabulce 9.1 . Zde jednotlivé symboly představují:

- $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ - **skupinové průměry** (průměry měřených hodnot ve skupinách – sloupcích),
- $n_1, n_2, \ldots, n_k$ - **skupinové četnosti** (nejlepší je, když jsou ve všech skupinách stejné, jednak to zaručuje maximální sílu a robustnost testu, jednak zjednodušuje výpočet, ovšem ANOVA se dá řešit i s rozdílnými četnostmi ve skupinách),
- $\bar{x}$ - **celkový aritmetický průměr** (je to průměrná hodnota skupinových průměrů (pro stejný počet pozorování ve skupinách) nebo vážený aritmetický průměr skupinových průměrů (pro rozdílný počet pozorování ve skupinách),
- N – **celkový počet všech prvků** ve všech skupinách (součet skupinových četností).

	Úroveň faktoru						Celkem
	A_1	A_2	…	A_i	…	A_k	
Opakování měření (jednotlivá pozorování)	x_{11}	x_{21}	…	x_{i1}	…	x_{k1}	
	x_{12}	x_{22}	…	x_{i2}	…	x_{k2}	
	…	…	…	…	…	…	
	…	…	…	…	…	…	
	x_{1n_1}	x_{2n_2}	…	x_{in_i}	…	x_{kn_k}	
průměry	$\bar{x}_1$	$\bar{x}_2$	…	$\bar{x}_i$	…	$\bar{x}_k$	$\bar{x}$
počet	n_1	n_2	…	n_i	…	n_k	N

Tabulka 9.1 – Uspořádání dat pro jednofaktorovou analýzu rozptylu

Základem řešení je tzv. tabulka analýzy rozptylu (viz tabulka 9.2):

Zdroj variability	Součet čtverců odchylek	Počet stupňů volnosti	Průměrný čtverec odchylek (rozptyl)	Testové kritérium
mezi skupinami	$S_G = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$	$DF_G = k - 1$	$M_G = \frac{S_G}{DF_G}$	$F = \frac{M_G}{M_R}$
uvnitř skupin (reziduální)	$S_R = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$	$DF_R = N - k$	$M_R = \frac{S_R}{DF_R}$	
Celkový	$S_C = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$	$DF_C = N - 1$		

Tabulka 9.2 – Schéma uspořádání tabulky analýzy rozptylu

V tabulce 9.2 se ve vzorci pro sumu čtverců odchylek mezi skupinami (S_G) používá člen n_i pouze tehdy, jsou-li četnosti ve třídách nestejné.

Nulovou hypotézu zamítáme, platí-li, že

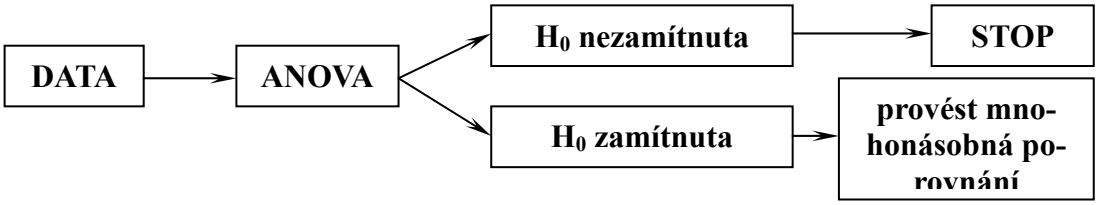
$$F \geq F_{1-\alpha;k-1;N-k},$$

kde $F_{1-\alpha;k-1;N-k}$ je kvantil Fisher-Snedecorova rozdělení na hladině významnosti $(1-\alpha)$ a se stupni volnosti (k-1) a (N-k).

Pokud nulovou hypotézu nezamítneme, potom výpočet končí – neprokázali jsme rozdíl střední hodnoty mezi jednotlivými skupinami a dále předpokládáme, že všechny výběry pochází z jednoho základního souboru nebo ze základních souborů se shodnou střední hodnotou.

V případě, že nulovou hypotézu zamítneme, potom se alespoň jedna skupina statisticky významně odlišuje od ostatních a nelze přijmout předpoklad, že všechny skupiny (výběry) pochází ze stejného základního souboru. V tomto případě nás zpravidla zajímá, mezi kterými skupinami nastal onen detekovaný rozdíl. K tomu slouží metody mnohonásobných porovnání.

Tento potup je uveden na obrázku 9.3 .



Obrázek 9.3 – Porovnání postupu analýzy rozptylu v případě zamítnutí a nezamítnutí H_0

9.1.2 Mnohonásobná porovnání

Metody mnohonásobných porovnání jsou vlastně také statistické testy, kterými porovnáváme vzájemné rozdíly mezi skupinovými průměry a posuzujeme statistickou významnost těchto rozdílů. Znamená to tedy, že mnohonásobných porovnání musíme udělat tolik, kolik je možných kombinací průměrů. Tyto testy nám odpoví na otázku – mezi kterými skupinami je statisticky významný rozdíl průměrů?

Metody mnohonásobných porovnání standardně používáme u Modelu I (tedy modelu s pevnými efekty). Zde máme přesně definovány jednotlivé úrovně faktorů a zajímají nás rozdíly právě mezi nimi. Pokud používáme Model II (model s náhodnými efekty), potom zpravidla mnohonásobná porovnání neprovádíme, protože pouze dokazujeme, že náhodně vybrané úrovně nějakého faktoru se od se liší, ale není nutné a účelné „přesně" testovat rozdíly mezi takto náhodně stanovenými úrovněmi – pokud úrovně vybereme v dalším výběru jinak, může být výsledek jiný.

Jako příklad si můžeme vzít první příklad z úvodu kapitoly 9 - pokus zaměřený na posouzení vlivu dávek hnojiva na růst semenáčků v lesní školce. Pokud založíme řízený pokus, kdy na jednotlivých záhonech použijeme přesně odstupňované dávky hnojiva (a zabezpečíme jinak plnou srovnatelnost podmínek), můžeme se ptát, jak takto pevně stanovené dávky hnojiva ovlivňují růst. To je příklad na Model I (s pevnými efekty) a zde má smysl použít metody mnohonásobného porovnání.

Naproti tomu, jestliže pouze získáme náhodné údaje o hnojení daným hnojivem ve školkách (např. dotazem – někde údaje poskytnou, někde neposkytnou, dávky hnojiva a jeho používání se liší – jsou dány nejen doporučením výrobce, ale i místní zkušeností, ekonomickými možnostmi, apod.), potom zřejmě testování konkrétních rozdílů mezi školkami nemá vypovídací schopnost. Údaje jsou náhodné a při další akci by mohlo dojít ke zcela jiným výsledkům na základě toho, jaké údaje bychom „sehnali". Na druhé straně, pokud ovšem víme, že školky, kde jsme se dotazovali, poskytují korektní údaje (např. z předchozí zkušenosti) a jsou např. reprezentanty určitého způsobu hospodaření v daných přírodních a klimatických podmínkách, a že rozdíly mezi nimi můžeme zevšeobecnit, potom za určitých okolností může mít i v tomto případě metoda mnohonásobného porovnání smysl.

Z předchozího příkladu vyplývá, že rozdíly mezi Modelem I a II jsou někdy velmi jemné a mnohdy závisí na kontextu a na otázkách, které si klademe. Jiným typickým příkladem může být posuzovávání rozdílů určité vlastnosti mezi lokalitami, odkud byly odebrány pokusné vzorky. Pokud vybereme např. 5 lokalit pevně (a máme zdůvodněno, proč právě tyto lokality), může se jejich výběr považovat za vliv s pevným efektem a provádíme mnohonásobná porovnání. Pokud ovšem lokality vybereme náhodně (např. podle dopravní dostupnosti apod.) z mnoha možných, které by přicházely v úvahu a jinak se podstatně neliší, potom pomocí analýzy rozptylu pouze dokážeme, že mezi náhodně vybranými lokalitami je (nebo není) statisticky významný rozdíl ve studované vlastnosti (tj. že daná vlastnost má na určitém území jistou míru variability), ale zkoumat rozdíly mezi konkrétními lokalitami už nemá smysl.

Kromě výše uvedených existuje ještě jen speciální typ mnohonásobného porovnání – srovnání pokusných zásahů s kontrolou (např. použití různých hnojiv a kontrolního pokusu bez aplikace hnojiva – zajímá nás hlavně vliv aplikace hnojiva oproti jeho nepoužití, ale ne už tolik rozdíly mezi jednotlivými hnojivy).

Metod mnohonásobného porovnání je celá řada – mezi nejznámější patří metoda Tukeyho, Scheffeho, Duncana, SKN (Student-Newman-Keuls) nebo Bonfferoniho. Každá z těchto metod má svoje výhody i nevýhody, jejich množství (předchozí výčet není zdaleka úplný) už samo o sobě naznačuje, že žádná z nich není naprosto všeobecně přijímána jako ideální.

V tomto textu si uvedeme ty metody, které si získaly nejvyšší „popularitu" a jsou také součástí většiny statistických programů – Tukeyho a Scheffeho metodu a specializovaný Dunnettův test pro porovnání s kontrolou.

Testy mnohonásobného porovnání mají obecně nižší sílu testu než ANOVA sama. To může někdy vést k paradoxní situaci, kdy ANOVA zamítne nulovou hypotézu (tj. indikuje statisticky významný rozdíl alespoň mezi dvěma průměry) a přitom testy mnohonásobného porovnání žádný rozdíl neukáží jako významný. K tomuto jevu dochází hlavně tehdy, je-li nulová hypotéza analýzou rozptylu zamítnuta „těsně" (tj. testové kritérium je jen o málo vyšší než kritická hodnota), potom testy s nižší silou (tj. méně „přísné", s větší tendencí nezamítnout nulovou hypotézu) nemusí detekovat žádný statisticky významný rozdíl.

9.1.2.1 Tukeyho metoda mnohonásobného porovnání

Je to vlastně obdoba t-testu a testuje se nulová hypotéza

H_0: $\mu_A = \mu_B$, (A ≠ B) oproti alternativní hypotéze $\mathbf{H_1}$: $\mu_A \neq \mu_B$,

tj. nulová hypotéza tvrdí, že střední hodnoty porovnávaných skupin *A* a *B* se neliší.

Testové kritérium má tvar

$$q = \frac{\overline{x}_A - \overline{x}_B}{SE} \tag{9.2}$$

kde SE (střední chyba - směrodatná odchylka - rozdílu průměrů skupin *A* a *B*) má tvar pro shodné počty pozorování (*n*) ve skupinách *A* a *B*

$$SE = \sqrt{\frac{M_R}{n}} \tag{9.3}$$

kde M_R je reziduální rozptyl (viz tabulku 9.2). Z hlediska síly testu a případné robustnosti (odolnosti, necitlivosti) k porušení předpokladů analýzy rozptylu je u tohoto testu doporučen stejný počet pozorování ještě důrazněji než u „základní" analýzy rozptylu (tj. u výpočtu podle tabulky 9.2) .

Pro různé počty pozorování (n_A, n_B) ve srovnávaných skupinách *A* a *B* platí tvar

$$SE = \sqrt{\frac{M_R}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} \tag{9.4}$$

Testové kritérium *q* se porovná s kritickou hodnotou $q_{\alpha;\ N-k;\ k;}$ (počet stupňů volnosti *N-k* se často označuje jako ν), která se nazývá „studentizované rozpětí" (studentized range) a je součástí podrobnějších statistických tabulek (zde **tabulka 1 v příloze**). Pokud je hodnota testového kritéria *q* menší než kritická hodnota, potom přijímáme nulovou hypotézu o rovnosti středních hodnot obou porovnávaných skupin.

Tento test musíme provést pro všechny možné kombinace skupin.

Tukeyho test patří k nejužívanějším a považuje se také za jeden z nejlepších z hlediska vhodného kompromisu síly testu a možnosti výskytu chyby I. druhu (o chybě I. a II. druhu a jejich vzájemných vztazích viz I.díl, kapitola 7.3). Jeho modifikací je SNK test, kdy výpočty testového kritéria jsou stejné, liší se pouze kritické hodnoty, které užívají jiných stupňů volnosti (podrobněji viz např. LEPŠ 1996 nebo ZAR 1984). Uvádí se (LEPŠ 1996), že SNK test má vyšší sílu testu (menší pravděpodobnost chyby II. druhu, je tedy „přísnější", má vyšší schopnost správně zamítnout ve skuteč-

nosti neplatnou hypotézu), ale na druhé straně má vyšší pravděpodobnost chyby I. druhu (skutečná pravděpodobnost chyby I druhu, tedy „nebezpečí“, že zamítneme ve skutečnosti platnou hypotézu, je u SNK testu vyšší než deklarovaná hladina významnosti).

9.1.2.2 Scheffeho metoda mnohonásobného porovnání

Tento test se také nazývá testem násobných kontrastů (multiple contrasts) a je považován za slabší než Tukeyho test (tj. má vyšší „náchylnost“ k chybě II. druhu, tedy obvykle detekuje méně rozdílů mezi středními hodnotami než Tukeyho test).

Nulová hypotéza je stejná jako u Tukeyho testu, testové kritérium se nazývá S a vypočítá se podle vztahu

$$S = \frac{\left|\bar{x}_A - \bar{x}_B\right|}{SE} \tag{9.5}$$

kde je

$$SE = \sqrt{M_R\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} \tag{9.6}$$

Kritická hodnota je

$$S_\alpha = \sqrt{(k-1)\cdot F_{\alpha;k-1;N-k}} \tag{9.7}$$

Určitou praktickou výhodou tohoto testu je fakt, že k jeho provedení nepotřebujeme žádné speciální hodnoty (zpravidla uváděné jen v rozsáhlejších specializovaných statistických tabulkách) jako jsou hodnoty *q* pro Tukeyho test, ale vystačíme s „běžnou“ hodnotou F, jejíž tabulky jsou součástí nejen všech statistických tabulek, ale i většiny učebnic, a také je možné je získat přímo v tabulkových kalkulátorech (např. v Excelu funkce FINV).

9.1.2.3 Dunnettova metoda mnohonásobného porovnání s kontrolou

Tento test slouží k testování jiné varianty než předchozí dva testy – nikoli k porovnání průměrů všech skupin mezi sebou, ale k porovnání jednotlivých skupin se skupinou kontrolní. Pokud máme celkem *k* skupin, z nichž jedna je kontrolní, potom pomocí Dunnettova testu provedeme *k-1* porovnání (ostatní skupiny versus kontrola).

Nulová hypotéza je formulována jako

$$\mathbf{H_0: \mu_A = \mu_{kontrola}}$$

oproti alternativní hypotéze (může být konstruována jako jednostranná – v tom případě nám záleží na tom, zda je průměr porovnávané skupiny vyšší nebo nižší než kontrolní - nebo oboustranná – v tom případě je důležité pouze to, že se oba průměry liší, který z nich je větší nebo menší, není již podstatné):

$\mathbf{H_1: \mu_A \neq \mu_{kontrola}}$ (oboustranná)

$\mathbf{H_1: \mu_A \geq \mu_{kontrola}}$ (jednostranná)

$\mathbf{H_1: \mu_A \leq \mu_{kontrola}}$ (jednostranná)

Testové kritérium je obdobné jako u Tukeyho testu

$$q = \frac{\overline{x}_{kontrola} - \overline{x}_A}{SE} \tag{9.8}$$

kde je SE (pro stejnou velikost porovnávané a kontrolní skupiny)

$$SE = \sqrt{\frac{2M_R}{n}} \tag{9.9}$$

a pro rozdílnou velikost kontrolní a porovnávané skupiny

$$SE = \sqrt{M_R\left(\frac{1}{n_A} + \frac{1}{n_{kontrola}}\right)} \tag{9.10}$$

Kritická hodnota je $q^*_{\alpha(1);N-k;p}$ pro jednostrannou hypotézu, kde $\alpha(1)$ znamená hodnoty jednostranného studentizovaného rozpětí q pro hladinu významnosti α; pro oboustrannou hypotézu je kritická hodnota $q^*_{\alpha(2);N-k;p}$ – symbol $\alpha(2)$ znamená hodnoty oboustranného studentizovaného rozpětí q^* pro hladinu významnosti α. Kritické hodnoty jsou tabelovány ve speciálních tabulkách (jiných než pro Tukeyho test) – zde v příloze **tabulka 2**.

Testování se provádí následujícím způsobem:

- všechny průměry uspořádáme podle velikosti od nejmenšího do největšího
- testujeme kontrolní skupinu postupně oproti ostatním, přičemž začínáme od největších rozdílů (pokud bude pro největší rozdíl přijata nulová hypotéza, menší rozdíly už nemusíme testovat, zde bude samozřejmě platit stejný výsledek)
- počet stupňů volnosti p se určí podle „vzdálenosti“ porovnávaných průměrů (jestliže např. porovnáváme druhý průměr s pátým, je hodnota $p = 4$ (2, 3, 4, 5), pokud první s druhým, je hodnota $p = 2$ (1, 2) apod.

Podobný způsob testování se používá u SNK testu.

V případě oboustranného testu zamítáme nulovou hypotézu, jestliže testové kritérium je menší než $q^*_{\alpha(2);N-k;p}$. Jestliže použijeme jednostranný test, potom záleží na typu alternativní hypotézy:

- pro H_1: $\mu_A \geq \mu_{kontrola}$ zamítáme nulovou hypotézu, jestliže platí $q \geq q^*_{\alpha(1);N-k;p}$
- pro H_1: $\mu_A \leq \mu_{kontrola}$ zamítáme nulovou hypotézu, jestliže platí $|q| \geq q^*_{\alpha(1);N-k;p}$,

tj. $q < -\ q^*_{\alpha(1);N-k;p}$,

Uvádí se, že síla Dunnettova testu je vyšší než u předchozích mnohonásobných porovnání (provádíme méně testů – jen $k-1$). Vzhledem k tomu, že porovnání kontrolní skupiny s ostatními je hlavním cílem tohoto testu, doporučuje se, aby kontrolní skupina měla více členů než ostatní, a to o něco méně než $\sqrt{k-1}$-krát více než ostatní skupiny, pro které platí požadavek stejného počtu pozorování. Např. máme-li 5 skupin (1 kontrolní a 4 „k porovnání“) a počet členů v ostatních skupinách je 8, potom doporučený počet členů kontrolní skupiny je „o něco méně než $8 \cdot \sqrt{5-1}$“, tj. asi 13-15.

Postup výpočtu jednofaktorové analýzy rozptylu si ukážeme na příkladu.

Příklad 9.1:

V rámci výzkumu vlastností dřeva z různých lokalit byla také porovnávána hustota dřeva (v kg/m^3). Rozhodněte, zdali mezi hustotou dřeva ze čtyř různých lokalit je statisticky významný rozdíl. Měřené údaje jsou v tabulce 9.3 .

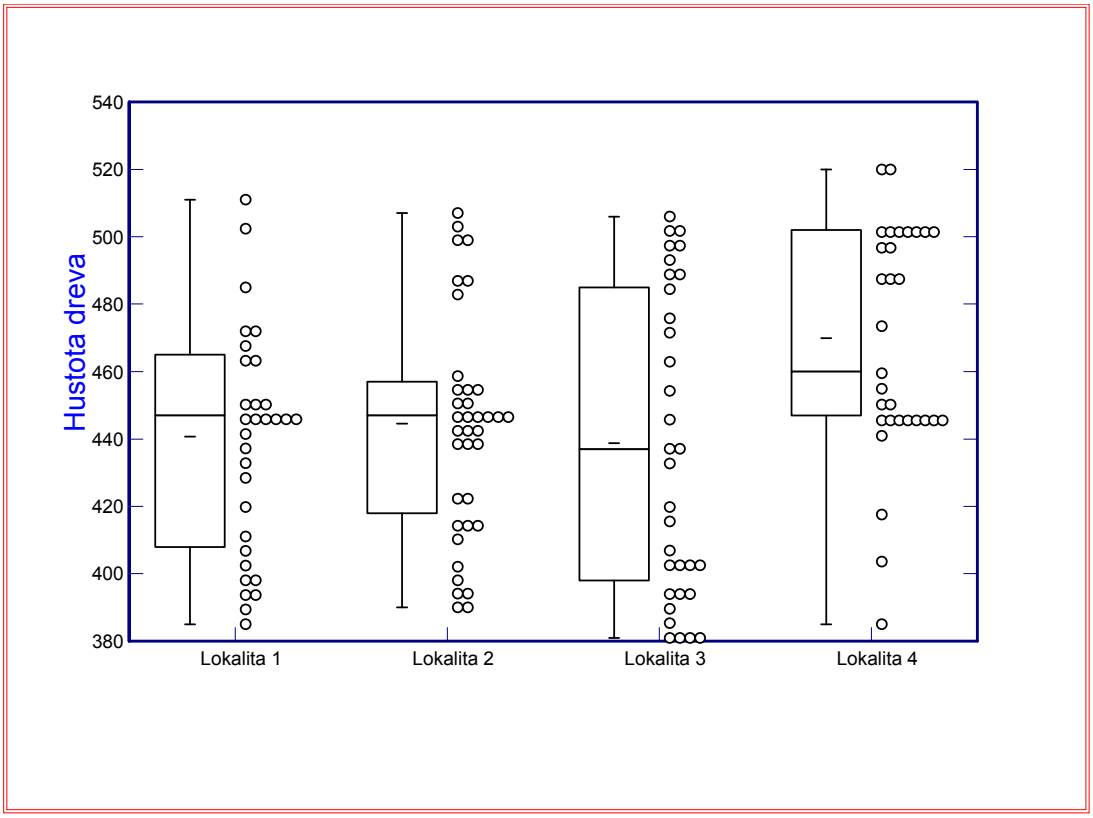
Číslo měření	Lokalita 1	Lokalita 2	Lokalita 3	Lokalita 4	Číslo měření	Lokalita 1	Lokalita 2	Lokalita 3	Lokalita 4
1	454	399	382	490	21	442	411	398	418
2	467	447	404	505	22	505	439	437	505
3	470	395	440	404	23	511	450	474	446
4	476	443	466	520	24	486	507	406	457
5	435	460	424	385	25	465	486	405	446
6	448	418	381	500	26	452	490	479	448
7	395	505	489	450	27	474	502	416	448
8	447	446	506	490	28	447	395	456	452
9	438	457	501	520	29	399	391	491	450
10	432	446	497	502	30	447	501	394	506
11	422	448	485	501	31	395	439	398	504
12	413	448	504	506	32		442	383	
13	408	452	448	452	33		423	383	
14	399	450	438	474	34		454		
15	454	406	499	447	35		426		
16	404	416	502	502	36		489		
17	390	416	408	447	37		390		
18	385	458	406	490	38				
19	450	456	391	443	39				
20	450	448	387	460	40				

Tabulka 9.3 – Měřené hodnoty hustoty dřeva (v kg/m^3)

V tomto případě považujeme jednotlivé lokality za „pevně“ vybrané a jde nám pouze o porovnání hustoty dřeva dané dřeviny mezi těmito lokalitami – zadání tedy považujeme za model s pevnými efekty (model I).

Nejprve si rozebereme podstatu úlohy, kterou máme řešit. Chceme zjistit, zda růstové podmínky na jednotlivých lokalitách mají vliv na hustotu dřeva. Tento vliv by se projevil odlišnými hodnotami aritmetických průměrů jednotlivých skupin. Naším úkolem je vyšetřit, zdali údaje, které máme k dispozici, nás opravňují k předpokladu, že hustota dřeva na všech lokalitách je stejná.

Před provedením samotné analýzy rozptylu musíme vyšetřit splnění předpokladů, především normality výběrů a homogenity rozptylu. Pokud použijeme testy normality (např. z kapitoly 8), zjistíme, že předpoklad normálního rozdělení je ve všech případech splněn. Bližší představu o jednotlivých výběrech podávají krabicové a tečkové grafy na obrázku 9.4 .

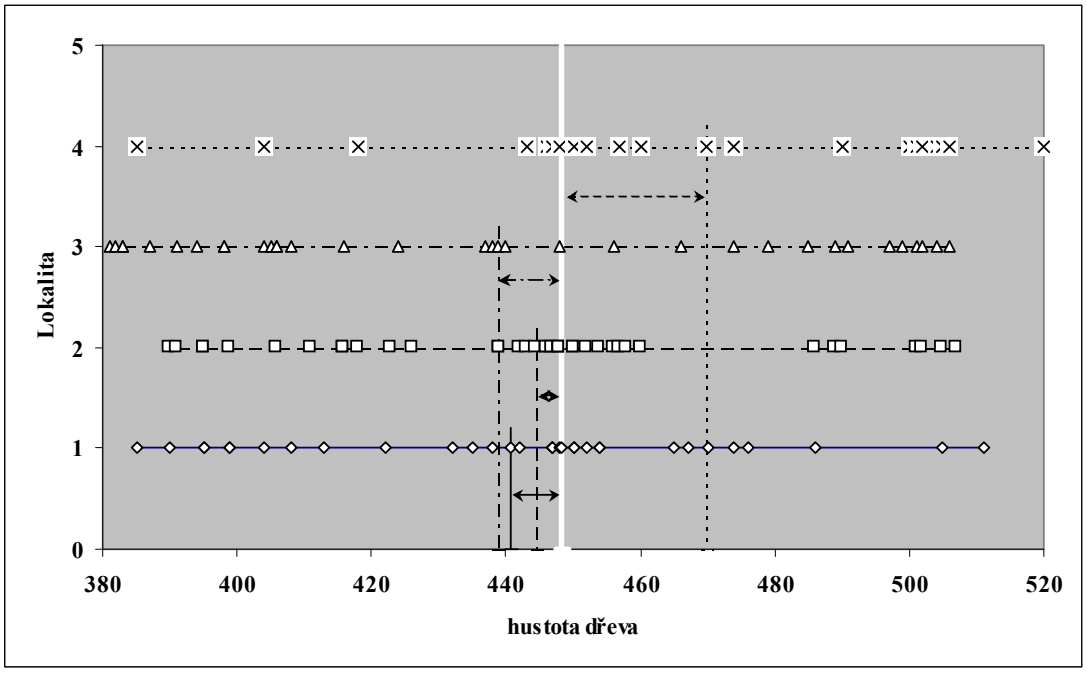


Obrázek 9.4 – Krabicové grafy hustoty dřeva z porovnávaných lokalit

Z grafů je zřejmé, že výběry mají oblasti lokálních koncentrací dat, ale především vzhledem k poměrně vysokému počtu hodnot v jednotlivých výběrech byly všechny přijaty jako normální. Také je zřejmé, že porovnání aritmetických průměrů (krátké čárky v krabicovém grafu) signalizuje, že výběry z lokality 1 – 3 jsou si z hlediska polohy dosti blízké (kolem 440 kg/m^3), lokalita 4 se poněkud odchyluje (kolem 470 kg/m^3). Také test homogenity rozptylu pro více výběrů (Bartletův test) nezamítl nulovou hypotézu, tedy všechny rozptyly budeme považovat za shodné. Tím jsou splněny základní předpoklady pro provedení analýzy rozptylu.

Rozložení měřených hodnot a skupinových průměrů vůči celkovému aritmetickému průměru je zobrazeno na obrázku 9.5 Na horizontálních úrovních (Lokalita 1 – 4) jsou zobrazeny jednotlivé měřené hodnoty. Každá lokalita má stejným typem čáry vyznačen svůj skupinový průměr ($\bar{x}_i$) – Lokalita 1 plnou čarou, Lokalita 2 dlouze čárkovaně, Lokalita 3 čerchovaně a Lokalita 4 krátce čárkovaně. Bílou čarou je zobrazen celkový aritmetický průměr ($\bar{x}$). Oboustranné šipky vyznačují vzdálenosti skupinových průměrů od průměru celkového. Tyto vzdálenosti vyjadřují tu část celkové variability, která je vysvětlitelná rozdíly mezi skupinami a tedy působením zkoumaného faktoru – Lokality. Čím je šipka delší, tím je větší odchylka od celkového průměru, což znamená vyšší vliv faktoru. Druhá část celkové variability – vnitroskupinová –

není na obrázku znázorněna (je tvořena odchylkami všech měřených hodnot dané skupiny od příslušného skupinového průměru – tedy např. u Lokality 1 odchylkami všech „kosočtverců“ od plné tenké čáry). Úkolem analýzy rozptylu je posoudit, zda podíl variability mezi skupinami oproti vnitroskupinové variabilitě je tak velký, že jej nelze vysvětlit náhodnými chybami a vlivy, ale také působením posuzovaného faktoru.



Obrázek 9.5 – Zobrazení měřených hodnot a skupinových průměrů vůči celkovému průměru

Základní potřebné charakteristiky všech výběrů jsou v tabulce 9.4 a výsledky výpočtu analýzy rozptylu jsou v tabulce 9.5 .

Výběr	Počet	Průměr	Rozptyl
Lokalita 1	31	440.6	1114.3
Lokalita 2	37	444.6	1094.8
Lokalita 3	33	438.7	2009.1
Lokalita 4	31	469.9	1207.7
Celkem	132	448.1	

Tabulka 9.4 – Základní charakteristiky porovnávaných výběrů

Zdroj variability	Součet čtverců odchylek	Počet stupňů volnosti	Průměrný čtverec odchylek (rozptyl)	Testové kritérium	Hodnota P	Kritická hodnota
Mezi skupinami	19863.67	3	6621.22	4.89	0.003	2.68
Uvnitř skupin	173364.59	128	1354.41			
Celkem	193228.27	131				

Tabulka 9.5 – Výsledky analýzy rozptylu

Z tabulky 9.5 vyplývá, že nulová hypotéza o „nulovém“ vlivu jednotlivých lokalit na hustotu dřeva byla zamítnuta (testové kritérium 4.89 je větší než kritická hodno-

ta 2.68) a znamená to, že alespoň mezi dvěmi lokalitami existuje statisticky významný rozdíl v hustotě dřeva.

Vzhledem k tomu, že tento příklad chápeme jako pokus s pevnými efekty, musíme pokračovat dále a zjistit, mezi kterými lokalitami tento rozdíl existuje. To provedeme pomocí metody mnohonásobného porovnání. Dále jsou uvedeny výsledky obou zde uváděných metod – Tukeyho i Scheffeho.

a) **Tukeyho metoda mnohonásobného porovnání**

Pro výpočet použijeme vzorce 9.2 a 9.4 (jednotlivé výběry mají různý počet prvků) a výsledky jsou uvedeny v tabulce 9.6 .

Porovnání (čísla označují jednotlivé Lokality)	Rozdíl průměrů	SE (podle vzorce 9.4)	Testové kritérium q	Kritická hodnota q	Výsledek porovnání (H_0 zamítáme/ /nezamítáme)
3 - 4	-31.21	6.51	4.79	3.68	**Zamítáme**
3 - 2	- 5.84	6.23	0.94	3.68	Nezamítáme
3 - 1	- 1.92	6.51	0.29	3.68	Nezamítáme
1 - 4	-29.29	6.61	4.43	3.68	**Zamítáme**
1 - 2	- 3.92	6.34	0.62	3.68	Nezamítáme
2 - 4	-25.37	6.34	4.00	3.68	**Zamítáme**

Tabulka 9.6 – Výsledky Tukeyho metody mnohonásobného porovnání (výsledné hodnoty jsou zaokrouhleny na dvě desetinná místa)

Pro výpočet SE se použila hodnota M_R = 1354.41, kritická hodnota byla převzata z tabulek pro $q_{0,05;4;128}$.

Z výsledků v tabulce je vidět, že významné rozdíly (tj. případy, kdy testové kritérium je vyšší než kritická hodnota) existují pro dvojice 3- 4, 1- 4 a 2 – 4. Mezi ostatními skupinami (lokalitami) nebyly významné rozdíly potvrzeny. Znamená to, že hustota dřeva na Lokalitě 4 se významně liší od všech ostatních lokalit, tedy že z hlediska hustoty dřeva tvoří Lokalita 4 jednu skupinu a ostatní lokality (1, 2, 3) skupinu druhou.

b) **Scheffeho metoda mnohonásobného porovnání**

V tomto případě použijeme vzorce 9.5 a 9.6 , jejichž výsledky jsou uvedeny v tabulce 9.7 .

Porovnání (čísla označují jednotlivé Lokality)	Rozdíl průměrů	SE (podle vzorce 9.6)	Testové kritérium S	Kritická hodnota S	Výsledek porovnání (H_0 zamítáme/ /nezamítáme)
3 - 4	-31.21	9.21	3.39	2.83	**Zamítáme**
3 - 2	- 5.84	8.81	0.66	2.83	Nezamítáme
3 - 1	- 1.92	9.21	0.21	2.83	Nezamítáme
1 - 4	-29.29	9.35	3.13	2.83	**Zamítáme**
1 - 2	- 3.92	8.96	0.44	2.83	Nezamítáme
2 - 4	-25.37	8.96	2.83	2.83	Nezamítáme

Tabulka 9.7 – Výsledky Scheffeho metody mnohonásobného porovnání

Kritická hodnota byla vypočítána podle vzorce 9.7, kde hodnota $F_{0,05;3;128} = 2.675$.

Vidíme, že výsledky jsou podobné, až na srovnání lokalit 2 a 4, kde nebyla (na rozdíl od Tukeyho testu) zamítnuta nulová hypotéza (nezaokrouhlené hodnoty jsou S = 2,831 a $S_\alpha = 2.833$). Tento výsledek souvisí s faktem, že Scheffeho test má nižší sílu, tedy i nižší schopnost zamítnout nesprávnou hypotézu, což se projeví právě v takovýchto hraničních případech (kdy testové kritérium a kritická hodnota si jsou dosti blízké).

9.2 Dvoufaktorová analýza rozptylu

9.2.1 Základní model dvoufaktorové analýzy rozptylu a její varianty

Pokud posuzujeme **vliv více faktorů na určitou veličinu**, používáme obecně vícefaktorovou analýzu rozptylu. Její principy si ukážeme na nejjednodušší variantě – **dvoufaktorové analýze rozptylu**, ANOVA s vyšším počtem parametrů se řeší obdobně, ale výpočet je technicky náročnější a interpretace složitější.

Základní model dvoufaktorové analýzy rozptylu je následující:

:

$$y_{ij} = \mu + \alpha_i + \beta_j + (\tau_{ij}) + \varepsilon_{ij} \qquad (9.11)$$

kde je

- y_{ij} měřená hodnota (pozorování) v ovlivněná *i-tou* úrovní faktoru A a *j-tou* úrovní faktoru B
- μ konstanta společná pro všechny pozorování, tj. průměrná teoretická hodnota měřené veličiny za předpokladu, že by nepůsobily žádné faktory (za předpokladu zanedbání náhodné chyby)
- α_i efekt - hodnota vyjadřující účinek úrovně A_i působícího faktoru A
- β_j efekt - hodnota vyjadřující účinek úrovně B_i působícího faktoru B
- τ_{ij} interakce mezi faktory (tento člen je volitelný, protože mohou existovat modely s interakcí i bez interakce)
- ε_{ij} náhodná chyba s $N(0,\sigma^2)$, tj. ta část hodnoty y_{ij}, kterou není možné vysvětlit ani konstantní úrovní (μ) ani působením faktorů

Z modelu 9.11 vyplývá, že v případě dvoufaktorové analýzy rozptylu testujeme více nulových hypotéz než v případě jednofaktorové analýzy rozptylu:

1) *„vliv faktoru A je nevýznamný“,*
2) *„vliv faktoru B je nevýznamný“,*
3) *„vliv interakce T je nevýznamný“* (tato hypotéza je „nepovinná“).

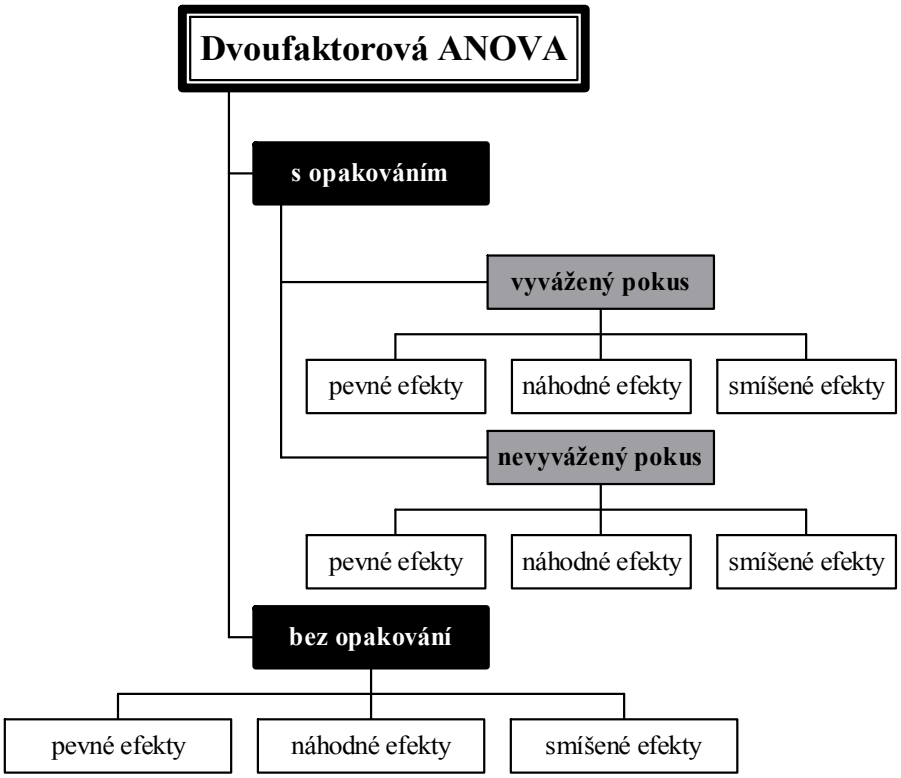
Pokud je vliv interakce (přesnější by bylo říci, že interakce sama) je nulový (nevýznamný), potom je vliv faktorů čistě aditivní. Znamená to, že rozdíl v průměrech mezi jednotlivými úrovněmi faktoru A je konstantní a není nijak ovlivněn hladinami faktoru B (a naopak, samozřejmě). Naopak pokud je interakce přítomna, potom jednotlivé úrovně jednoho faktoru ovlivňují hodnoty úrovní druhého faktoru. Zda je in-

terakce „přítomna“ (tj. zda má logický smysl a do modelu reálně patří) je nutné rozhodnout na základě analýzy problému, který se pomocí analýzy rozptylu řeší. Pokud se dospěje k názoru, že interakce je z hlediska podstaty problému nelogická nebo nepodstatná, model se zúží a řeší se jako ANOVA bez interakce.

Dvoufaktorová ANOVA zahrnuje několik možných variant uspořádání pokusu, z nichž nejdůležitější jsou uvedeny na obrázku 9.6 :

- **ANOVA s opakováním** – pro každou kombinaci úrovní obou faktorů (pro každou buňku, celu) je změřeno několik hodnot,
  - **vyvážený pokus** – počet měřených hodnot ve všech buňkách je stejný,
  - **nevyvážený pokus** - počet měřených hodnot v buňkách je různý,
- **ANOVA bez opakování** – v každé buňce je jen jedna měřená hodnota.

Nejvýhodnější je používat vyváženou analýzu rozptylu s opakováním – tato varianta je nejjednodušší na výpočet a má nejvyšší sílu testu. Pokud není možné vyvážený model pokusu dodržet, je vhodné používat tzv. proporční uspořádání (podrobněji v kapitole 9.2.3).



Obrázek 9.6 – Základní členění dvoufaktorové analýzy rozptylu

Pokusy s opakováním mohou mít pevné, náhodné nebo smíšené faktory – zde je toto členění důležité, protože pro jednotlivé varianty platí různé způsoby výpočtu (a také se liší interpretace, jak již bylo vysvětleno v předchozí kapitole).

9.2.2 Dvoufaktorová ANOVA s opakováním a vyváženým modelem

Toto uspořádání je základním typem dvoufaktorové analýzy rozptylu. Uspořádání měřených hodnot a faktorů je uvedeno v tabulce 9.8 . Faktor A má *a* úrovní (A_1,

A_2, …, A_a), faktor B má *b* úrovní (B_1, B_2, …, B_b). V každé buňce je *n* opakování měřené veličiny (např. v buňce tvořené kombinací úrovní A_1 a B_1 jsou opakování x_{111}, x_{112}, …, x_{11n}).

Tabulka dvoufaktorové analýzy rozptylu je velmi podobná jednofaktorové analýze rozptylu, pouze je zde více řádků pro faktory a přibyl řádek pro interakci. Její schéma je v tabulce 9.9 . Pro lepší pochopení vztahů v této tabulce jsou na obrázku 9.7 graficky znázorněny vztahy mezi jednotlivými zdroji variability.

Testová kritéria se v tomto případě počítají tři (pro významnost faktoru A, faktoru B a pro interakci) a jejich konkrétní vzorec závisí na typu faktorů (zda se jedná o faktory pevné, náhodné nebo smíšené). Přehled testových kritérií pro jednotlivé kombinace faktorů udává tabulka 9.10 .

Testová kritéria se porovnávají s kritickými hodnotami $F_{\alpha;f1;f2}$ F-rozdělení, kde je α hladina významnosti, *f1* jsou stupně volnosti dané čitatelem výrazu pro výpočet F-kritéria v tabulce 9.10 a *f2* jsou stupně volnosti dané jmenovatelem výrazu pro výpočet F-kritéria v tabulce 9.10 (příslušné výrazy pro výpočet stupňů volnosti jsou ve sloupci „počet stupňů volnosti“ tabulky 9.9). Např. kritická hodnota pro Model I a pro faktor A (vzorec M_A/M_R podle tabulky 9.10) je $F_{\alpha;a-1;ab(n-1)}$, protože počet stupňů volnosti pro M_A je *a-1* (podle tabulky 9.9), pro M_R je to *ab(n-1)*.

		FAKTOR A			
		A_1	A_2	…	A_a
FAKTOR B	B_1	x_{111} x_{112} … x_{11n}	x_{211} x_{212} … x_{21n}	… … … ….	x_{a11} x_{a12} … x_{a1n}
	B_2	x_{121} x_{122} … x_{12n}	x_{221} x_{222} … x_{22n}	… … … ….	x_{a21} x_{a22} … x_{a2n}
	…	… … … …	… … … ….	… … … ….	… … … …
	B_b	x_{1b1} x_{1b2} … x_{1bn}	x_{2b1} x_{2b2} … x_{2bn}	… … … ….	x_{ab1} x_{ab2} … x_{abn}

Tabulka 9.8 – Uspořádání dat pro dvoufaktorovou analýzu rozptylu s opakováním a s vyváženým modelem

Zdroj variability	Součet čtverců odchylek	Počet stupňů volnosti	Průměrný čtverec odchylek (rozptyl)
Faktor A	$S_A = \frac{\sum_{i=1}^{a}\left(\sum_{j=1}^{b}\sum_{k=1}^{n} x_{ijk}\right)^2}{bn} - C$	$DF_A = a - 1$	$M_A = \frac{S_A}{DF_A}$
Faktor B	$S_B = \frac{\sum_{j=1}^{b}\left(\sum_{i=1}^{a}\sum_{k=1}^{n} x_{ijk}\right)^2}{an} - C$	$DF_B = b - 1$	$M_B = \frac{S_B}{DF_B}$
Interakce A x B	$S_{AB} = S_F - S_A - S_B$	$DF_{AB} = (a-1)(b-1)$	$M_{AB} = \frac{S_{AB}}{DF_{AB}}$
Variabilita vysvětlená faktory a interakcí	$S_F = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\sum_{k=1}^{n} x_{ijk}\right)^2}{n} - C$	$DF_F = ab - 1$	
Variabilita uvnitř buněk (reziduální)	$S_R = S_C - S_F$	$DF_R = ab(n-1)$	$M_R = \frac{S_R}{DF_R}$
Celková variabilita	$S_C = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} x_{ijk}^2 - C$	$DF_C = N - 1$	

$$\text{kde je } C = \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} x_{ijk}\right)^2}{N} \quad \text{a} \quad N = abn$$

Tabulka 9.9 – Schéma uspořádání dvoufaktorové analýzy rozptylu s opakováním a s vyváženým modelem

Pokud platí, že testové kritérium je vyšší než kritická hodnota, potom zamítáme nulovou hypotézu o nevýznamnosti příslušného faktoru nebo interakce a dále pracujeme s předpokladem, že daný faktor nebo interakce má na měřenou veličinu statisticky významný vliv.

Tabulka 9.9 obsahuje výpočet vlivu interakce. Používá se v těch případech, kdy je interakce má zřejmý reálný smysl nebo si nejsme jisti (ale nemůžeme ji určitě vyloučit). V některých případech ovšem předem (apriori) víme, že interakce do modelu nepatří, že oba faktory působí čistě aditivně. Potom se použije výpočet bez interakce.

Ten se liší od schématu uvedeného v tabulce 9.9 ve způsobu výpočtu testového kritéria F pro faktory A a B - získáme je tak, že hodnoty M_A a M_B dělíme hodnotou

$$M_R^* = \frac{S_{AB} + S_R}{DF_{AB} + DF_R} \tag{9.12}$$

z čehož vyplývá, že v tomto případě veškerou variabilitu nevysvětlenou působením faktorů A a B považujeme za náhodnou složku. Pochopitelně odpadá třetí kritérium pro interakci.

Obrázek 9.7 – Vztahy mezi jednotlivými zdroji variability ve dvoufaktorové analýze rozptylu s opakováním

Testovaný efekt	Model I (oba faktory jsou pevné)	Model II (oba faktory jsou náhodné)	Model III (faktor A je pevný, faktor B je náhodný)
faktor A	$\frac{M_A}{M_R}$	$\frac{M_A}{M_{AB}}$	$\frac{M_A}{M_{AB}}$
faktor B	$\frac{M_B}{M_R}$	$\frac{M_B}{M_{AB}}$	$\frac{M_B}{M_R}$
interakce AxB	$\frac{M_{AB}}{M_R}$	$\frac{M_{AB}}{M_R}$	$\frac{M_{AB}}{M_R}$

Tabulka 9.10 – Výpočet testového kritéria F pro různé typy dvoufaktorové analýzy rozptylu s opakováním (podle ZAR 1984)

Pokud je vliv určitého faktoru považován za významný, potom je možné použít metod mnohonásobného porovnání ke zjištění, mezi kterými skupinami tento statisticky významný rozdíl existuje. Používá se poněkud upravených metod mnohonásobného porovnání jednofaktorové analýzy rozptylu, jež jsou uvedeny v kapitole 9.1.2. Každý faktor se posuzuje samostatně. Jestliže např. vyjdou oba faktory jako významné, potom se zvlášť posuzují rozdíly mezi úrovněmi faktoru A a zvlášť mezi úrovněmi faktoru B.

Použití metod mnohonásobného porovnání si ukážeme na Tukeyho metodě, jejíž základní verze je popsána v kapitole 9.1.2.1. Vycházíme ze vzorců 9.2 a 9.3, kde se upraví hodnota *n* ve jmenovateli vzorce 9.3 tak, že používáme počet hodnot toho fak-

toru, který právě neposuzujeme (např. jestliže zjišťujeme rozdíly mezi úrovněmi faktoru A, potom ve jmenovateli vzorce 9.3 bude výraz *bn*, pokud posuzujeme rozdíly mezi úrovněmi faktoru B, pak použijeme výraz *an*, kde *a* (resp. *b*) je počet úrovní faktoru A (resp. B) a *n* je počet opakování (měření, pozorování) v buňce.

Příklad 9.2:

V rámci akreditace zkušebny dřeva byla statisticky zkoumána správnost práce laborantů a příslušných měřících zařízení. V tabulce 9.11 jsou výsledky měření vlhkosti dřeva (v %) provedené 10x jednotlivými laboranty na všech měřících zařízeních. Rozhodněte, zdali je možné předpokládat, že výsledky dosažené na všech zařízeních a provedené všemi laboranty zaručují stejné výsledky.

Zadání tohoto příkladu patří mezi typické příklady použití dvoufaktorové analýzy rozptylu s opakováním. Pokud má nějaká laboratoř poskytovat věrohodné výsledky, musí zaručit, že ať měření provádí kdokoli na jakémkoli přístroji (se srovnatelnými parametry a přesností), budou výsledky stejné. Pouhé „okulární“ posouzení dosažených výsledků nestačí, výsledky musí být porovnány pomocí co nejobjektivnější metody. ANOVA je jedním z nejlepších statistických nástrojů takové kontroly.

V našem případě máme faktor A (zařízení) se čtyřmi úrovněmi ($a = 4$) a faktor B (laboranti) se třemi úrovněmi ($b = 3$). Počet opakování v buňce je $n = 10$.

Provedeme analýzu rozptylu podle schématu v tabulce 9.9 . Musíme také rozhodnout, zda do výpočtu zahrneme interakci. V tomto případě je zřejmé, že „přesnost“ přístrojů a
laborantů je nezávislá, že je možné použít model bez interakce. Pro srovnání budou vypočítány oba modely, aby bylo možné posoudit jejich rozdíly. Model s interakcí je v tabulce 9.12 a model bez interakce (který budeme považovat za konečný výsledek) v tabulce 9.13 .

Z tabulky 9.12 vyplývá, že interakce je opravdu zanedbatelná a můžeme použít jako výslednou tabulku 9.13 (bez interakce). Porovnáním prvních dvou řádků zjistíme, že statisticky významný vliv má faktor A, tj. jednotlivá zařízení. Znamená to, že alespoň jedno zařízení se svými výsledky významně odlišuje od ostatních. Naopak faktor B - tj. laboranti – nevykazují statisticky významné rozdíly a všichni měří na stejných zařízeních stejně přesně.

Analyzovaný model má typické pevné efekty (zajímají nás čtyři konkrétní zařízení a tři konkrétní laboranti v určité laboratoři, nejedná se o náhodný výběr z různých laboratoří), můžeme provést mnohonásobné porovnání, abychom zjistili, které (případně která) zařízení se vzájemně významně liší ve svých výsledcích. Použijeme modifikovanou Tukeyho metodu.

	Zařízení 1	Zařízení 2	Zařízení 3	Zařízení 4
Laborant 1	7.8	8.6	7.5	8.6
	7.8	6.9	9.4	8.7
	7.8	8.5	7.4	8.5
	8.0	7.6	8.9	8.1
	7.9	7.4	7.1	9.4
	8.1	7.6	7.3	10.0
	8.0	7.1	9.1	8.9
	7.8	6.7	7.2	9.5
	8.0	6.6	7.3	7.9
	8.0	8.0	6.9	9.0
Laborant 2	7.8	7.0	8.7	7.9
	8.2	8.2	8.5	8.9
	8.3	8.0	7.5	8.9
	8.0	7.1	8.4	7.0
	8.2	8.0	8.1	9.2
	8.3	7.0	7.2	9.6
	8.2	8.0	7.5	9.5
	8.3	8.4	6.6	8.9
	8.4	7.3	8.1	8.9
	7.8	9.0	7.7	9.6
Laborant 3	8.3	7.9	6.8	7.0
	8.3	8.2	8.0	6.4
	8.3	8.0	6.8	6.3
	8.2	8.0	7.1	8.9
	8.1	8.0	8.0	9.0
	8.1	8.0	8.4	9.0
	8.3	7.1	7.9	9.3
	8.2	7.2	6.5	9.8
	8.1	7.2	8.6	9.9
	8.1	8.3	9.1	10.0

Tabulka 9.11 – Zadání příkladu na dvoufaktorovou analýzu rozptylu s opakováním

Jako základ se použije vzorec 9.2, pomocí kterého vypočítáme testové kritérium *q*, přičemž vzorec 9.3 pro výpočet SE modifikujeme do podoby

$$SE = \sqrt{\frac{M_R}{bn}} = \sqrt{\frac{0{,}520}{3 \cdot 10}} = 0.132$$

Výsledky mnohonásobného porovnání jsou v tabulce 9.14 . Z jejích výsledků vyplývá, že statisticky významné rozdíly jsou ve dvojicích 2 – 4, 1 - 4 a 3 – 4, jinými slovy, že zařízení číslo čtyři dává významně odlišné výsledky, rozdíly mezi ostatními zařízeními jsou náhodné. Testová kritéria se porovnávají s kritickou hodnotou q rozdělení $q_{0,05;4;114} = 3.686$.

Zdroj variability	Součet čtverců odchylek	Počet stupňů volnosti	Průměrný čtverec odchylek (rozptyl)	Testové kritérium	Kritická hodnota	Hodnota P	Vliv zdroje variability je
faktor A (zařízení)	20.594	3	6.865	12.794	2.689	0.000	významný
faktor B (laborant)	0.363	2	0.182	0.338	3.080	0.714	nevýznamný
Interakce (laborant X zařízení)	1.295	6	0.216	0.402	2.184	0.876	nevýznamný
Celková variabilita vysvětlená faktory a interakcí	22.252	11					
Variabilita uvnitř buněk (reziduální)	57.948	108	0.537				
Celková variabilita pokusu	80.200	119					

Tabulka 9.12 – Výsledná tabulka analýzy rozptylu s interakcí

Zdroj variability	Součet čtverců odchylek	Počet stupňů volnosti	Průměrný čtverec odchylek (rozptyl)	Testové kritérium	Kritická hodnota	Hodnota P	Vliv zdroje variability je
faktor A (zařízení)	20.594	3	6.865	13.209	2.684	0.000	významný
faktor B (laborant)	0.363	2	0.182	0.349	3.076	0.706	nevýznamný
Celková variabilita vysvětlená faktory	20.957	5					
Variabilita uvnitř buněk (reziduální)	59.243	114	0.520				
Celková variabilita pokusu	80.200	119					

Tabulka 9.13 - Výsledná tabulka analýzy rozptylu bez interakce

Porovnání (čísla označují jednotlivá Zařízení)	Rozdíl průměrů	SE	Testové kritérium q	Kritická hodnota q	Výsledek porovnání (H_0 zamítáme/ /nezamítáme)
2 - 4	-1.057	0.132	8.026	3.686	**Zamítáme**
2 - 1	-0.393	0.132	2.988	3.686	Nezamítáme
2 - 3	-0.090	0.132	0.684	3.686	Nezamítáme
3 - 4	-0.967	0.132	7.342	3.686	**Zamítáme**
3 - 1	-0.303	0.132	2.304	3.686	Nezamítáme
1 - 4	-0.663	0.132	5.038	3.686	**Zamítáme**

Tabulka 9.14 - Výsledky mnohonásobného porovnání pomocí Tukeyho metody

Závěr můžeme formulovat tak, že všichni laboranti měří stejně (rozdíly jejich měření jsou náhodné), zařízení 4 poskytuje významně odlišné výsledky měření vlhkosti dřeva než ostatní zařízení.

9.2.3 Dvoufaktorová ANOVA s opakováním a nevyváženým modelem

Pokud není možné dodržet vyvážený model měření, což je velmi doporučeno (je zaručena nejvyšší síla testu), je možné pracovat i s nevyváženým modelem, ale s určitými omezeními. Nejlepší případ nevyváženého modelu je tzv. **proporční uspořádání**, kdy četnosti v jednotlivých buňkách odpovídají vzorci

$$n_{ij} = \frac{s \cdot r}{N} \tag{9.13}$$

kde je

s počet pozorování s hladinou 1. faktoru
r počet pozorování s hladinou 2. faktoru
N celkový počet všech pozorování

Příklad proporčního uspořádání je uveden v tabulce 9.15 . Např. pro buňku $x_{11} = 24 \cdot 9/72 = 3$.

	A_1	A_2	A_3	A_4	Σ
B_1	3	6	9	6	24
B_2	4	8	12	8	32
B_3	2	4	6	4	16
Σ	9	18	27	18	72

Tabulka 9.15 – Příklad proporčního uspořádání dat pro nevyvážený model analýzy rozptylu. Čísla v buňkách udávají počty hodnot.

Pokud je toto uspořádání dodrženo, je možné použít řešení podle tabulky 9.9 s úpravami vyplývajícími ze změněných počtů hodnot v buňkách. Upravené řešení je v tabulce 9.16 . Výpočty se provádí obdobně jako u vyváženého modelu.

Pokud chybí jedna nebo několik málo hodnot (obvykle ne víc než je počet úrovní) do vyváženého nebo proporčního modelu, je možné tyto hodnoty dopočítat. Podrobnosti jsou uvedeny např. v ZAR 1984.

Silně nevyvážené (neproporční) modely je možné spočítat jen pomocí přibližného rozkladu pomocí tzv. ekvivalentních četností (podrobnosti viz MELOUN-MILITKÝ 1994) nebo pomocí speciálních postupů regresní analýzy (viz ZAR 1984).

9.2.4 Dvoufaktorová ANOVA bez opakování měření

V některých případech se stává, že měření pro jednotlivé kombinace faktorů nelze replikovat, takže v každé buňce je jen jedno měření. I tento případ lze statisticky zpracovat, i když zde jsou – oproti metodě s vyváženým experimentem s opakováním – jistá omezení, především nemůžeme uvažovat s interakcí. Je tomu tak proto, že uvnitř buněk není žádná variabilita a tedy nemůžeme vypočítat M_R (tj. variabilitu uvnitř buněk). K odhadu celkové variability slouží pouze tzv. „zbytková“ variabilita (v an-

glicky psané literatuře označovaná jako „remainder“), která se získá jako rozdíl celkové variability a aditivně působících faktorů.

Uspořádání tabulky analýzy rozptylu pro dvoufaktorovou analýzu rozptylu bez opakování je uvedeno v tabulce 9.17 .

Zdroj variability	Součet čtverců odchylek	Počet stupňů volnosti	Průměrný čtverec odchylek (rozptyl)
Faktor A	$S_A = \sum_{i=1}^{a} \frac{\left(\sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} x_{ijk} \right)^2}{\sum_{j=1}^{b} n_{ij}} - C$	$DF_A = a - 1$	$M_A = \frac{S_A}{DF_A}$
Faktor B	$S_B = \sum_{j=1}^{b} \frac{\left(\sum_{i=1}^{a} \sum_{k=1}^{n_{ij}} x_{ijk} \right)^2}{\sum_{i=1}^{a} n_{ij}} - C$	$DF_B = b - 1$	$M_B = \frac{S_B}{DF_B}$
Interakce A x B	$S_{AB} = S_F - S_A - S_B$	$DF_{AB} = (a-1)(b-1)$	$M_{AB} = \frac{S_{AB}}{DF_{AB}}$
Variabilita vysvětlená faktory a interakcí	$S_F = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{\left(\sum_{k=1}^{n_{ij}} x_{ijk} \right)^2}{n_{ij}} - C$	$DF_F = ab - 1$	
Variabilita uvnitř buněk (reziduální)	$S_R = S_C - S_F$	$DF_R = ab(n-1)$	$M_R = \frac{S_R}{DF_R}$
Celková variabilita	$S_C = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} x_{ijk}^2 - C$	$DF_C = N - 1$	

kde je $C = \frac{\left(\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} x_{ijk} \right)^2}{N}$ a $N = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij}$

Tabulka 9.16 – Tabulka dvoufaktorové analýzy rozptylu s nevyváženým modelem a proporčním uspořádáním

Výpočet testových kritérií pro obě nulové hypotézy: (1) vliv faktoru A je nulový; (2) vliv faktoru B je nulový, je založen na vztazích

$$F = \frac{M_A}{M_E} \text{ nebo } \frac{M_B}{M_E} \tag{9.14}$$

Pokud model skutečně neobsahuje interakci, potom vztah 9.14 platí s obvyklými hodnotami chyby I. i II. druhu pro všechny typy modelů (s pevnými, náhodnými i smíšenými efekty). Pokud je možné, že model reálně interakci obsahuje (ale my ji nemůžeme spočítat), potom hodnoty ze vztahu 9.14 platí jen pro model s náhodnými efekty, v případě modelu s pevnými efekty a pro pevný efekt v modelu se smíšenými efekty vzrůstá pravděpodobnost chyby II. druhu – test je tedy „měkčí“, má zvýšenou schopnost nezamítnout nulovou hypotézu.

V případě pevných efektů můžeme provádět mnohonásobná porovnání obdobně jako u předchozích metod, např. pomocí Tukeyho testu, pouze upravíme výpočet chyby SE na tvar pro faktor A:

$$SE = \sqrt{\frac{M_E}{b}} \tag{9.15}$$

a obdobně pro faktor B dosadíme ve jmenovateli počet úrovní faktoru A – *a*. Kritická hodnota bude pro faktor A $q_{\alpha;a;(a-1)(b-1)}$ a pro faktor B $q_{\alpha;b;(a-1)(b-1)}$

Zdroj variability	Součet čtverců odchylek	Počet stupňů volnosti	Průměrný čtverec odchylek (rozptyl)
Faktor A	$S_A = \frac{\sum_{i=1}^{a}\left(\sum_{j=1}^{b} x_{ij}\right)^2}{b} - C$	$DF_A = a - 1$	$M_A = \frac{S_A}{DF_A}$
Faktor B	$S_B = \frac{\sum_{j=1}^{b}\left(\sum_{i=1}^{a} x_{ij}\right)^2}{a} - C$	$DF_B = b - 1$	$M_B = \frac{S_B}{DF_B}$
Variabilita nevysvětlená působením faktorů	$S_E = S_C - S_A - S_B$	$DF_E = (a-1)(b-1)$	$M_E = \frac{S_E}{DF_E}$
Celková variabilita	$S_C = \sum_{i=1}^{a}\sum_{j=1}^{b} x_{ij}^2 - C$	$DF_C = N - 1$	

$$C = \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b} x_{ij}\right)^2}{N} \qquad a \qquad N = ab$$

Tabulka 9.17 – Uspořádání tabulky pro dvoufaktorovou analýzu rozptylu bez opakování

Příklad 9.3:

V rámci biometrického výzkumu byl zkoumán vliv sociálního postavení stromů v porostu a světové strany na velikost tloušťkového přírůstu. V tabulce 9.18 *jsou hodnoty tloušťkového přírůstu za 10 let v mm. Rozhodněte, zda některý posuzovaný faktor má statisticky významný vliv.*

		Třída sociálního postavení stromu v porostu				
		1	2	3	4	5
Světová strana	S	36	28	21	20	18
	Z	35	28	22	21	20
	J	32	26	23	22	18
	V	35	27	22	22	19

Tabulka 9.18 – Zadání příkladu 9.3:

Cílem analýzy je prozkoumat, zda dva vybrané faktory – světová strana a třída sociálního postavení stromů v porostu – mají statisticky významný vliv na velikost tloušťkového přírůstu v posledních 10 letech.

Předpokládá se, že sociální postavení stromů v porostu bude mít vliv na velikost přírůstu (třída 1 – stromy nadúrovňové, výrazně vyšší než ostatní a obvykle nejsilnější; třída 2 – stromy úrovňové, obvykle nejpočetnější skupina tvořící „hlavní úroveň"; třída 3 – stromy do úrovně vrůstající; třída 4 – stromy podúrovňové; třída 5 – stromy odumírající), neboť velikost přírůstu souvisí s možnostmi výživy i přístupu světla.

Světová strana má vliv na ukládání tloušťkového přírůstu především tam, kde převládá určitý směr větru (souvisí s tvorbou tahového a tlakového dřeva) nebo působí jiné vlivy s tímto faktorem spojené.

V tomto případě se nepředpokládá významný vliv interakce.

Provedeme výpočet podle tabulky 9.17 , jejíž výsledky jsou uvedeny v tabulce 9.19 .

Zdroj variability	**Součet čtverců odchylek**	**Počet stupňů volnosti**	**Průměrný čtverec odchylek (rozptyl)**	**Testové kritérium**	**Kritická hodnota**	**Hodnota P**	**Vliv zdroje variability je**
Faktor A (třída sociálního postavení)	628.500	4	157.125	115.675	3.259	0.000	**významný**
Faktor B (světová strana)	2.950	3	0.983	0.724	3.490	0.557	nevýznamný
Variabilita nevysvětlená faktory	16.300	12	1.358				
Celkem	647.750	19					

Tabulka 9.19 – Výsledky dvoufaktorové analýzy rozptylu bez opakování

Porovnání (čísla označují jednotlivé třídy sociálního postavení)	Rozdíl průměrů	SE	Testové kritérium q	Kritická hodnota q	Výsledek porovnání (H_0 zamítáme/ /nezamítáme)
5 - 1	-15.750	0.583	27.031	4.508	**Zamítáme**
5 - 2	- 8.500	0.583	14.588	4.508	**Zamítáme**
5 - 3	- 3.250	0.583	5.578	4.508	**Zamítáme**
5 - 4	- 2.500	0.583	4.291	4.508	Nezamítáme
4 - 1	-13.250	0.583	22.740	4.508	**Zamítáme**
4 - 2	- 6.000	0.583	10.297	4.508	**Zamítáme**
4 - 3	- 0.750	0.583	1.287	4.508	Nezamítáme
3 - 1	-12.500	0.583	21.453	4.508	**Zamítáme**
3 - 2	- 5.250	0.583	9.010	4.508	**Zamítáme**
2 - 1	- 7.250	0.583	12.443	4.508	**Zamítáme**

Tabulka 9.20 – Výsledky mnohonásobného porovnání pomocí Tukeyho metody

Z výsledků vidíme, že statisticky významný vliv má faktor A (třída sociálního postavení), protože testové kritérium F (115,675) je větší než kritická hodnota $F_{0,05;4;12} = 3.259$. Naopak vliv světové strany nebyl potvrzen, tedy tloušťkový přírůst se u stromů ve studovaném území ukládá z hlediska postavení vůči světovým stranám rovnoměrně.

Jestliže považujeme třídy sociálního postavení za pevné faktory, můžeme provést Tukeyho metodu mnohonásobného porovnání, jejíž výsledky jsou v tabulce 9.20 . Vidíme, že třída 1 a třída 2 tvoří samostatné skupiny (liší se významně od všech ostatních), nevýznamné rozdíly jsou mezi třídou 4 a 3 a také těsně mezi třídami 4 a 5 (toto zamítnutí nulové hypotézy je zřejmě „výsledkem“ chyby II. druhu, protože není možné, aby třída 4 byla nevýznamně odlišná jak od třídy 3 tak od třídy 5, přičemž třídy 5 a 3 se od sebe významně liší). Je tedy možné uzavřít, že třídy 1, 2 a 5 tvoří samostatné skupiny, které se liší od všech ostatních, třídy 4 a 3 tvoří další homogenní skupinu.

9.2.5 Využití analýzy rozptylu v plánování pokusů

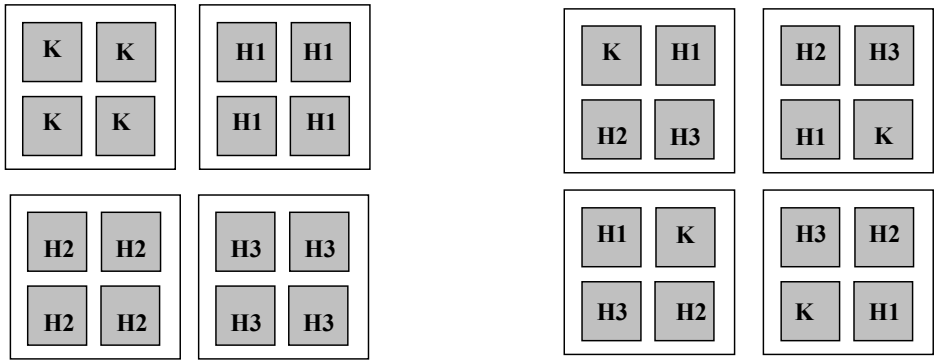
9.2.5.1 Uspořádání základních pokusných plánů

Jestliže zakládáme reálné pokusy, v mnoha případech se musíme vyrovnat se skutečností, že není možné zaručit naprosto shodné podmínky na různých pokusných plochách nebo pro různé pokusné jedince. Tyto případy jsou velmi časté v zemědělství, lesnictví, ekologii, biologii, lékařství apod.

Typickým příkladem jsou terénní pokusy. Většina biotických i abiotických podmínek prostředí se kontinuálně mění podle toho, kde založíme pokusné plochy – obvykle platí, že blízké plochy jsou si podobnější než plochy vzdálené. Jestliže např. chceme sledovat vliv hnojení na růst semenáčků v lesní školce, je možné založit pokus, kde se určité záhony budou hnojit určitým přípravkem. Abychom však vyloučili vliv všech ostatních faktorů kromě hnojiva, musíme zabezpečit, aby všechny další faktory působící na růst byly identické. Což je v praktických podmínkách, např. lesní školky, velmi obtížné. Každý záhon se poněkud liší např. půdou, zastíněním, obsahem vody apod. Proto je nutné pokusné záhony rozmístit tak, aby se na každém záhonu aplikovala všechna hnojiva, a to v náhodném pořadí. Cílem tohoto uspořádání **je minimalizovat vliv heterogenních podmínek** na jednotlivých plochách při zachování základní podmínky statistické analýzy – **nezávislosti opakování**. Takovým uspořádáním pokusů, které splnění tohoto cíle umožní (a také následným vyhodnocením) se zabývá speciální odvětví statistiky – **plánování pokusů**.

Metodika plánování pokusů je značně rozsáhlá a zahrnuje velký počet různých typů uspořádání pokusů a způsobů jejich vyhodnocení. Zde se pouze dotkneme nejpoužívanějších metod. Kromě nich se používá celá řada speciálních technik plánování pokusů, jejichž naplánování i vyhodnocení je velmi technicky náročné a je umožněno specializovanými statistickými programy, např. neúplné náhodné bloky, split-plot techniky, řecko-latinské čtverce a další.

Obrázek 9.8 ukazuje rozdíl mezi chybným založením pokusu (vlevo) a správným (vpravo) pomocí tzv. **úplných znáhodněných bloků**. Zkoumáme vliv tří druhů hnojiva na růst (skupiny H1 – H3) a porovnáváme je s kontrolní skupinou (K) bez hnojení. Vlevo každý záhon obsahuje všechna opakování jedné skupiny – to je chybné založení pokusu, protože naprosto nezohledňuje rozdíly růstových podmínek po ploše školky a např. kontrolní skupina a skupina hnojená hnojivem H3 mohou mít značně rozdílné podmínky. Správné uspořádání ukazuje pravá část obrázku 9.8 , kde je jasně vidět, že všechna opakování jednotlivých skupin jsou rovnoměrně rozložena po celé ploše pokusného pozemku. Jednotlivé druhy ošetření (např. hnojiva) se v bloku rozdělení náhodně, např. pomocí losování, tabulek náhodných čísel nebo pomocí schémat uváděných v odborné literatuře. Důležitou podmínkou je maximální homogenita podmínek každého bloku. Toto uspořádání zohledňuje vliv různých růstových podmínek a dává maximálně objektivní podmínky pro vyhodnocení toho vlivu, který nás zajímá – v tomto případě druhu hnojiva.



Obrázek 9.8 – Schéma nesprávného (vlevo) a správného (vpravo) uspořádání znáhodněných bloků

K	H1	H2	H3
K	H1	H2	H3
K	H1	H2	H3
K	H1	H2	H3

K	H1	H3	H2
H1	K	H2	H3
H2	H3	H1	K
H3	H2	K	H1

Obrázek 9.9 – Vlevo je nesprávné uspořádání pokusu, vpravo správné uspořádání - jedna z možných variant latinského čtverce

Jiným způsobem uspořádání jsou tzv. **latinské čtverce**, které obsahují v každém řádku a každém sloupci jedno opakování každé skupiny. Používají se především tehdy, jestliže ani v rámci bloku není možné dodržet dostatečnou homogenitu podmínek a je nezbytné heterogenitu eliminovat. Toto uspořádání je také ekonomické, protože vyžaduje minimální počet pokusných plošek a různorodost podmínek se hodnotí ve dvou na sebe kolmých směrech. Umožňuje vyloučit vliv existujících rozdílů v podmínkách pokusu mezi jednotlivými řádky a sloupci, protože celková hodnota výsledku pokusu pro jednotlivá ošetření (např. druh hnojiva, různé kultivary, různé druhy výchovných zásahů, apod.) je daná součtem hodnot z pokusných jednotek (plošek, políček, zkusných ploch, pokusných jedinců, apod.), které jsou umístěné vždy jiném řádku a v jiném sloupci. Možné uspořádání latinského čtverce ukazuje obrázek 9.9 . Možností uspořádání latinských čtverců je hodně, jejich počet se stanoví podle vztahu $n!(n-1)!$, kde n je počet sloupců a řádků, tedy např. pro $n = 3$ je to 12 čtverců, pro $n = 4$ už 576, pro $n = 5$ je to 161 280 kombinací atd.

9.2.5.2 Vyhodnocení základních pokusných plánů

Vyhodnocení znáhodněných bloků i latinských čtverců se provádí pomocí analýzy rozptylu.

Znáhodněné bloky jsou vlastně zobecněním párového t-testu pro více skupin než dvě (stejně jako je jednofaktorová ANOVA zobecněním t-testu pro nezávislé výběry). Výpočet se provádí v podstatě podle schématu dvoufaktorové analýzy rozptylu bez opakování (tabulka 9.17), kde jeden faktor (pevný) jsou pokusné zásahy (např. hnojení), druhý faktor (náhodný) je zařazení do bloku. Jedná se tedy vlastně o Model III (smíšené faktory) dvoufaktorové analýzy rozptylu bez opakování. Používají se i stejné metody mnohonásobného porovnání (např. Tukeyho pro porovnání skupin mezi sebou nebo Dunnettova pro srovnání s kontrolou).

Příklad 9.4:

Při výzkumu účinku nových druhů hnojiv byl založen pokus metodou znáhodněných bloků pro porovnání účinků tří druhů hnojiva (H1, H2 a H3). Jedna skupina byla ponechána bez hnojení jako kontrolní (K). Posuďte, zda hnojiva zlepšují růst sadebního materiálu v lesní školce oproti kontrolní skupině a zda se liší mezi sebou. Měřenou veličinou v tabulce 9.21 je průměrná výška sazenic na každé dílčí plošce v cm.

Tzv. „polní“ pokusy jsou typickým příkladem použití znáhodněných bloků. Obvykle porovnávají vliv nějakého opatření (hnojení, výchovného zásahu, způsobu ošetření nebo také kultivaru, provenience, apod.) na produkci (nebo jinou měřitelnou vlastnost) daného pokusného materiálu.

Blok 1	Výška	Blok 2	Výška	Blok 3	Výška	Blok 4	Výška	Blok 5	Výška
K	21.6	H1	24.0	H2	26.0	H1	23.9	H3	23.3
H1	24.1	H3	29.4	K	22.1	H3	21.6	H2	26.6
H2	26.3	K	19.4	H3	23.1	H2	24.5	K	19.8
H3	25.8	H2	28.5	H1	27.5	K	17.9	H1	24.3

Tabulka 9.21 – Blokové uspořádání zadání příkladu 9.4

Pro analýzu rozptylu musíme data z tabulky 9.21 uspořádat do tvaru vhodného pro výpočet (viz tabulku 9.22), čímž vznikne tabulka dvoufaktorové analýzy rozptylu bez opakování. Zde už jsou samozřejmě hodnoty seřazeny podle bloků a způsobů ošetření.

Blok	Ošetření (hnojivo)			
	K	H1	H2	H3
1	21.6	24.1	26.3	25.8
2	19.4	24.0	28.5	29.4
3	22.1	27.5	26.0	23.1
4	17.9	23.9	24.5	21.6
5	19.8	24.3	26.6	23.3

Tabulka 9.22 – Zadání příkladu 9.4 ve tvaru vhodném pro analýzu rozptylu

K výpočtu použijeme schéma tabulky 9.17 a výsledek je v tabulce 9.23 . Porovnáním kritických hodnot a testových kritérií zjistíme, že faktor „způsob ošetření – druh hnojiva“ má statisticky významný vliv na výšku sazenic, zatímco druhý faktor – bloky – nikoliv. Znamená to, že mezi bloky nebyly zásadní rozdíly v růstových podmínkách, což dokazuje splnění základní podmínky použití znáhodněných bloků – homogenitu podmínek v rámci bloku. To umožní objektivní posouzení vlivu faktoru „druh hnojiva“. Máme za úkol posoudit jednak významnost vlivu použití hnojiva vůči kontrole, jednak druhy hnojiva mezi sebou. První úkol vyřešíme Dunnettovým testem, druhý „klasickým“ Tukeyovým testem. Výsledky Dunnettova testu jsou v tabulce 9.24 a Tukeyova testu v tabulce 9.25 .

V tabulky 9.24 vyplývá, že účinky všech hnojiv významně zvyšují výškový růst, protože ve všech případech byla zamítnuta nulová hypotéza o nevýznamném vlivu jednotlivých hnojiv. Byla použita jednostranná hypotéza $\mathbf{H_0}$: $\mu_A = \mu_{kontrola}$ oproti $\mathbf{H_1}$: $\mu_A \geq \mu_{kontrola}$, protože všechny průměry hnojených skupin byly vyšší. Kritické hodnoty byly stanoveny jako $q^*_{0,05;N-k;p}$, kde N-k = 16 a p „vzdálenost“ porovnávaných skupin (pro *K - H2* se p = 4, pro další srovnávanou dvojici *K - H1* se p = 3 a pro *K – H3* se p = 2). Můžeme tedy tuto část uzavřít tvrzením, že všechna hnojiva statisticky významně zlepšují výškový růst sazenic.

Tabulka 9.25 udává výsledky Tukeyova testu pro porovnání všech skupin mezi sebou. Vidíme zde potvrzení výsledku Dunnettova testu (všechna hnojiva se statisticky významně liší od kontroly) a navíc zde máme i porovnání všech hnojiv mezi sebou. Výsledky ukazují, že se jednotlivá hnojiva mezi sebou neliší (veškeré rozdíly mezi H1, H2 a H3 jsou nevýznamné). Pokud tedy použijeme jakékoli hnojivo, zlepšení růstu bude prakticky stejné.

Zdroj variability	Součet čtverců odchylek	Počet stupňů volnosti	Průměrný čtverec odchylek (rozptyl)	Testové kritérium	Kritická hodnota	Hodnota P	Vliv zdroje variability je
Ošetření (hnojivo)	106.981	3	35.660	10.776	3.490	0.001	**významný**
Blok	27.053	4	6.763	2.044	3.259	0.152	nevýznamný
Variabilita nevysvětlená ošetřením a blokem	39.711	12	3.309				
Celkem	173.745	19					

Tabulka 9.23 – Výsledky analýzy rozptylu pro zadání příkladu 9.4

Porovnání (písmena označují jednotlivé druhy hnojiva a kontrolu (K))	Rozdíl průměrů	SE	Testové kritérium q	Kritická hodnota q	Výsledek porovnání (H_0 zamítáme/ /nezamítáme)
K - H2	-6.22	1.150	5.406	2.230	**Zamítáme**
K - H1	-4.60	1.150	3.998	2.060	**Zamítáme**
K - H3	-4.48	1.150	3.894	1.750	**Zamítáme**

Tabulka 9.24 – Výsledky Dunnettova testu pro zadání příkladu 9.4

Porovnání (písmena označují jednotlivé druhy hnojiva a kontrolu (K))	Rozdíl průměrů	SE	Testové kritérium q	Kritická hodnota q	Výsledek porovnání (H_0 zamítáme/ /nezamítáme)
K - H2	-6.22	0.814	7.646	4.046	**Zamítáme**
K - H1	-4.60	0.814	5.655	4.046	**Zamítáme**
K - H3	-4.48	0.814	5.507	4.046	**Zamítáme**
H3 - H2	-1.74	0.814	2.139	4.046	Nezamítáme
H3 - H1	-0.12	0.814	0.148	4.046	Nezamítáme
H1 - H2	-1.62	0.814	1.991	4.046	Nezamítáme

Tabulka 9.25 - Výsledky Tukeyho testu pro zadání příkladu 9.4

Vyhodnocení latinských čtverců je poněkud komplikovanější, protože se zde uvažuje vlastně se třemi faktory – postavením „pokusné plochy“, tj. políčka tabulky na obrázku 9.9 , které je dané pozicí řádku (první faktor) a sloupce (druhý faktor) a dále typem ošetření (hnojivo, kultivar, …), což je třetí faktor. K řešení tedy potřebujeme třífaktorovou analýzu rozptylu bez interakce, což již přesahuje rozsah tohoto textu. Zájemci najdou podrobnosti o plánování experimentů (i daleko složitějších než zde

popsané základní typy) včetně podrobně řešených příkladů v rozsáhlé literatuře k tomuto tématu, z našich např. MYSLIVEC 1957, GROFÍK 1987, ze zahraničních např. MONTGOMERY 1991, MEAD 1988 a mnohé další.

9.3 Neparametrická ANOVA

Neparametrická ANOVA se používá především tehdy, jsou-li výrazně narušeny základní předpoklady pro provedení parametrické analýzy rozptylu, tedy především normalita a homogenita rozptylu. Nutno podotknout, že parametrická ANOVA sama je vůči narušení předpokladů poměrně robustní („odolná“) a neparametrické metody používáme zpravidla při výrazném porušení předpokladů a také tehdy, mají-li jednotlivé výběry velmi málo prvků (nebo jsou jejich počty silně nevyvážené) a normalitu není možné spolehlivě stanovit. Při použití neparametrické analýzy rozptylu musíme, tak jako i u jiných neparametrických testů, počítat s nižší silou testu (a tedy slabší schopností zamítnout nulovou hypotézu). Uvádí se (LEPŠ 1996), že při splnění předpokladů normality a homogenity má neparametrická ANOVA asi 95 % síly testu parametrické analýzy rozptylu. V tomto případě obvykle použijeme běžnou parametrickou analýzu rozptylu. Ovšem v případě, že předpoklady pro parametrickou analýzu rozptylu jsou výrazně narušeny, je neparametrická ANOVA silnějším testem než parametrická.

9.3.1 Kruskal-Wallisův test (K-W test)

Tento test je neparametrická obdoba jednofaktorové analýzy rozptylu, podobně jako je Mann-Whitneyův (Wilcoxonův) test neparametrickou obdobou t-testu.

K-W test je založen, tak jako většina neparametrických testů, na pořadí prvků. Postup provedení testu je následující:

- prvky všech výběrů (skupin) sloučíme do jednoho sdruženého výběru (musíme zachovat informaci o tom, ze kterého výběru který prvek pochází);
- prvky sdruženého výběru seřadíme podle velikosti od nejmenšího k nejvyššímu;
- takto seřazené prvky očíslujeme podle pořadí (nejmenší prvek dostane číslo 1, druhý nejmenší 2, atd), přičemž prvky stejné hodnoty obdrží průměrné pořadí těchto prvků;
- dále již **pracujeme pouze s pořadím** – pořadí jednotlivých prvků rozdělíme znovu do původních výběrů (skupin);
- v jednotlivých skupinách pořadí prvků sečteme – získáme hodnoty $R_{i;}$
- vypočítáme testové kritérium

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \qquad (9.16)$$

kde je

N celkový počet všech prvků ve všech výběrech dohromady

n_i počet prvků v *i-tém* výběru
R_i součet pořadí v *i-tém* výběru;

- pokud jsou mezi prvky skupiny stejných hodnot (a tedy stejných pořadí), opravíme kritérium H podle vztahu

$$H_C = \frac{H}{1 - \frac{\sum_{i=1}^{m}\left(t_i^3 - t_i\right)}{N^3 - N}} \tag{9.17}$$

kde je
t_i počet stejných hodnot v *i-té* skupině stejných hodnot
m počet skupin stejných hodnot

- testové kritérium H (nebo H_C, pokud porovnáváme výběry, kde jsou skupiny stejných hodnot) porovnáme s kritickou hodnotou H (resp. χ^2). Pokud je testové kritérium menší než kritická hodnota, nezamítáme nulovou hypotézu o rovnosti průměrů.

Kritická hodnota pro K-W test je dvojí:

- pro malé výběry (do $n_i \leq 8$ pro 3 výběry, $n_i \leq 4$ pro 4 výběry a do $n_i \leq 3$ pro 5 výběrů) je to tabelované speciální kritérium H (**tabulka 3** v příloze),
- pro větší výběry a pro větší počet výběrů než 5 je to statistika χ^2 pro *k-1* stupňů volnosti.

Pokud je nulová hypotéza zamítnuta, je možné stejně jako u parametrické analýzy rozptylu zjistit, mezi kterými skupinami (výběry) existují statisticky významné rozdíly.

a) pro stejné počty prvků ve všech skupinách

V tomto případě používáme test založený na Tukeyho testu (viz kapitolu 9.1.2.1). Vycházíme z obdoby testového kritéria pro Tukeyho test (rovnice 9.2), ale místo průměrů použijeme součty pořadí R_i (tedy pro srovnání skupin *A* a *B* hodnoty R_A a R_B)

$$q = \frac{R_A - R_B}{SE} \tag{9.18}$$

kde je

$$SE = \sqrt{\frac{n(nk)(nk+1)}{12}} \tag{9.19}$$

Testové kritérium *q* porovnáme s kritickou hodnotou $q_{\alpha;\infty;k}$, kde *k* je počet všech porovnávaných skupin v celém K-W testu.

b) pro nestejně velké skupiny (počty prvků v jednotlivých skupinách se liší)

V tomto případě použijeme **Dunnův test**, který je založený na testovém kritériu

$$Q = \frac{\overline{R}_B - \overline{R}_A}{SE} \tag{9.20}$$

kde je

$\overline{R}$ průměrné pořadí v porovnávaných skupinách (tedy $\overline{R}_A = R_A/n_A$ a $\overline{R}_B = R_B/n_B$)

SE se vypočítá podle vztahu

$$SE = \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} \tag{9.21}$$

a pokud výběry obsahují skupiny stejných hodnot, potom se SE vypočítá podle upraveného vztahu

$$SE = \sqrt{\left(\frac{N(N+1)}{12} - \frac{\sum_{i=1}^{m}\left(t_i^3 - t_i\right)}{12(N-1)}\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} \tag{9.22}$$

kde je symbolika stejná jako u vztahu 9.17.

Testové kritérium Q se porovná s kritickou hodnotou $Q_{\alpha;k}$, která je tabelována ve speciálních tabulkách (v příloze **Tabulka 5**).

Příklad 9.5:

V rámci limnologického výzkumu byla ve čtyřech vodních nádržích měřena hodnota pH. Z každé nádrže bylo odebráno 8 vzorků (jeden vzorek z nádrže číslo 3 byl znehodnocen) a zjištěné hodnoty pH jsou v tabulce 9.26 . Posuďte, zda se hodnoty pH v jednotlivých nádržích liší.

Vstupní hodnoty měření z tabulky 9.26 musíme upravit do podoby vhodné k výpočtu – všechny čtyři výběry spojit dohromady, seřadit podle velikosti a jednotlivým měřením přiřadit jejich pořadí. Výsledky tohoto postupu jsou v tabulce 9.27 .

Tabulka 9.26 – Zadání příkladu 9.5 – hodnoty pH ze čtyř nádrží

Nádrž 1	Nádrž 2	Nádrž 3	Nádrž 4
7.73	7.80	7.84	7.87
7.69	7.73	7.75	7.71
7.68	7.71	7.74	7.71
7.76	7.81	7.77	7.91
7.70	7.74	7.80	7.74
7.72	7.78	7.78	7.81
7.70	7.74	7.81	7.79
7.73	7.78		7.85

V případě skupin stejných hodnot je nutno dát pozor na správné přiřazení pořadí. Např. se ve sdruženém souboru vyskytují čtyři hodnoty 7.74. Podle seřazených hodnot by měly „teoreticky“ dostat pořadí 12, 13, 14 a 15. Vypočítáme průměr těchto pořadí ((12+13+14+15)/4=13.5) a toto pořadí se přiřadí všem hodnotám 7.74. Stejně se postupuje i s ostatními skupinami stejných hodnot (v příkladu je jich celkem 7). Poté se jednotlivá pořadí (od této chvíle již nepracujeme s měřenými hod-

notami, pouze s **pořadími s korekcí!!)** rozdělí do původních výběrů, což je uvedeno v tabulce 9.28 . Dole v této tabulce jsou vedeny součty pořadí R_i a četnosti jednotlivých výběrů n_i. Podle vzorce 9.16 vypočítáme kritérium H a vzhledem k tomu, že ve výběrech jsou skupiny stejných hodnot, musíme provést korekci H_C podle vztahu 9.17.

$$H = \frac{12}{31 \cdot 32}\left[\frac{55^2}{8} + \frac{132.5^2}{8} + \frac{145^2}{7} + \frac{163.5^2}{8}\right] - 3 \cdot 32 = 11.876$$

$$H_C = \frac{11.876}{1 - \frac{(2^3 - 2) + (3^3 - 3) + (3^3 - 3) + (4^3 - 4) + (3^3 - 3) + (2^3 - 2) + (3^3 - 3)}{31^3 - 31}} = 11.943$$

Z výpočtu je vidět, že korekce H_C je významná pouze při velkém počtu rozsáhlých skupin stejných hodnot, jinak je její vliv zanedbatelný. Výsledné testové kritérium porovnáme s kritickou hodnotou $\chi^2_{0.05;3} = 7.815$. Výsledkem je zamítnutí nulové hypotézy a přijímáme závěr, že hodnoty pH se v jednotlivých nádržích liší.

Měřená hodnota	Pořadí bez korekce	Pořadí s korekcí na skupiny stejných hodnot	Měřená hodnota	Pořadí bez korekce	Pořadí s korekcí na skupiny stejných hodnot	Měřená hodnota	Pořadí bez korekce	Pořadí s korekcí na skupiny stejných hodnot
7.68	1	1	7.74	***12***	13.5	7.80	***23***	23.5
7.69	2	2	7.74	***13***	13.5	7.80	***24***	23.5
7.70	***3***	3.5	7.74	***14***	13.5	7.81	***25***	26
7.70	***4***	3.5	7.74	***15***	13.5	7.81	***26***	26
7.71	***5***	6	7.75	16	16	7.81	***27***	26
7.71	***6***	6	7.76	17	17	7.84	28	28
7.71	***7***	6	7.77	18	18	7.85	29	29
7.72	8	8	7.78	***19***	20	7.87	30	30
7.73	***9***	10	7.78	***20***	20	7.91	31	31
7.73	***10***	10	7.78	***21***	20			
7.73	***11***	10	7.79	22	22			

Tabulka 9.27 – Vzestupně uspořádané hodnoty příkladu 9.5. Skupiny stejných hodnot jsou vyznačeny šedě a jejich pořadí bez korekce tučnou kurzívou. Výsledná pořadí, se kterými se bude dále pracovat, jsou v pravém sloupci (pořadí s korekcí)

Vzhledem k tomu, že nemáme stejný počet prvků ve všech výběrech, musíme jako metodu mnohonásobného porovnání použít Dunnův test. Použijeme vztahy 9.20 a 9.22 (musíme provést korekci na skupiny hodnot) s výsledky, které jsou uvedeny v tabulce 9.29 . Z těchto výsledků vyplývá, že statisticky významné rozdíly jsou mezi hodnotami pH v nádrži 1 oproti nádržím 3 a 4. Ostatní rozdíly se nejeví jako významné. Nevýznamný výsledek týkající se srovnání mezi nádržemi 1 a 2 je zřejmě důsledkem chyby II. druhu (i zde mají testy mnohonásobného porovnání menší sílu než samotný Kruskal-Wallisův test). Můžeme tedy uzavřít, že hodnota pH se odlišuje významně v nádrži 1, v ostatních je stejná.

Nádrž 1		Nádrž 2		Nádrž 3		Nádrž 4	
měřená hodnota	pořadí	měřená hodnota	pořadí	měřená hodnota	pořadí	měřená hodnota	pořadí
7.68	1	7.71	6	7.74	13.5	7.71	6
7.69	2	7.73	10	7.75	16	7.71	6
7.70	3.5	7.74	13.5	7.77	18	7.74	13.5
7.70	3.5	7.74	13.5	7.78	20	7.79	22
7.72	8	7.78	20	7.80	23.5	7.81	26
7.73	10	7.78	20	7.81	26	7.85	29
7.73	10	7.80	23.5	7.84	28	7.87	30
7.76	17	7.81	26			7.91	31
n_1	8	n_2	8	n_3	7	n_4	8
R_1	55	R_2	132.5	R_3	145	R_4	163.5

Tabulka 9.28 – Pořadí v rámci jednotlivých výběrů (dole jsou uvedeny součty pořadí R_i)

Tabulka 9.29 – Výsledky Dunnovy metody mnohonásobného porovnání

Porovnání mezi výběry	Rozdíl průměrných pořadí	SE	Testové kritérium Q	Kritická hodnota Q	Výsledek porovnání (H_0 zamítáme/ /nezamítáme)
Nádrž 1 - Nádrž 3	-13.839	4.692	2.949	2.639	Zamítáme
Nádrž 1 - Nádrž 4	-13.563	4.533	2.992	2.639	Zamítáme
Nádrž 1 - Nádrž 2	- 9.688	4.533	2.137	2.639	Nezamítáme
Nádrž 2 - Nádrž 3	- 4.152	4.692	0.885	2.639	Nezamítáme
Nádrž 2 - Nádrž 4	- 3.875	4.533	0.855	2.639	Nezamítáme
Nádrž 4 - Nádrž 3	- 0.277	4.692	0.059	2.639	Nezamítáme

9.3.2 Dvoufaktorová neparametrická ANOVA

Dvoufaktorová neparametrická ANOVA pro klasické modely s opakováním nebo bez opakování není příliš častá. Technicky je to vlastně rozšíření a úprava Kruskal-Wallisova testu. Podstata spočívá v tom, že se stejně jako u K-W testu všechny hodnoty nahradí pořadím. Poté se pro každou buňku spočítají sumy pořadí, stejně jako pro řádky a pro sloupce (tedy pro oba faktory). Samotné součty čtverců odchylek pořadí a následné výpočty se pak provádějí v podstatě stejně jako v případě parametrické analýzy rozptylu (podle schémat v tabulkách 9.9 a 9.17) s určitými korekcemi, pokud se vyskytují skupiny stejných hodnot. Podrobný postup včetně příkladu uvádí např. ZAR (1984).

Častější je užití neparametrického testu v případě znáhodněných bloků – zde se tento postup nazývá **Friedmanův test**. Používá se především v případě, když *k* úrovní pevného faktoru (tj. toho, jehož vliv na měřenou veličinu zkoumáme, druhým faktorem pak jsou bloky) nepochází z normálního rozdělení a předpoklad normality je silně narušen. Je nutné si znovu uvědomit, že i Friedmanův test má menší sílu testu než parametrická metoda (např. pro k = 2 je to asi 64 % síly parametického testu, pro k = 3 je to asi 73 % a se vzrůstajícím počtem výběrů síla testu stoupá až na 95 % pro velký – teoreticky nekonečný - počet výběrů). Proto se používá jen tehdy, je-li jeho použití nutné, ale v těchto případech je obvykle silnější než parametrický test.

Předpokládáme *a* úrovní pevného faktoru a *b* bloků. Jako obvykle, měřená data nahradíme pořadím, ale jinak než u K-W testu. V rámci každého bloku seřadíme hodnoty podle velikosti a přiřadíme jim pořadí. Potom sečteme hodnoty pořadí pro každý pevný faktor a získáme *a* hodnot, obvykle označovaných R_i. Testové kritérium se stanoví

$$\chi_r^2 = \frac{12}{ba(a+1)} \sum_{i=1}^{a} R_i^2 - 3b(a+1) \qquad (9.23)$$

kde je

a počet úrovní pevného (zkoumaného) faktoru
b počet bloků
Ri součet pořadí pro *i-tou* úroveň pevného faktoru

Testové kritérium porovnáme s kritickou hodnotou $\chi^2_{\alpha;a-1}$. Pro některé kombinace *a* a *b* existují speciální hodnoty Friedmanova rozdělení (zvláště pro a = 3 – viz **Tabulka 4** v příloze). Pokud je testové kritérium vyšší, zamítáme nulovou hypotézu o nevýznamném vlivu studovaného (pevného) faktoru.

Pokud se vyskytnou skupiny stejných pořadí, použijeme vzorec

$$\left(\chi_r^2\right)_c = \frac{\sum_{i=1}^{a} R_i^2 - \frac{\left(\sum_{i=1}^{a} R_i\right)^2}{a}}{\frac{ba(a+1)}{12} - \frac{\sum T}{a-1}} \tag{9.24}$$

kde je

$$\sum T = \frac{\sum_{i=1}^{m}\left(t_i^3 - t_i\right)}{12} \tag{9.25}$$

Jestliže je nulová hypotéza zamítnuta, můžeme zjistit, mezi kterými úrovněmi pevného faktoru existuje statisticky významný rozdíl. Použijeme modifikaci Tukeyho nebo Dunnettova (pro srovnání s kontrolní skupinou) testu. Testové kritérium je v případě Tukeyho testu

$$q = \frac{R_A - R_B}{SE} \tag{9.26}$$

kde je

$$SE = \sqrt{\frac{ba(a+1)}{12}} \tag{9.27}$$

a testové kritérium q se porovnává s kritickou hodnotou studentizovaného rozpětí $q_{\alpha;\infty;k}$.

Pro případ Dunnettova testu (porovnání s kontrolou – viz kapitola 9.1.2.3) se SE vypočítá

$$SE = \sqrt{\frac{ba(a+1)}{6}} \tag{9.28}$$

Příklad 9.6:

Použijte Friedmanův test pro zadání příkladu 9.4. Měřené hodnoty jsou v tabulkách 9.21 a 9.22 .

V tomto příkladu znovu posoudíme vliv tří druhů hnojiva na výškový růst semenáčků, tentokrát pomocí neparametrického testu. Výsledky potom porovnáme se závěry příkladu 9.4.

Blok	Ošetření (hnojivo)							
	K		H1		H2		H3	
	hodnota	pořadí	hodnota	pořadí	hodnota	pořadí	hodnota	pořadí
1	21.6	1	24.1	2	26.3	4	25.8	3
2	19.4	1	24.0	2	28.5	3	29.4	4
3	22.1	1	27.5	4	26.0	3	23.1	2
4	17.9	1	23.9	3	24.5	4	21.6	2
5	19.8	1	24.3	3	26.6	4	23.3	2
Součet		5		14		18		13

Tabulka 9.30 – Pořadí pro Friedmanův test

Pro Friedmanův test musíme měřené hodnoty nahradit pořadím. Vycházíme z tabulky 9.22 , kde jsou měřené hodnoty uspořádány podle úrovní pevného faktoru (druh hnojiva) a podle bloků. Každé hodnotě v každém řádku (bloku) přidělíme pořadí. Například v 1. bloku – 1. řádku tabulky – je nejmenší hodnota 21.6 (úroveň faktoru „hnojivo“ K), tedy dostane pořadí 1, následují hodnoty 24.1 úrovně H1 (pořadí 2), 25.8 úrovně H3 (pořadí 3) a 26.3 úrovně H2 (pořadí 4). Potom sečteme pořadí ve sloupcích (pro jednotlivé úrovně faktoru „hnojivo“) a získáme tak hodnoty R_i. Tento postup je zřejmý z tabulky 9.30 .

Poté vypočítáme testové kritérium

$$\chi_r^2 = \frac{12}{5 \cdot 4 \cdot 5} \cdot \left(5^2 + 14^2 + 18^2 + 13^2\right) - 3 \cdot 5 \cdot 5 = 10.68$$

Kritická hodnota je podle speciálních tabulek Friedmanova rozdělení 7.8, podle hodnoty $\chi^2_{0,05;3} = 7.815$, tedy testové kritérium je vyšší než kritická hodnota, nulovou hypotézu proto zamítáme – použitá hnojiva mají statisticky významný vliv na měřenou veličinu.

Pro metodu mnohonásobného porovnání použijeme Dunnettův test (porovnáváme s kontrolou), jehož výsledky jsou v tabulce 9.31 .

Porovnání mezi výběrem bez hnojení (K) a výběry hnojenými (H1 - H3)	Rozdíl součtů pořadí	SE	Testové kritérium q	Kritická hodnota q	Výsledek porovnání (H_0 zamítáme/ /nezamítáme)
K - H2	-13.000	4.082	3.184	2.060	**Zamítáme**
K - H1	-9.000	4.082	2.205	2.060	**Zamítáme**
K - H3	-8.000	4.082	1.960	2.060	Nezamítáme

Tabulka 9.31 – Výsledky Dunnettova testu

Dunnettův test prokázal statisticky významné rozdíly mezi kontrolou a výběry H1 a H2, neprokázal významný rozdíl mezi kontrolou a výběrem H3. Oproti parametrické analýze rozptylu se zde projevila menší síla neparametrického testu, protože v příkladu 9.4 byl prokázán významný rozdíl mezi kontrolou a všemi výběry.

10 Korelační a regresní analýza

V předcházejících kapitolách jsme zkoumali jednotlivé jevy (statistické znaky) izolovaně – zabývali jsme se tzv. **jednorozměrnými soubory,** tj. soubory popisujícími pouze jeden statistický znak a nezajímaly nás jeho vazby a vztahy k jiným jevům. V reálném světě (v přírodě, společnosti, ekonomice,…) se ovšem jevy nacházejí ve více nebo méně složitých vzájemných vztazích – navzájem na sobě závisí a podmiňují se. Proto se statistická analýza nemůže omezit pouze na zkoumání izolovaných jevů, ale musí se také zabývat analýzou jejich vzájemných vztahů.

Vztahy mezi jevy se možné zkoumat jak pro znaky kvantitativní (měřitelné) tak i pro znaky kvalitativní. Vzhledem k tomu, že v oborech studovaných na LDF MZLU v Brně analyzujeme v největší míře znaky **kvantitativní**, bude se tato kapitola zabý-

vat především jimi a metodami, které jejich vzájemné vztahy popisují – **korelační a regresní analýzou**.

Vztahy mezi jednotlivými znaky zkoumáme obvykle na **vícerozměrném statistickém souboru**.

10.1 Vícerozměrný statistický soubor

Vícerozměrný statistický soubor je množina *C* souběžných realizací určitého počtu veličin $X_1, X_2, \ldots, X_m$. Tento název přísluší též množině *n* objektů, jejichž *m* určitých vlastností je souběžně předmětem statistického šetření. Množina *C* vznikne získáním hodnot znaků $X_1, X_2, \ldots, X_m$ na prvcích množiny *n*. *C* je potom množina uspořádaných *m-tic* hodnot $[x_1, x_2, \ldots, x_m]$ znaků $X_1, X_2, \ldots, X_m$.

Vícerozměrný statistický soubor si můžeme představit např. na veličinách měřených při biometrických analýzách porostů. Například v určitém porostu je na jistém počtu stromů (množina *n* objektů) měřena výčetní tloušťka stromu, výška stromu, výška nasazení koruny a šířka koruny (*m* určitých vlastností). Znakem X_1 je potom výčetní tloušťka, znakem X_2 výška stromu, znakem X_3 výška nasazení koruny a znakem X_4 je šířka koruny. Počet zkoumaných vlastností $m = 4$. Výsledkem takového měření je množina uspořádaných *m-tic* hodnot konkrétních měřených hodnot $[x_1, x_2, \ldots, x_m]$, kde x_1 je konkrétní měřená výčetní tloušťka, x_2 výška, atd. Množinu C můžeme zapsat takto[1]

$$\mathbf{C} = \begin{bmatrix} \mathbf{x}_1^{\mathbf{T}} \\ \vdots \\ \mathbf{x}_j^{\mathbf{T}} \\ \vdots \\ \mathbf{x}_n^{\mathbf{T}} \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,i} & \cdots & x_{1,m} \\ \vdots & & \vdots & & \vdots \\ x_{j,1} & \cdots & x_{j,i} & \cdots & x_{j,m} \\ \vdots & & \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,i} & \cdots & x_{n,m} \end{bmatrix} \qquad (10.1)$$

Ve vztahu 10.1 si každý **řádek** můžeme představit jako **hodnoty všech veličin měřené na jednotlivém stromu** (např. $x_{1,1}$ je výčetní tloušťka 1. stromu, $x_{1,i}$ je i-tá vlastnost (např. výška nasazení koruny) 1. stromu atd. Každý **sloupec** představuje **jednu měřenou veličinu** (např. sloupec tvořený hodnotami $x_{1,1}, \ldots, x_{j,1}, \ldots, x_{n,1}$ jsou hodnoty první měřené veličiny (výčetní tloušťky) pro všech *n* měřených stromů). Všechny měřené hodnoty tohoto sloupce tvoří veličinu X_1 – výčetní tloušťku.

Tak jako u jednorozměrných veličin, i zde je typické, že jednotlivé veličiny jsou náhodné, tj. jejich konkrétní měřené hodnoty jsou výsledkem působení náhodných vlivů. Proto v této souvislosti mluvíme o **vícerozměrné náhodné veličině,** pro které platí stejné vlastnosti jako pro jednorozměrné náhodné veličiny (frekvenční a distribuční funkce, odhady parametrů polohy, rozptýlení i tvaru, statistické testy apod.), pouze jejich matematické vyjádření a manipulace s nimi je technicky obtížnější, protože obvykle vychází z operací s vektory a maticemi. V této kapitole teorii vícerozměrných náhodných veličin omezíme na minimální možnou míru a budeme se přede-

[1] Vzhledem k tomu, že v této kapitole budeme pracovat s vícerozměrným souborem, musíme v zápisech vzorců nějak odlišit zápis matic a vektorů od ostatních prvků vzorců – tedy **matice a vektory budou zapisovány tučně.**

vším zbývat výpočetními postupy těch metod, které jsou z praktického hlediska rozhodující a interpretací jejich výsledků. Podrobnější informace o vlastnostech vícerozměrných náhodných veličin je možné získat např. v MELOUN-MILITKÝ 1994 nebo ve specializovaných monografiích věnovaných tomuto tématu, např. SIOTANI ET ALL. 1985, KENDALL-STUART 1966, MORRISON 1984 a mnoho dalších.

Ve vícerozměrných souborech kromě analýzy jejich vlastností také zkoumáme vztahy mezi nimi, a to prostřednictvím statistické závislosti.

10.2 Statistická závislost a korelace

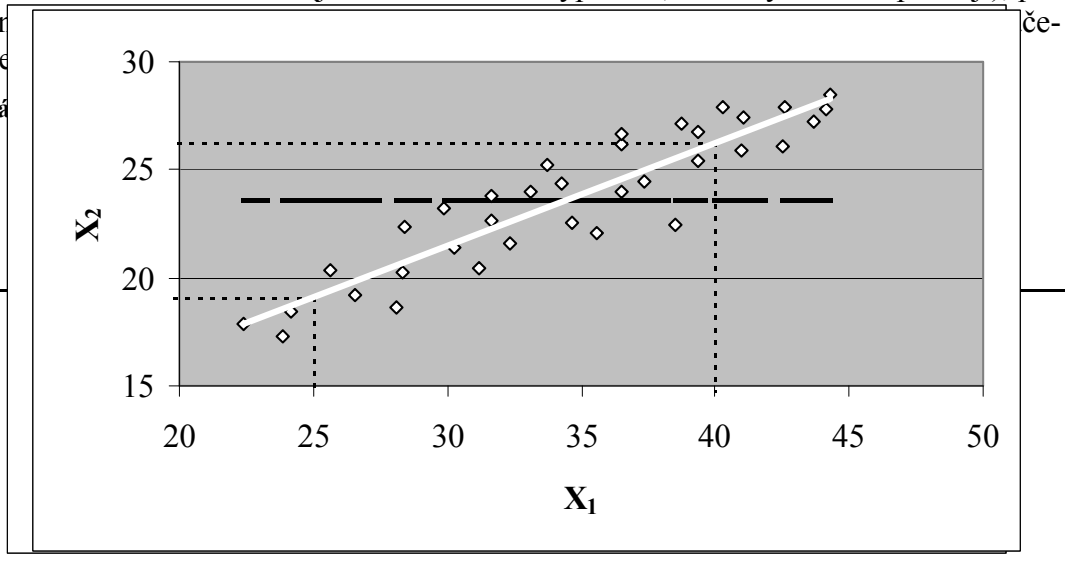
O statistické závislosti znaků $X_1, X_2, ..., X_m$ mluvíme, když hodnotě znaku X_i přísluší vždy nejméně jedna hodnota každého z ostatních znaků:

- jestliže hodnotě znaku X_i přísluší libovolné hodnoty všech ostatních znaků, nazýváme všechny znaky $X_1, X_2, ..., X_m$ **statisticky nezávislé**;
- jestliže hodnotě znaku X_i přísluší hodnoty všech ostatních znaků podle určitého pořádku, nazýváme všechny znaky $X_1, X_2, ..., X_m$ **stochasticky závislé**;
- jestliže hodnotě znaku X_i přísluší právě jedna hodnota všech ostatních znaků, nazýváme všechny znaky $X_1, X_2, ..., X_m$ **funkčně závislé**.

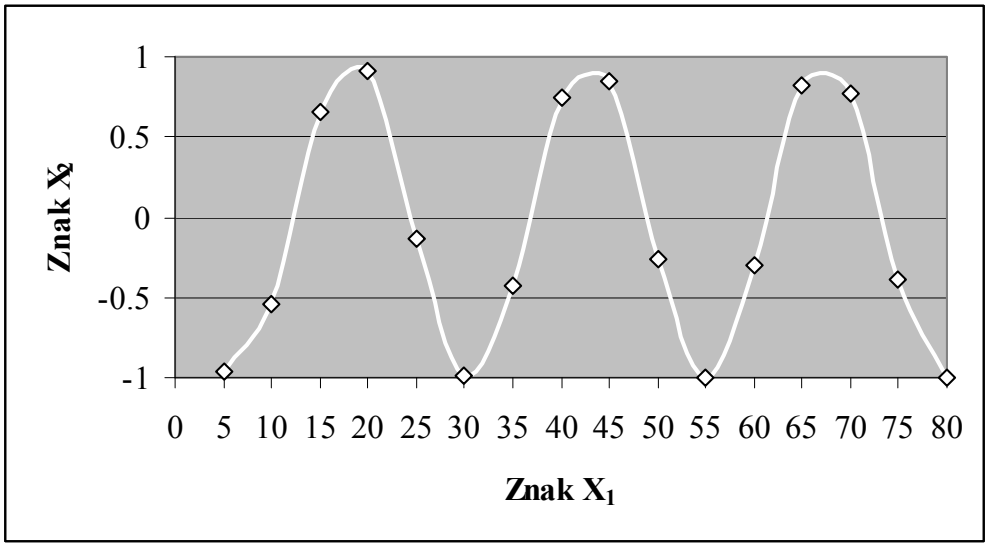
Sledujeme-li závislost znaku X_i na ostatních znacích $X_1, X_2, ..., X_{i-1}, X_{i+1}, ..., X_m$, nazýváme znak X_i závislý znak (**závislá proměnná**), ostatní znaky tvoří soubor nezávislých znaků (**nezávislých proměnných**).

Výše uvedené typy statistické závislosti jsou graficky znázorněny na obrázcích 10.1 , 10.2 a 10.3 . Na obrázku 10.1 je příklad nezávislých znaků, kdy zjevně mezi znaky X_1 a X_2 neexistuje žádný podstatný vztah, tedy jakékoli hodnotě jednoho znaku můžeme přiřadit libovolnou hodnotu znaku druhého a z hodnoty jednoho znaku nemůžeme odvodit hodnotu znaku druhého. Vzájemná nezávislost je vyjádřena na obrázku bílou přerušovanou čarou, která je rovnoběžná s osou X, což naznačuje neexistenci statistické závislosti. Znamená to, že „model" závislosti nepřinese jakékoli zlepšení oproti tomu, když pro jakoukoli hodnotu X_1 vyjádříme hodnotu X_2 pomocí aritmetického průměru. Na obrázku 10.2 je znázorněn případ stochastické závislosti, kdy je vidět mezi znaky X_1 a X_2 zřetelný vztah (určitý trend – ten je vyjádřen na obrázku bílou čarou), kdy je zřejmě možné najít vhodné vyjádření tohoto vztahu (obvykle matematický model – to je ten „určitý pořádek" z výše uvedené definice statistické závislosti) a pomocí tohoto modelu („pořádku") přiřadit známé hodnotě jednoho znaku hodnotu znaku druhého s určitou pravděpodobností. Zde je zřejmé, že oproti aritmetickému průměru přinese použití modelu zpřesnění určení hodnoty X_2 pro určitou hodnotu X_1. Například pro hodnoty $X_1 = 20$ a $X_1 = 40$ je hodnota X_2 vyjádřená aritmetickým průměrem stejná - 23.6 – naznačeno černou čárkovanou čarou. Pokud použijeme modelové závislosti (jak model určit a vypočítat, bude vysvětleno později), pro hodn[illegible] če-
no te[illegible]

Obrá[illegible]



Obrázek 10.2 – Příklad stochastické závislosti



Obrázek 10.3 – Příklad funkční závislosti

Na obrázku 10.3 je příklad poslední možnosti – funkční závislosti, kdy je každé hodnotě jednoho znaku přiřazena druhá hodnota zcela jednoznačně – podle funkčního předpisu, v tomto případě je to funkce $x_2 = \sin(x_1)$. V tomto případě je každé hodnotě jednoho znaku přiřazena pouze jedna hodnota druhého znaku. Je zřejmé, že funkční závislost je zvláštní případ stochastické závislosti, kdy je hodnota druhého znaku přiřazována ne s „určitou“, ale s jednoznačnou, tj. „stoprocentní“ pravděpodobností.

Statistická analýza nejčastěji zkoumá stochastické závislosti. Je tomu tak proto, že v reálném světě jsou prakticky všechny měřené nebo jinak zjišťované veličiny v různé míře zatíženy náhodnými vlivy. Tento prvek náhodnosti se samozřejmě promítá i do jejich vzájemných vztahů, které je proto možné charakterizovat jako pravděpodobnostní – stochastické.

Statistické závislosti se dělí do několika skupin podle typu znaků, jejichž závislost popisují:

- **korelace –** popisuje vliv změny úrovně nezávisle proměnných znaků na změnu úrovně závislého znaku a platí pro **kvantitativní (měřené) znaky;**
- **kontingence** – popisuje závislost kvalitativních (slovních, popisných) znaků, které mají více než dvě alternativy, tzv. **množných znaků** (např. druh dřeviny, národnost, apod.);
- **asociace** - popisuje závislost kvalitativních (slovních, popisných) znaků, které mají pouze dvě alternativy, tzv. **alternativních znaků** (např. pohlaví, odpovědi typu ano/ne, …).

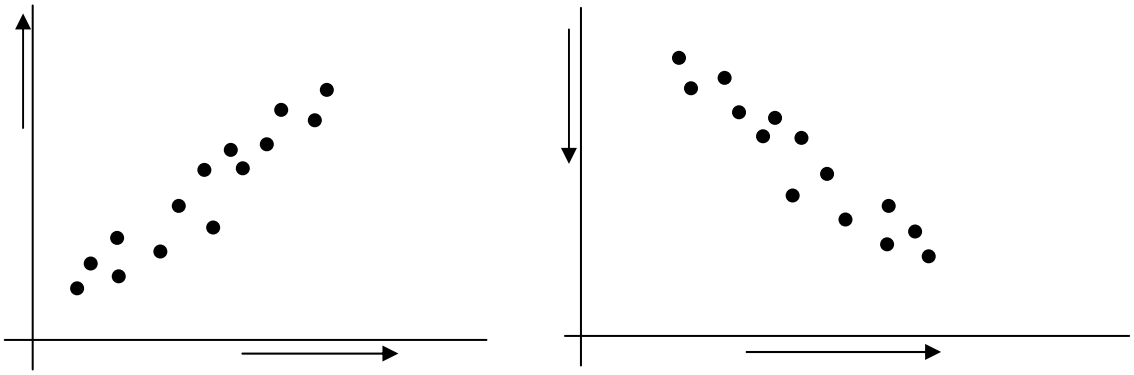
V následujícím textu se budeme z důvodů uvedených v úvodu této kapitoly zabývat především korelační závislostí.

Korelaci dělíme podle různých kritérií, např.:

- **počtu korelovaných znaků**
  - *jednoduchá* – popisuje vztah dvou znaků,

- *mnohonásobná* – popisuje vztahy více než dvou znaků,
- *parciální* – popisuje závislost dvou znaků ve vícerozměrném statistickém souboru při vyloučení závislosti ostatních znaků;

- **smyslu změny hodnot** analyzovaných znaků
  - *kladná* – se zvyšováním hodnot jednoho znaku se zvyšují i hodnoty druhého znaku (obrázek 10.4 vlevo),
  - *záporná* - se zvyšováním hodnot jednoho znaku se zmenšují hodnoty druhého znaku (obrázek 10.4 vpravo).



Obrázek 10.4 – Příklad kladné (vlevo) a záporné korelace (vpravo

Úlohy spojené s korelační závislostí řeší soubor metod a početních úkonů, které se souhrnně nazývají **korelační počet**. Podle zaměření úloh jej dělíme na

- **korelační analýzu**
  - zjišťuje *existenci závislosti* a její druhy,
  - měří *těsnost závislosti*,
  - ověřuje *hypotézy o statistické významnosti závislosti*;
- **regresní analýzu**
  - zabývá se *vytvořením vhodného matematického modelu* závislosti,
  - stanoví *potřebné parametry* tohoto *modelu*,
  - ověřuje *hypotézy o vhodnosti a důležitých vlastnostech modelu*.

V souvislosti s korelační a regresní analýzou se často hovoří také o korelačních a regresních modelech.

Závěrem této kapitoly je nutné zdůraznit, že **prokázání statistické závislosti ještě nemusí znamenat příčinnou (kauzální) závislost**. Lze najít mnoho případů, kdy určité veličiny vykazují statistickou závislost a přitom mezi nimi není možné prokázat žádnou skutečnou příčinnou závislost. Mnohé statistické příručky uvádějí různé humorné až absurdní případy „korelací“, např. mezi počtem mrazových dní a počtem vražd v USA (LEPŠ 1996) nebo příjmy kněží a spotřebou alkoholu, mezi růstem počtu televizí a počtem chovanců psychiatrických léčeben, apod. V takovýchto případech je zdánlivá korelace způsobena jinými, do této závislosti nezahrnutými faktory, a mluvíme o zprostředkované korelaci. **Je tedy vždy nutno pozorně zvážit na základě podrobné znalosti studovaného problému, jestli příslušná (statisticky prokázaná!) korelace má skutečně logické zdůvodnění a je možné ji rozumně interpretovat**.

10.3 Formulace korelačních a regresních modelů

Rozdíl mezi korelačními a regresními modely si můžeme dobře ukázat na dvou skupinách úloh, které se liší vzájemným postavením jednotlivých veličin.

10.3.1 Korelační modely

Pro tuto skupinu úloh je typické, že se jedná o vícerozměrný soubor, kde nám jde o postižení závislostí mezi jednotlivými proměnnými. Nemáme dopředu určeno, která proměnná na které závisí, tj. která je nezávislá a která je závislá. Tyto úlohy může převzít kterákoli proměnná. Model tohoto typu se nazývá **korelační**. Jednotlivé údaje pro korelační model se získávají obvykle měřením a jejich hodnoty jsou experimentátorem neovlivnitelné. Konkrétní měřené hodnoty jsou zpravidla náhodným výběrem ze základního souboru. Mezi typické příklady patří:

- vztah mezi tloušťkou a výškou měřenou na náhodně vybraných stromech,
- vztah mezi tloušťkovým přírůstem a klimatickými faktory měřenými na náhodně vybraných bodech,
- vztah mezi délkou a šířkou listů měřenou na náhodně vybraných listech,

apod.

Obecně lze korelační model maticově zapsat jako *n* x *m* rozměrné pole dat, kde *m* je počet proměnných (tj. sloupců matice **X**), kde *j* = 1,2, ..., *m* a *n* je počet hodnot každé proměnné (tj. počet *m*-rozměrných bodů x_i), kde *i* = 1.2. ..., *n*.:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & X_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nm} \end{bmatrix} \qquad (10.2)$$

10.3.2 Regresní modely

Druhá skupina úloh se liší v tom, že **vysvětlující (nezávislá) proměnná je předem nastavovaná** a experimentátorem ovlivnitelná, tedy je nenáhodná. Znamená to, že máme dopředu určeno, které veličiny jsou nezávisle proměnné a která je závisle proměnná. Tyto modely se nazývají **regresní**. Jako typické příklady můžeme uvést:

- vztah mezi taxační veličinou porostu a věkem, kdy věk je předem určen (např. růstová řada porostů s věkem odstupňovaným po 10 letech, ve kterých měříme danou taxační veličinu);
- vztah mezi výškou a tloušťkou (tloušťka je dopředu určena, např. tloušťkovými stupni a měříme pouze výšku pro stromy určené tloušťky);
- vztah mezi výškovým růstem sazenic a odstupňovanými dávkami hnojiva (dávky hnojiva jsou pevně nastaveny, výškový růst se měří jako náhodná veličina);

apod.

V zásadě se regresní modely dělí na dvě hlavní skupiny - **lineární** a **nelineární**, a to buď z hlediska parametrů nebo proměnných. Nejdůležitější je hledisko linearity z hlediska parametrů.

Za **lineární regresní model** se považuje takový, který **má parametry v lineárním postavení**. Z toho vyplývá, že za lineární model se považuje i takový, jehož grafickým obrazem je křivka, ale jehož parametry jsou ve vzájemném lineárním postavení. Např. model $y = a + bx + cx^2$ je lineárním modelem, protože parametry *a, b, c* má v lineárním postavení, ačkoli jeho obrazem je křivka - parabola. Nejběžnějším lineárním modelem je samozřejmě přímkový model $y = a + bx$.

Lineární regresní model je možné formulovat takto

$$E(y/\mathbf{x}) = \sum_{j=0}^{m} \beta_j \cdot f_j(\mathbf{x}) \qquad (10.3)$$

kde je

$E(y/\mathbf{x})$ podmíněná střední hodnota náhodné veličiny *y* v místě ***x***

β_j *j*-tý parametr regresní funkce

$f(x_j)$ regresor regresní funkce (nějaká funkce nezávisle proměnné *x*)

Regresory již nesmějí obsahovat žádný neznámý parametr regresní funkce (možné regresory jsou např. x, x^2, cos x, ...). Pokud regresní model obsahuje absolutní člen, potom je *j = 0, 1, 2, ..., m*, pokud ho neobsahuje, potom *j = 1, 2, ..., m*.

Regresní model je buď aproximací „ideálního“ (teoretického) modelu $f_T(\mathbf{x_i},\boldsymbol{\beta})$ nebo je odvozen na základě znalosti chování modelovaného experimentálního systému. Problémem je to, že zpravidla „ideální“ model neznáme, a proto ho nahrazujeme více nebo méně přesnou aproximací. To znamená, že místo neznámých teoretických parametrů $\boldsymbol{\beta}$ vypočítáme vhodnou metodou „pouze“ jejich odhady **b**, tedy skutečný regresní model se může vyjádřit

$$y' = \sum_{j=0}^{m} b_j f_j(\mathbf{x}) \qquad (10.4)$$

Za **nelineární modely** se považují takové, jejichž **parametry nejsou ve vzájemném lineárním postavení**, např. $y = a \cdot x^b$ nebo $y = a \cdot e^{bx}$ a mnoho a mnoho dalších. Výpočet jejich parametrů je obtížnější než v předchozím případě a proto se jejich použití se rozšířilo až v poslední době, kdy problém s obtížným výpočtem parametrů díky výpočetní technice pomalu mizí. Ze skupiny nelineárních modelů se může z praktických důvodů vyčlenit skupina **linearizovatelných** modelů, což jsou nelineární modely, které lze vhodnou transformací převést na lineární model. Např. Michajlovovu růstovou funkci

$$y = a \cdot e^{\frac{k}{x}} \qquad (10.5)$$

lze převést logaritmickou transformací na lineární tvar $\ln y = \ln a + k \cdot (1/t) \cdot \ln e$. Tím lze zjednodušit výpočet parametrů, které se vypočítají pro zlinearizovaný tvar modelu a zpětně se retransformují na parametry nelineárního modelu. Tento způsob ovšem není statisticky zcela „čistý“ a v současné době, kdy je možné používat profesionální

statistické programy s kvalitními algoritmy pro výpočet nelineárních modelů, se používá jen v případech, kdy tyto programy nejsou k dispozici nebo pro prvotní odhad parametrů.

Na tomto místě je nutno zdůraznit, **že není vždy možné korelační a regresní modely od sebe přísně oddělovat**. Většina hodnot vstupujících do statistické analýzy je získána přímým pozorováním (měřením), a tedy možnost předem nastavit nezávisle proměnnou bývá v mnoha případech omezená nebo nemožná. Například pro veličiny výčetní tloušťka a výška, které byly získány náhodným výběrem, se také obvykle stanovuje regresní model, i když, přísně vzato, zde není možné pokládat jednu proměnnou za závislou na druhé a náhodná variabilita obou proměnných bude přibližně stejná. V těchto případech je zřejmě vhodnější hovořit o **vysvětlující proměnné** (místo o nezávislé) a o **vysvětlované proměnné** (místo o závislé). V případě že X je náhodná proměnná, chápeme regresi Y na X jako studium závislosti odpovědi (reakce) veličiny Y na zjištěných hodnotách vysvětlující proměnné X. Vztahy používané při řešení korelačních a regresních modelů jsou v podstatě shodné, liší se hlavně jejich význam a interpretace.

10.4 Korelační analýza lineárního modelu

Základním prostředkem korelační analýzy jsou míry korelace. Jsou to statistické charakteristiky, které popisují těsnost studované závislosti. Jak bylo uvedeno v předchozí kapitole, z hlediska korelační analýzy není podstatné, která veličina je vysvětlovaná a která vysvětlující.

10.4.1 Korelační koeficient

Korelační koeficient je základní mírou lineární závislosti. Rozlišujeme několik typů korelačních koeficientů, z nichž mezi nejdůležitější patří:

- **vícenásobný** - definuje míru lineární stochastické závislosti mezi náhodnou veličinou X_1 a nejlepší lineární kombinací složek $X_2, X_3, ..., X_m$ náhodného vektoru **X**,
- **párový** - zvláštní případ vícenásobného korelačního koeficientu, kdy vyjadřuje míru lineární stochastické závislosti mezi náhodnými veličinami X_i a X_j,
- **parciální** - definuje míru lineární stochastické závislosti mezi náhodnými veličinami X_i a X_j při zkonstantnění dalších složek vektoru **X**,
- **pořadový** - neparametrická modifikace párového korelačního koeficientu.

K pochopení významu a funkce korelačního koeficientu v korelační a regresní analýze je vhodné blíže osvětlit jeho podstatu.

Uvažujme nejjednodušší případ párového korelačního koeficientu (všechny následující úvahy platí i pro vícerozměrné výběry, ale tyto případy není možné graficky znázornit).

Mějme náhodný vektor **X** se dvěma složkami x_1 a x_2. Na obrázku 10.5 jsou černými body znázorněny experimentální (měřené) hodnoty (x_{2i}), bílými kolečky jsou znázorněny odpovídající hodnoty vypočítané na základě regresního modelu (x'_{2i}) – o

způsobu jeho výpočtu bude pojednáno v kapitole 10.5. Čárkovaně je vyznačena poloha aritmetického průměru závisle proměnné ($\overline{x}_2$).

Při odvození korelačního koeficientu vycházíme z rozkladu celkového rozptylu experimentálních bodů okolo aritmetického průměru

$$S^2_{x_2} = \frac{\sum_{i=1}^{n} \left(x_{2i} - \overline{x}_2\right)^2}{n} \tag{10.6}$$

na dvě složky:

- **rozptyl vysvětlený regresním modelem** (rozptyl bodů regresního modelu okolo celkového aritmetického průměru)

$$S^2_{x'_2} = \frac{\sum_{i=1}^{n} \left(x'_{2i} - \overline{x}_2\right)^2}{n} \tag{10.7}$$

- **rozptyl reziduální** (rozptyl experimentálních bodů okolo vypočítaných hodnot regresního modelu)

$$S^2_{x_2x_1} = \frac{\sum_{i=1}^{n} \left(x_{2i} - x'_{2i}\right)^2}{n} \tag{10.8}$$

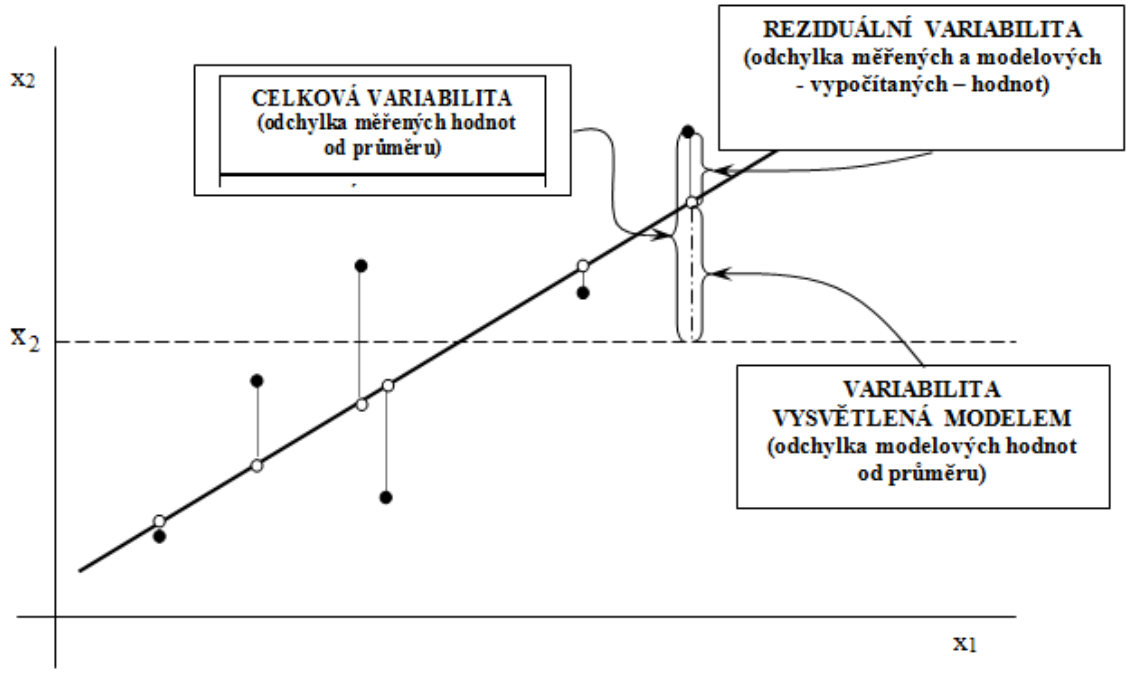
Zatímco první složku (rozptyl vysvětlený modelem) můžeme vysvětlit závislostí x_2 na x_1, druhou složku (rozptyl reziduální) musíme přisoudit vlivu neuvažovaných nebo neznámých činitelů.

Na stupeň korelace můžeme tedy usuzovat podle toho, **jakým dílem se obě složky podílejí na celkovém rozptylu hodnot x_2**. Kdyby totiž model dokonale vystihoval danou závislost, byl by celkový rozptyl plně vysvětlen prvou složkou (rozptylem vysvětleným modelem). Platí tedy, že čím je větší podíl první složky, tím je korelace těsnější a znak x_1 přispívá ke zpřesnění odhadu hodnoty znaku x_2. K číselnému vyjádření stupně (míry) korelace se tedy může použít poměr

$$R^2 = \frac{S^2_{x'_2}}{S^2_{x_2}} = 1 - \frac{S^2_{x_1x_2}}{S^2_{x_2}} \tag{10.9}$$

kde R^2 se nazývá **koeficient determinace**. Vyjadřuje, jaká **část celkového rozptylu je vysvětlena modelem**. Jeho odmocnina se nazývá **koeficient korelace** a používá se jako nejběžnější míra **lineární** korelace

$$R = \sqrt{\frac{S^2_{x'_2}}{S^2_{x_2}}} = \sqrt{1 - \frac{S^2_{x_2x_1}}{S^2_{x_2}}} \tag{10.10}$$



Obrázek 10.5 - Rozklad celkového rozptylu v regresním modelu

10.4.1.1 Párový korelační koeficient

Párový korelační koeficient slouží v případě jednoduché korelace (závislosti dvou veličin) k posouzení těsnosti závislosti. Korelačních koeficientů je několik druhů, mezi nejběžnější patří:

- **Pearsonův**
- **Spearmanův** (korelace pořadí)

10.4.1.1.1 Pearsonův korelační koeficient

Pearsonův korelační koeficient je základní mírou lineární korelace. Vzhledem k tomu, že jeho výpočet vychází z momentových charakteristik polohy a variability, je nezbytnou podmínkou jeho použití **dvourozměrné normální rozdělení.** Korelační koeficient se stanoví se podle vzorce

$$r_{x_1x_2} = r_{x_2x_1} = \frac{cov_{x_1x_2}}{S_{x_1} \cdot S_{x_2}} \qquad (10.11)$$

kde výraz

$$cov_{x_1x_2} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{1i} - \bar{x}_1\right)\cdot\left(x_{2i} - \bar{x}_2\right) \qquad (10.12)$$

se nazývá **kovariance.** Obvyklejší výraz (používá se, jestliže kovarianci počítáme z výběru, což je nejčastější případ, je to vlastně bodový odhad kovariance základního souboru) je

$$\mathrm{cov}_{x_1x_2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{1i}-\bar{x}_1\right)\cdot\left(x_{2i}-\bar{x}_2\right) \qquad (10.13)$$

Korelační koeficient je tedy **standardizovaná (normovaná) kovariance**. Tato standardizace se provádí proto, aby bylo možné pracovat s veličinami měřenými v různých jednotkách a bylo možné těsnost závislosti porovnávat. Výsledný vztah pro Pearsonův korelační koeficient je

$$r_{x_1x_2} = r_{x_2x_1} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(x_{1i}-\bar{x}_1\right)}{S_{x_1}}\cdot\frac{\left(x_{2i}-\bar{x}_2\right)}{S_{x_2}} \qquad (10.14)$$

nebo pro bodový odhad kovariance

$$r_{x_1x_2} = r_{x_2x_1} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{\left(x_{1i}-\bar{x}_1\right)}{\hat{S}_{x_1}}\cdot\frac{\left(x_{2i}-\bar{x}_2\right)}{\hat{S}_{x_2}} \qquad (10.15)$$

kde je $\hat{S}_{x_i}$ pro *i = 1,2* bodový odhad směrodatné odchylky podle vztahu

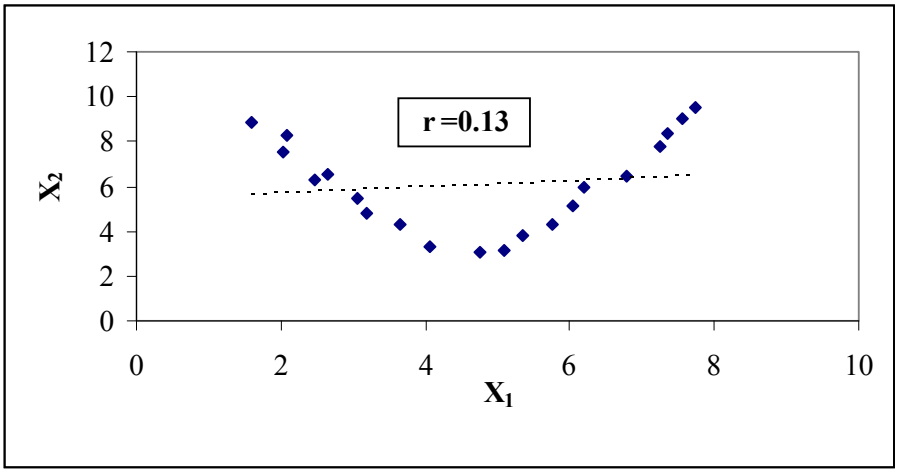
$$\hat{S}_{x_i} = \frac{\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)^2}{n-1} \qquad (10.16)$$

Základní vlastnosti korelačního koeficientu jsou následující:

- je to bezrozměrná míra lineární korelace;
- nabývá hodnoty 0 –1 pro kladnou korelaci, 0 – (-1) pro zápornou korelaci;
- hodnota 0 znamená, že mezi posuzovanými veličinami není žádný **lineární** vztah (může být nelineární) nebo tento vztah zůstal na základě dat, které máme k dispozici, neprokázán;
- hodnota 1 nebo (-1) indikuje funkční závislost;
- hodnota korelačního koeficientu je stejná pro závislost x_1 na x_2 i pro opačnou závislost x_2 na x_1.

Je nutno zdůraznit, že **nekorelovanost** (tj. r = 0) **ještě nemusí znamenat nezávislost posuzovaných veličin!** Musíme si uvědomit, že korelační koeficient měří jen přímočarou závislost a pokud je závislost křivočará, nemusí Pearsonův korelační koeficient ukazovat žádnou těsnou vazbu mezi veličinami (i když mezi nimi ve skutečnosti existuje!!). Příklad takových dat ukazuje obrázek 10.6 .
Zde je jasně vidět výrazná kvadratická závislost. Korelační koeficient se ale blíží 0, což indikuje nekorelovanost. Pokud použijeme vhodný kvadratický model, míra korelace bude 0.96, tedy velmi silná.

Sílu závislosti musíme posuzovat vždy v kontextu posuzovaných veličin, především vzhledem k tomu, **co víme o závislostech mezi nimi**! Mnohé statistické příručky uvádějí „univerzální“ stupnice hodnocení síly korelace,ale není vhodné tyto stupnice opravdu univerzálně používat. Jsou to stupnice typu: $r < 0.2$ – žádná nebo velmi slabá korelace, 0.2 – 0.4 slabá korelace, atd. Je nutné si uvědomit, že v určitém typu závislosti bude r = 0.8 považováno za slabší závislost (za předpokladu, že obvyklé hodnoty *r* se pohybují třeba v rozmezí 0.90 – 0.95), naopak u jiných veličin může být $r = 0.5$ považováno za silnou závislost. **Toto hodnocení je nutné vždy provádět s hlubokou znalostí řešené problematiky, ne schématicky podle obecných stupnic!**



Obrázek 10.6 – Příklad dat s křivočarou závislostí, kde Pearsonův korelační koeficient indikuje lineární nekorelovanost (*r* se blíží nule)

10.4.1.1.2 Spearmanův korelační koeficient

Tento koeficient se také nazývá **korelace pořadí**, protože vychází nikoli z měřených hodnot, ale z jejich pořadí. Je to tedy obdoba kvantilových charakteristik nebo neparametrických testů. Má také obdobné použití. Spearmanův korelační koeficient se používá tehdy, je-li hrubě porušen předpoklad dvojrozměrné normality a není tedy možné použít Pearsonův korelační koeficient. Nejčastější příčinou porušení normality jsou odlehlá měření, která mohou Pearsonův korelační koeficient naprosto „zmást“ a pokud bychom soudili podle jeho výsledku, dospěli bychom pravděpodobně k velmi nesprávným závěrům. Takovým bodům se říká vlivné a jejich detekci a dalšímu zpracování bude věnována část kapitoly o regresní diagnostice. Pro normálně rozložená data jsou hodnoty obou koeficientů velmi blízké. Pokud se jejich hodnoty značně liší, je v datech nějaký problém, který je záhodno prozkoumat.

Výpočet Spearmanova korelačního koeficientu vychází z pořadí hodnot, které stanovíme pro oba soubory zvlášť. Postupujeme podle stejných zásad jako u neparametrických testů (viz např. Wilcoxonův test v I. dílu, str. 126 nebo Kruskal - Wallisův test v 9. kapitole tohoto dílu), tj. seřadíme hodnoty od nejmenší k největší, s tím, že stejným hodnotám přiřadíme průměrná pořadí. Poté spočítáme diference jednotlivých pořadí d_i a vypočítáme hodnotu Spearmanova korelačního koeficientu podle vztahu

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n^3 - n} \qquad (10.17)$$

Pokud se v datech vyskytují skupiny stejných hodnot, počítá se Spearmanův korelační koeficient podle upraveného vzorce

$$r_S = \frac{\frac{n^3 - n}{6} - \sum_{i=1}^{n} d_i^2 - \sum T_{X_1} - \sum T_{X_2}}{\left(\frac{n^3 - n}{6} - 2\sum T_{X1}\right)\left(\frac{n^3 - n}{6} - 2\sum T_{X_2}\right)} \qquad (10.18)$$

kde je

$$\sum T_{X_1} = \frac{\sum_{i=1}^{m}\left(t_i^3 - t_i\right)}{12} \qquad (10.19)$$

$$\sum T_{X_2} = \frac{\sum_{i=1}^{m}\left(t_i^3 - t_i\right)}{12} \qquad (10.20)$$

kde je t_i počet stejných hodnot v *i-té* skupině stejných hodnot v souboru X_1, resp. X_2. Výsledek podle vztahu 10.18 se od výsledku podle vztahu 10.17 podstatněji liší jen tehdy, je-li v obou souborech velké množství skupin stejných dat.

Příklad 10.1:

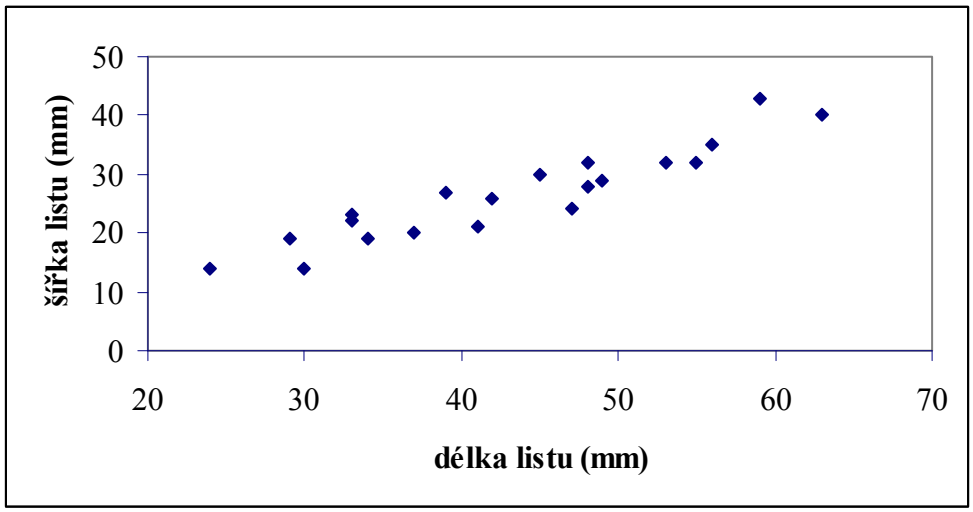
Je dán výběr velikosti n = 20, kde jedna proměnná představuje měřené hodnoty délky bukových listů, druhá proměnná šířky listů (v mm). Data jsou v tabulce 10.1 . Vypočítejte Pearsonův a Spearmanův korelační koeficient.

Zadání - měřené hodnoty			**Pořadí pro výpočet Spearmanova korelačního koeficientu**		
Číslo měření	**Délka listu (x_1)**	**Šířka listu (x_2)**	**Délka listu (K_1)**	**Šířka listu (K_2)**	**$d^2 = (K_1 - K_2)^2$**
1	24	14	1	1.5	0.25
2	29	19	2	3.5	2.25
3	30	14	3	1.5	2.25
4	33	22	4.5	7	6.25
5	33	23	4.5	8	12.25
6	34	19	6	3.5	6.25
7	37	20	7	5	4.00
8	39	27	8	11	9.00
9	41	21	9	6	9.00
10	42	26	10	10	0.00
11	45	30	11	14	9.00
12	47	24	12	9	9.00
13	48	28	13.5	12	2.25
14	48	32	13.5	16	6.25
15	49	29	15	13	4.00
16	53	32	16	16	0.00
17	55	32	17	16	1.00
18	56	35	18	18	0.00
19	59	43	19	20	1.00
20	63	40	20	19	1.00

Tabulka 10.1 - Zadání příkladu 10.1 a pořadí pro výpočet Spearmanova korelačního koeficientu

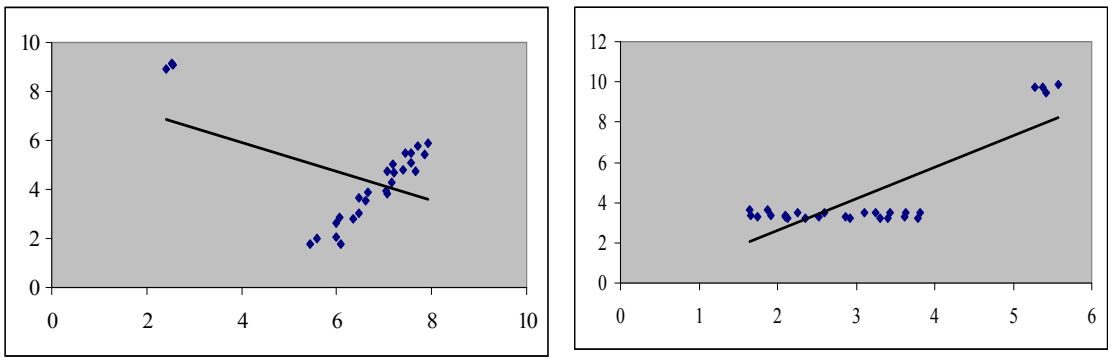
Pearsonův korelační koeficient vypočítáme podle vzorce 10.14, kde je $\bar{x}_1$ = 43.25 mm, $\bar{x}_2$ = 26.50 mm, S_{x1} = 10.58 mm a S_{x2} = 7.665 mm. Výsledek je 0.9334. Toto číslo svědčí o vysokém stupni korelace mezi oběma veličinami.

Pořadí pro Spearmanův korelační koeficient jsou již připravena v tabulce 10.1 . Použijeme vzorec 10.17, protože skupin stejných dat je jen málo (v každém výběru dvě), takže korekce podle vzorce 10.18 je jen nepatrná. Hodnota $\Sigma d_i^2 = 85$ výsledná hodnota je $r_s = 0.9361$. Pokud bychom použili korekci podle vzorce 10.18, byl by výsledek 0.9359. Data jsou graficky znázorněna na obrázku 10.7 . Vidíme, že se jedná o datové soubory bez vybočujících (vlivných) bodů, což potvrzují téměř shodné hodnoty obou korelačních koeficientů (0.933 a 0.936).



Obrázek 10.7 – Grafické znázornění dat Příkladu 10.1

Následující příklady na obrázku 10.8 ukazují rozdíly mezi použitím Pearsonova a Spearmanova koeficientu pro data s vlivnými body.



Obrázek 10.8 - Příklady dat s vlivnými body. Čára ukazuje zdánlivý přímkový trend.

Obrázek 10.8 vlevo ukazuje případ poměrně silné kladné korelace (body vpravo) s několika odlehlými body vlevo nahoře. Pearsonův korelační koeficient to ihned „zaznamená“ a jeho hodnota je –0.449, tj. záporná korelace! Odlehlé body (pouze tři z celkového počtu 28) zcela obrátily smysl korelace. Spearmanův korelační koeficient má hodnotu 0.391, tedy zachová smysl korelace.

Příklad napravo ukazuje jinou možnost – dva v podstatě samostatné soubory, které každý sám nemají významný korelační vztah – body jdou rovnoběžně s osou X. Jejich spojením vznikne zdánlivě poměrně silná korelace, kterou Pearsonův korelační koeficient „ocení“ hodnotou 0.818, zatímco Spearmanův korelační koeficient dosáhne nízké hodnoty (statisticky nevýznamné) 0.283.

Obecně tedy platí, že pro normální data bez vlivných bodů používáme Pearsonův korelační koeficient, pro data s vlivnými body nebo s dvourozměrným výrazně nenormálním rozdělením je vhodnější neparametrický Spearmanův koeficient. V každém případě je vhodné znázornit data graficky a analyzovat případné problémové hodnoty.

10.4.1.2 Mnohonásobný korelační koeficient

Mnohonásobný korelační koeficient používáme tehdy, zkoumáme-li závislost více veličin než dvou. V případě korelačních modelů to znamená, že matice **X** má více než dva sloupce (náhodné veličiny), v případě regresních modelů je zkoumána závislost mezi vysvětlovanou (závislou) proměnnou a dvěma a více vysvětlujícími (nezávislými) proměnnými.

Základem pro výpočet mnohonásobného korelačního koeficientu je korelační matice párových (jednoduchých) korelačních koeficientů **R**. Je to symetrická čtvercová matice řádu *m x m*, kde na diagonále jsou jedničky a mimodiagonální prvky jsou tvořeny párovými korelačními koeficienty R_{ij}

$$\mathbf{R} = \begin{bmatrix} 1 & R_{12} & \cdots & R_{1i} & \cdots & R_{1m} \\ R_{21} & 1 & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & 1 & \cdots & \cdots & \cdots \\ R_{i1} & \cdots & \cdots & 1 & \cdots & R_{im} \\ \cdots & \cdots & \cdots & \cdots & 1 & \cdots \\ R_{m1} & R_{m2} & \cdots & R_{mi} & \cdots & 1 \end{bmatrix} \qquad (10.21)$$

V matici **R** platí, že $R_{ij} = R_{ji}$. Z matice **R** lze vypočítat mnohonásobný korelační koeficient pro závislost mezi x_1 a vektorem $\mathbf{x}^*$ (tvořeným složkami x_2, ..., x_m) podle vzorce

$$R_{1(2,3,\ldots,m)} = \sqrt{1 - \frac{\det(\mathbf{R})}{\det(\mathbf{R_{(11)}})}} \qquad (10.22)$$

kde je

det(.) determinant výrazu v závorce

$R_{(ij)}$ matice vzniklá vypuštěním *i*-tého řádku a *j*-tého sloupce (v tomto případě prvního řádku a sloupce, obecně vždy čísla před závorkou, které označuje závisle proměnnou)

Mezi základní vlastnosti mnohonásobného korelačního koeficientu patří:

- $0 \le R \le 1$
- pokud je R = 1, znamená to, že závisle proměnná x_1 je přesně lineární kombinací veličin $x_2, ..., x_m$
- pokud je R = 0, potom jsou také všechny párové korelační koeficienty nulové
- s růstem počtu vysvětlujících (nezávislých) proměnných hodnota vícenásobného korelačního koeficientu neklesá, tj. platí

$$R_{1(2)} \le R_{1(2,3)} \le ... \le R_{1(2,\ ...,\ m)}$$

Výpočet mnohonásobného korelačního koeficientu podle vzorce 10.22 je velmi rychlý a výhodný např. pomocí tabulkového kalkulátoru. Běžně používané kalkulátory (např. Excel) jsou schopny samy spočítat korelační matici i determinant matice, což umožní velmi rychlý výpočet mnohonásobného korelačního koeficientu.

10.4.1.3 Parciální korelační koeficient

V řadě případů je potřebné sledovat ve vícerozměrném souboru (výběru) intenzitu vztahu mezi dvěma proměnnými při vyloučení vlivu ostatních. K tomuto účelu se používá **parciální korelační koeficient**, který **definuje míru lineární stochastické závislosti mezi náhodnými veličinami x_i a x_j při zkonstantnění dalších složek vektoru X**. Podle toho, kolik dalších proměnných je z hodnocení závislosti „vyloučeno“, rozlišují se parciální korelační koeficienty různých řádů. Párový korelační koeficient je vlastně parciální korelační koeficient nultého řádu (žádná proměnná není vyloučena). Parciální korelační koeficient prvního řádu sleduje závislost mezi dvěma proměnnými při vyloučení vlivu třetí (její označení se dává do závorky) - např. $R_{12(3)}$.

Výpočet parciálních korelačních koeficientů je také založen na párových korelačních koeficientech.

Parciální korelační koeficient prvního řádu se obecně vypočítá podle vztahu

$$R_{ij(k)} = \frac{R_{ij} - R_{ik}R_{jk}}{\sqrt{\left(1 - R_{ij}^2\right)\left(1 - R_{jk}^2\right)}} \qquad (10.23)$$

Parciální korelační koeficient druhého řádu se obecně vypočítá podle vztahu

$$R_{ij(kl)} = \frac{R_{ij(k)} - R_{ik(l)}R_{jk(l)}}{\sqrt{\left(1 - R_{ik(l)}^2\right)\left(1 - R_{jk(l)}^2\right)}} \qquad (10.24)$$

Ze vzorců10.23 a 10.24 vyplývá obecný výraz pro výpočet parciálního korelačního koeficientu (*m-1*)-ho řádu

$$R_{ij(1,2,...,i-1,i+1,...,j-1)} = \frac{A - B \cdot C}{\sqrt{\left(1 - B^2\right)\left(1 - C^2\right)}} \qquad (10.25)$$

kde je

A $R_{ij(1,2,...,i-1,i+1,...,j-2)}$

B $R_{i,j-1(1,2,...,i-1,i+1,...,j-2)}$

C $R_{j,j-1(1,2,...,j-1)}$

Ze vztahu 10.25 je zřejmé, že se jedná o vzorec, kdy k výpočtu parciálního korelačního koeficientu určitého řádu, např. *r-tého*, je nutné znát parciální korelační koeficienty 1., 2., ..., *r*-1. řádu. Tyto vztahy jsou vhodné především pro „ruční“ výpočty na kalkulačce, ale jsou (zvláště při výpočtu parciálních korelačních koeficientů vyšších řádů) velmi zdlouhavé a náročné na přesnost (jakákoli chyba se přenáší do výpočtu parciálních korelačních koeficientů vyšších řádů).

Pro výpočet na počítači je vhodné užít vzorce

$$R_{ij(1,2,...,m)} = \frac{(-1)^j \cdot \det(\mathbf{R}_{(ij)})}{\sqrt{\det(\mathbf{R}_{(ii)}) \cdot \det(\mathbf{R}_{(jj)})}} \qquad (10.26)$$

kde je

$\det(\mathbf{R}_{(ij)})$ determinant korelační matice **R** (viz vzorec) s vynechaným *i*-tým řádkem a *j*-tým sloupcem

Výpočet pomocí vzorce 10.26 se také velmi pohodlně provádí v tabulkovém kalkulátoru.

Význam parciálních korelačních koeficientů je např. v tom, že s jejich pomocí lze odhalit klamné (zdánlivé) korelace. Je nutné si uvědomit, že významná párová korelace není důkazem skutečné příčinné souvislosti. Představme si situaci, kdy zkoumáme stupeň korelace v korelačním modelu se složkami x_1, x_2 a x_3, kdy za vysvětlovanou proměnnou považujeme x_1 a zkoumáme stupeň závislosti na ostatních dvou složkách. Vysoká míra korelace, např. R_{12}, nemusí ještě znamenat, že jev vyjádřený náhodnou veličinou x_2 má skutečnou příčinnou souvislost s jevem vyjádřeným náhodnou veličinou x_1. Je nutné také zkoumat vzájemný vztah vysvětlujících proměnných x_2 a x_3, protože v případě jejich silné korelace by vysoká hodnota R_{12} mohla být způsobena právě silnou vzájemnou korelací vysvětlujících proměnných a nikoli vlivem x_2 na x_1. Právě v takovýchto případech mohou pomoci parciální korelační koeficienty.

Využití korelačních koeficientů pro hodnocení závislosti si ukážeme na následujícím příkladu.

Příklad 10.2:

V rámci biometrického výzkumu byl na jednotlivých stromech výzkumné plochy zkoumán vztah mezi veličinami objem (v), výčetní tloušťka ($d_{1.3}$, zde zjednodušeně označeno d), výška (h) a délka zelené koruny (k). Vyšetřete těsnost korelační závislosti objemu na tloušťce, výšce a délce zelené koruny. Měřené hodnoty jsou v tabulce 10.2 .

Zadání budeme považovat za korelační model se čtyřmi náhodnými veličinami v, d h, k (všechny hodnoty byly měřeny nebo vypočítány z naměřených veličin na náhodně vybraných stromech). Naším úkolem je zjistit, zda je oprávněný předpoklad, že objem stromu je v korelační závislosti na výčetní tloušťce, výšce stromu a délce zelené koruny.

Prvním krokem bude výpočet mnohonásobného korelačního koeficientu $R_{v(d,h,k)}$. Využijeme výpočtu podle vzorce 10.22. K tomuto výpočtu potřebujeme nejprve sestavit korelační matici **R**, která bude zahrnovat párové korelační koeficienty pro všechny možné kombinace proměnných. Pro náš příklad vypadá korelační matice takto:

	v	d	h	k
v	1	0.98096	0.93911	0.92014
d	0.98096	1	0.92987	0.90576
h	0.93911	0.92987	1	0.93457
k	0.92014	0.90576	0.93457	1

Vidíme, že korelační koeficienty jsou ve všech případech vysoké, to znamená, že existují statisticky významné korelační závislosti nejen mezi vysvětlovanou a vysvětlujícími proměnnými, ale i mezi vysvětlujícími proměnnými navzájem. Je zřejmé, že mnohonásobný korelační koeficient bude také vysoký (musí být nejméně roven nejvyššímu z párových korelačních koeficientů). Pro další výpočty a snazší označování řádků a sloupců matic označíme objem (v) jako proměnnou 1, tloušťku (d) jako proměnnou 2, výšku (h) jako proměnnou 3 a délku zelené koruny (k) jako proměnnou 4. K výpočtu $R_{1(2,3,4)}$ potřebujeme kromě základní korelační matice R také matici s vypuštěným řádkem a sloupcem vysvětlované proměnné (v našem případě 1. proměnná - objem, tedy $R_{(11)}$).

	2	3	4
2	1	0.92987	0.90576
3	0.92987	1	0.93457
4	0.90576	0.93457	1

Vypočítáme determinanty matice R a matice $R_{(11)}$ (buď pomocí „křížového pravidla“ nebo můžeme např. využít speciální funkce DETERMINANT tabulkového kalkulátoru Excel) a dosadíme do vzorce

$$R = \sqrt{1 - \frac{0.000487}{0.015782}} = 0.98445$$

Vzhledem k vysokým korelacím mezi vysvětlujícími proměnnými je zde reálný předpoklad, že vysoké hodnoty korelačních koeficientů mezi objemem a vysvětlujícími proměnnými nemusí být pouze důsledkem reálné příčinné závislosti, ale mohou být způsobeny právě těsnými vazbami mezi vysvětlujícími proměnnými. Abychom zjistili skutečný stupeň závislosti objemu na zadaných vysvětlujících proměnných, musíme zjistit stupeň závislosti mezi dvojicí proměnných s vyloučením vlivu ostatních, tedy použít parciálních korelačních koeficientů. V tomto případě je nutné vypočítat parciální korelační koeficienty II. řádu, protože vždy vylučujeme dvě proměnné. Vzhledem k tomu, že výpočet pomocí vzorce 10.25 je velmi zdlouhavý a složitý (je nutno nejprve vypočítat několik parciálních korelačních koeficientů I. řádu a následně počítat koeficienty II. řádu), využijeme maticového vzorce 10.26. Způsob výpočtu pomocí determinantů matic si ukážeme podrobně pro případ závislosti $R_{1,2(3,4)}$, tj. závislost objemu na tloušťce při zkonstantnění (tj. vyloučení vlivu) výšky a délky koruny. K výpočtu potřebujeme matice $R_{(12)}$, $R_{(22)}$ a $R_{(11)}$. Vypočítáme jednotlivé determinanty stejně jako v případě mnohonásobného koeficientu a dosadíme do vzorce

$$R_{1,2(3,4)} = \frac{(-1)^i \cdot \det(\mathbf{R}_{(12)})}{\sqrt{\det(\mathbf{R}_{(11)}) \cdot \det(\mathbf{R}_{(22)})}} = \frac{(-1)^2 \cdot 0.01207}{\sqrt{0.01578 \cdot 0.01314}} = 0.83836$$

Obdobně vypočítáme i další parciální koeficienty:

$R_{1,3(2,4)} = 0.20109$ $R_{2,4(1,3)} = -0.03136$
$R_{1,4(2,3)} = 0.21558$ $R_{3,4(1,2)} = 0.52316$
$R_{2,3(1,4)} = 0.12668$

Číslo stromu	Objem (m^3)	Výčetní tloušťka (cm)	Výška (m)	Délka zelené koruny (m)	Číslo stromu	Objem (m^3)	Výčetní tloušťka (cm)	Výška (m)	Délka zelené koruny (m)
i	v	d	h	k	i	v	d	h	k
1	0.077	12.5	12.0	6.4	26	0.003	6.0	7.8	1.1
2	0.003	6.0	7.4	2.4	27	0.080	12.0	13.5	6.8
3	0.077	12.5	12.2	6.0	28	0.009	7.0	7.8	1.4
4	0.007	7.0	6.8	1.2	29	0.012	7.0	9.6	3.8
5	0.014	7.0	9.9	3.5	30	0.029	8.5	9.9	3.4
6	0.005	6.0	9.0	2.0	31	0.065	12.0	11.2	4.6
7	0.029	9.0	10.0	3.9	32	0.071	12.0	12.0	5.1
8	0.009	7.0	8.0	2.3	33	0.009	7.0	7.9	1.5
9	0.013	8.0	9.8	3.6	34	0.017	8.0	9.0	2.2
10	0.095	13.0	13.5	6.4	35	0.102	13.0	13.5	6.9
11	0.044	10.0	11.0	3.0	36	0.003	6.0	7.8	.4
12	0.034	9.0	10.8	4.1	37	0.048	10.0	12.1	4.6
13	0.115	14.0	13.9	8.3	38	0.049	11.0	10.8	4.3
14	0.021	8.0	10.2	3.6	39	0.014	7.0	10.0	3.7
15	0.025	9.0	9.3	3.1	40	0.061	11.0	12.5	5.9
16	0.132	15.0	14.0	8.6	41	0.098	13.0	13.8	9.5
17	0.011	7.0	9.0	2.8	42	0.071	12.0	12.0	6.7
18	0.048	11.0	10.0	5.0	43	0.011	7.0	8.9	3.9
19	0.010	7.0	8.6	2.8	44	0.017	8.0	9.3	3.5
20	0.132	15.0	14.0	7.6	45	0.023	8.0	10.4	3.6
21	0.065	11.5	11.9	5.5	46	0.065	11.5	12.2	5.8
22	0.029	9.0	10.0	3.4	47	0.077	12.0	13.0	6.0
23	0.065	11.5	11.8	5.6	48	0.091	13.0	13.0	5.3
24	0.071	12.0	12.0	5.3	49	0.059	11.0	12.0	4.8
25	0.012	7.0	9.4	3.0	50	0.036	9.5	10.5	3.7

Tabulka 10.2 – Zadání příkladu 10.2

Je vidět, že použití parciálních korelačních koeficientů ukázalo jiný obrázek o závislostech v tomto vícerozměrném výběru. Z hodnot těchto koeficientů vyplývá, že skutečná příčinná závislost zřejmě bude hlavně mezi objemem stromu a jeho výčetní tloušťkou, ostatní závislosti budou zřejmě statisticky nevýznamné (kromě závislosti mezi výškou a délkou koruny). Všechny hodnoty korelačních koeficientů je nutné testovat, což bude podrobněji popsáno v kapitole 10.7.1.

10.5 Regresní analýza lineárního modelu

10.5.1 Základní tvar lineárního regresního modelu

Základní úlohou regresní analýzy je nalezení vhodného lineárního modelu studované závislosti. V čem tato úloha – ve statistice nazývaná **regresní úloha** – spočívá? Při jejím řešení se snažíme **nahradit každou měřenou** (experimentální, empirickou, zjištěnou) **hodnotu závisle proměnné** (vysvětlované proměnné) ***Y*** **hodnotou teoretickou** (modelovou, vyrovnanou, predikovanou), tj. hodnotou **ležící na spojité funkci** (modelu) **nezávisle proměnné** (vysvětlující proměnné) *X* (***X***) – jeden normální a jeden tučný symbol je zde proto, že nezávisle proměnná X může být jedna veličina nebo matice více veličin.

Při formulaci lineárního regresního modelu (tj. spojité funkce nezávisle proměnné) vycházíme z rovnice 10.3. Tento model je možné rozepsat

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nm} \end{bmatrix}}_{\mathbf{X}} \cdot \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_m \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\boldsymbol{\varepsilon}} \qquad (10.27)$$

což lze zapsat v maticové formě

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (10.28)$$

kde je

y *n x 1*-rozměrná závisle proměnná (je to jeden sloupec *n* měřených hodnot - řádků)

X *n* x *m*-rozměrná nezávisle proměnná (teoreticky nastavované, nenáhodné hodnoty, v praxi ovšem někdy také měřené, je to *m* sloupců po *n* hodnotách)

β *m x 1*-rozměrný vektor regresních koeficientů (ty určují velikost změny závisle proměnné na jednotkové změně nezávisle proměnné, je jich tolik, kolik je nezávislých proměnných – tedy *m*)

ε *n x 1*-rozměrný vektor chyb (vyjadřují tu část celkové variability, která není vysvětlena modelem – každá hodnota má svou chybu, je jich tedy stejně jako hodnot – *n*)

Pokud je počet nezávislých proměnných $m = 1$, potom se jedná o **jednoduchý lineární regresní model** (závislost jedné závisle proměnné na jedné nezávisle pro-

měnné), pokud je $m > 1$, potom se jedná o **mnohonásobný lineární regresní model** (závislost jedné závisle proměnné na několika nezávisle proměnných)[2].

Z výše uvedených vztahů vyplývá, že **podstatou regresní analýzy je**

- **stanovit nejvhodnější tvar regresního modelu** (tedy určit příslušnou rovnici, která bude popisovat závislost Y na **X**)
- **vypočítat jeho parametry** (tj. stanovit konkrétní hodnoty parametrů β).

Jak již bylo uvedeno v kapitole 10.3.2 o formulaci regresních modelů, teoretický („ideální“) model neznáme (ideální model platí pro základní soubor a my ve valné většině případů, prakticky vždy, pracujeme s výběrem, tedy s určitým „výsekem“ základního souboru). Při stanovení konkrétní rovnice modelu máme v podstatě dvě možnosti:

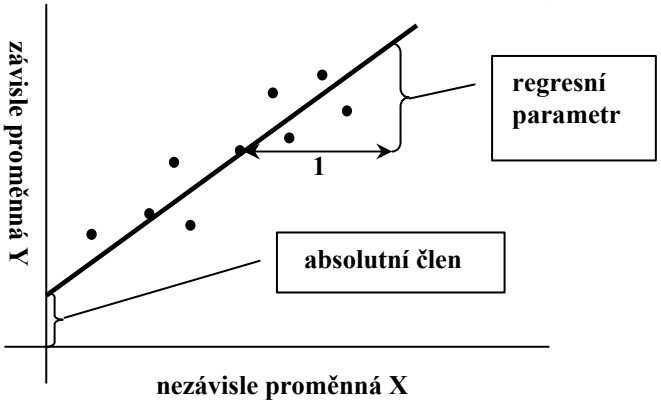
1) Použít model, který **vizuálně a/nebo podle statistických kritérií nejlépe odpovídá měřeným hodnotám**. Vizuální odhad modelu můžeme z technických důvodů aplikovat pouze na jednoduchou regresi (kdy můžeme veličiny Y a X lehce graficky znázornit) nebo maximálně na model se dvěma nezávisle proměnnými (můžeme zobrazit v trojrozměrném grafu). Statistická kritéria vhodnosti modelu, která budou probrána v dalších kapitolách, lze uplatnit i pro mnohonásobné modely. Daleko důležitější podmínkou je to, že **tento typ modelu můžeme uplatnit pouze tehdy, připouští-li povaha řešeného problému jakýkoli tvar modelu a nejsme tedy vázáni omezeními a podmínkami danými reálným systémem**, který modelujeme. Například jestliže modelujeme závislost výšky porostu na věku (růstová funkce), nemůžeme použít přímku, i kdyby naměřená data z růstové řady porostů náhodou nejlépe vyhovovala tomuto „modelu“. Tím bychom připustili, že růst je neukončený, roste nade všechny meze, je stále (v mládí i ve stáří) stejně rychlý a pod. a získali bychom naprosto nesmyslné hodnoty, zvláště predikované. V tomto případě musíme použít některou z růstových funkcí, které splňují požadavky kladené na modelování růstu živých organismů.

2) Nejprve **najít množinu modelů, které svými vlastnostmi vyhovují řešenému problému** (v příkladě uvedeném v bodě 1) by to byly růstové funkce – např. Michajlovova, Korfova, Chapman-Richardsova a další) a **teprve mezi nimi najít podle statistických kritérií ten model, která nejlépe vyhovuje měřeným datům**. Tento přístup je možné považovat za nejlepší, neboť splní jednak požadavek, že model musí být reálný, jednak vyhoví statistickým požadavkům na regresní modely.

Stejně jako tvar modelu, neznáme ani teoretické hodnoty parametrů β. Proto je při praktickém výpočtu modelu nahrazujeme jejich (co možná nejlepšími) odhady b. Musíme tedy použít takovou metodu, která nám umožní získat „co možná nejlepší“ odhady parametrů regresního modelu. K tomuto účelu se obvykle používá **metoda nejmenších čtverců** (viz kapitola 10.5.2).

[2] Kromě těchto případů, kterými se budeme dále podrobněji zabývat, ve statistice existují i metody pro řešení vztahů mezi více závisle proměnnými a více nezávisle proměnnými – v případě, že jak matice závislých, tak i nezávislých proměnných jsou náhodné (obdoba korelačního modelu), řeší tuto úlohu **kanonická korelace**, v případě, že matice nezávislých proměnných je nenáhodná (obdoba regresního modelu), potom se používá **metoda projekce latentních proměnných**. Tyto metody jsou velmi složité teoreticky i interpretačně a k jejich praktickému řešení je potřeba specializovaný software. V tomto učebním textu se jimi nebudeme zabývat.

Nejjednodušším regresním modelem (a také jedním z nejpoužívanějším) je přímka. Její rovnice je $y' = a + bx$, kde koeficient a se nazývá **absolutní člen** a je to souřadnice průsečíku přímku s osou Y (tedy hodnota závisle proměnné Y pro nulovou hodnotu X), koeficient b se nazývá **regresní parametr** a je to hodnota určující sklon přímky (směrnice přímky). Regresní parametr vyjadřuje, o kolik se změní hodnota y, jestliže se x změní o jednotku. Geometrická interpretace je na obrázku 10.9 .



Obrázek 10.9 – Geometrická interpretace členů regresního modelu přímky

10.5.2 Metoda nejmenších čtverců (MNČ)

10.5.2.1 Princip MNČ

MNČ je založena na principu, který je graficky znázorněn na obrázku 10.10 . Vzhledem k možnostem grafického znázornění je princip ukázán pro jednoduchý regresní model. Černými body jsou znázorněny měřené hodnoty, přičemž poloha *i-této* bodu je dána uspořádanou dvojicí $[x_i,y_i]$, přičemž hodnota x_i je nastavovaná, pevná a hodnota y_i je měřená. Těmto hodnotám odpovídají jejich „protějšky“ na regresní čáře, které jsou zobrazeny bílými kroužky. Jsou to hodnoty, které byly vypočítány pomocí použitého regresního modelu a označují se y'_i. Pro každou hodnotu x_i mohu pomocí rovnice regresního modelu vypočítat hodnotu y'_i. Naší snahou je, aby rozdíly mezi měřenou hodnotou y_i a vypočítanou (modelovou) hodnotou y'_i byly co nejmenší, tj. aby model co nejlépe prokládal měřená data.

To se nám povede, jestliže nalezneme takový tvar regresní funkce, který minimalizuje hodnotu součtu čtverců (druhých mocnin) odchylek skutečných (měřených) a modelem vypočtených hodnot závisle proměnné Y, podle vztahu

$$\sum_{i=1}^{n}\left(y_i - y'_i\right)^2 = \min. \qquad (10.29)$$

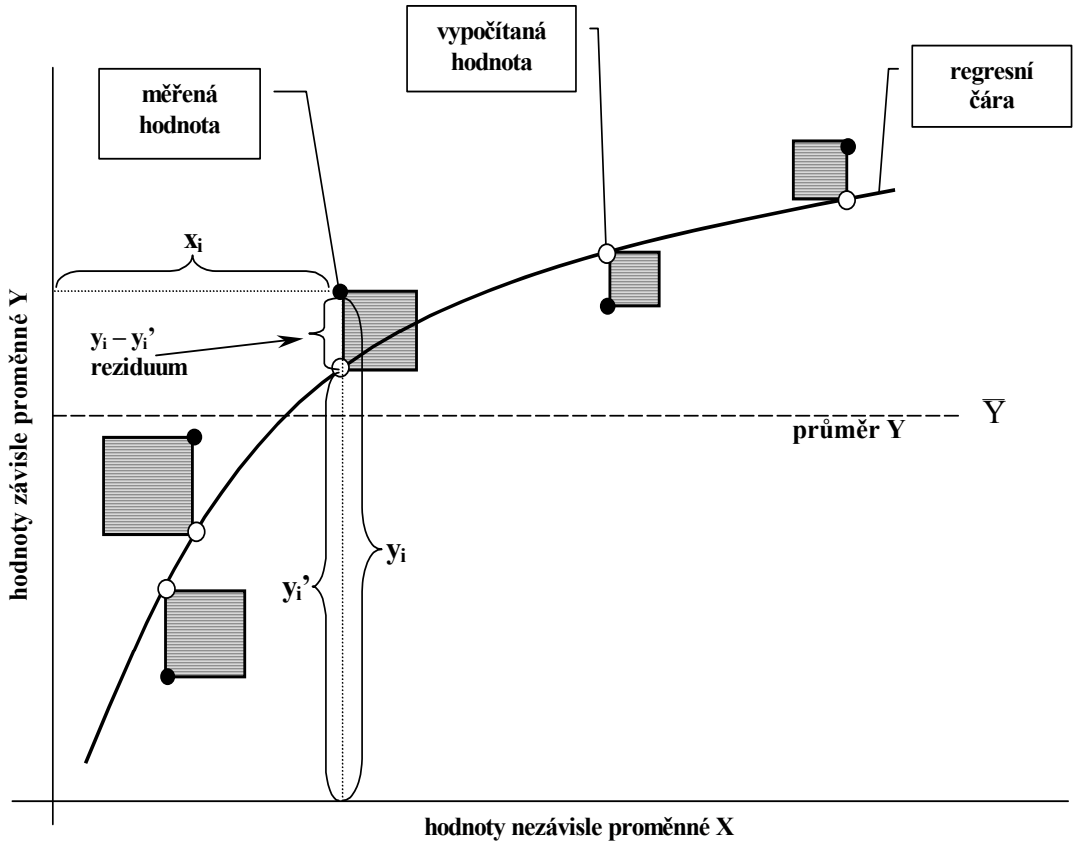
Rozdíl $y_i - y_i'$ se nazývá **reziduum** (je to tedy **rozdíl mezi měřenou a modelovou hodnotou**). Toto kritérium by se tedy mělo přesněji nazývat kritériem nejmenšího součtu reziduálních čtverců. V grafickém vyjádření podle obrázku 10.10 musíme minimalizovat plochu vodorovně vyšrafovaných čtverců.

Je nutné podotknout, že tato metoda nehledá absolutně nejlepší matematický model, ale nejlepší z dané třídy modelů (např. nejlepší přímku, parabolu, …). Volba nejlepší třídy modelu je na statistikovi.

Z matematického hlediska nalezneme minimum (tj. extrém funkce) tak, že provedeme postupně parciální derivace podle všech parametrů. Ukážeme si postup pro nejjednodušší případ – přímku.

Regresní model přímky má tvar $y_i' = a + b \cdot x_i$. Pokud dosadíme tento výraz do vztahu 10.29 místo y_i', dostaneme výraz

$$\sum_{i=1}^{n} \left(y_i - a + b \cdot x_i \right)^2 = \min.$$



Obrázek 10.10 – Grafické znázornění principu MNČ (podle MINAŘÍK 1995)

Provedeme parciální derivaci podle *a* a potom podle *b*

$$\frac{\partial \sum_{i=1}^{n} \left(y_i - a + b \cdot x_i \right)^2}{\partial a} = 2 \sum_{i=1}^{n} \left(y_i - a - b \cdot x_i \right) \cdot \left(-1 \right) = 0$$

$$\frac{\partial \sum_{i=1}^{n} \left(y_i - a + b \cdot x_i \right)^2}{\partial b} = 2 \sum_{i=1}^{n} \left(y_i - a - b \cdot x_i \right) \cdot \left(-x_i \right) = 0$$

Úpravou těchto vztahů získáme normální rovnice přímky ve tvaru

$$\sum_{i=1}^{n} y_i = n \cdot a + b \sum_{i=1}^{n} x_i \qquad (10.30)$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \qquad (10.31)$$

Normální rovnice přímky tvoří soustavu dvou lineárních rovnic o dvou neznámých *a, b*, kterou můžeme řešit známými metodami lineární algebry (např. pomocí determinantů).

Nevýhodou tohoto postupu je fakt, že soustavy normálních rovnic musíme zvlášť sestavovat a řešit pro každou třídu modelů (pro přímky, paraboly, hyperboly, …) a také pro různý počet nezávisle proměnných. Proto je vhodné používat **obecné maticové vyjádření MNČ**, jejíž hlavní výhodou je naprostá univerzálnost použití bez ohledu na typ použitého **lineárního** modelu a počtu nezávislých proměnných.

Normální rovnice přímky 10.30 a 10.31 (stejně jako jakékoliv jiné normální rovnice) můžeme přepsat

$$\underbrace{\begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}}_{g} = \underbrace{\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}}_{A} \cdot \underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{b} \qquad (10.32)$$

do maticového zápisu

$$g - A \cdot b = 0 \qquad (10.33)$$

Jednotlivé členy rovnice 10.33 můžeme vypočítat takto

$$\mathbf{g} = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} = \mathbf{X^T} \cdot \mathbf{y} \qquad (10.34)$$

$$\mathbf{A} = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \cdot \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} = \mathbf{X^T} \cdot \mathbf{X} \qquad (10.35)$$

Jestliže dosadíme do maticového zápisu 10.33 pravé strany vztahů 10.34 a 10.35, dostaneme

$$\mathbf{X^T y} - \mathbf{X^T X} \cdot \mathbf{b} = \mathbf{0} \qquad (10.36)$$

z čehož jednoduchou úpravou získáme **obecný výraz pro výpočet vektoru regresních koeficientů *b***

$$\mathbf{b} = \left(\mathbf{X^T} \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X^T} \cdot \mathbf{y} \qquad (10.37)$$

Ve výše uvedených vzorcích výraz $\mathbf{X^T}$ znamená transpozici matice **X,** výraz $\mathbf{X^{-1}}$ znamená inverzi matice **X.** Matice **X** je matice nezávisle proměnných (jestliže počítáme tzv. absolutní člen – v rovnici přímky *a* – pak musíme přidat vektor jedniček),

vektor **y** je řada měřených hodnot závisle proměnné (viz rovnici 10.27). Velkou výhodou výpočtu pomocí výrazu 10.37 je jednak jeho již zmíněná univerzálnost a také fakt, že potřebné maticové operace (násobení, transpozice a inverze) zvládají bez problémů běžné tabulkové kalkulátory (např. Excel nebo Quattro Pro), takže není naprosto problémem pomocí nich velmi rychle vypočítat regresní koeficienty jakéhokoliv lineárního regresního modelu bez nutnosti provádět derivace a sestavovat soustavy normálních rovnic.

Pomocí maticových operací se mohou také přímo vypočítat modelové hodnoty y_i'. Použije se výraz

$$\mathbf{y}' = \mathbf{X} \cdot \mathbf{b} \tag{10.38}$$

kdy se za **b** dosadí vztah 10.37 a získáme

$$\mathbf{y}' = \mathbf{X}\left(\mathbf{X}^{\mathbf{T}} \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}^{\mathbf{T}} \cdot \mathbf{y} \tag{10.39}$$

kde výraz

$$\mathbf{X}\left(\mathbf{X}^{\mathbf{T}} \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}^{\mathbf{T}} \tag{10.40}$$

se nazývá **projekční matice H.** Tato matice se nazývá „projekční“ proto, že je schopna promítnout libovolný vektor do „roviny“ nezávisle proměnných, tj. z měřených hodnot **y** stanovit modelové hodnoty **y'.** Má také značné použití v tzv. regresní diagnostice.

Příklad 10.3:

Při výzkumu závislosti tloušťky kůry na různých faktorech byl také zkoumán vztah mezi tloušťkou kůry (Y), výčetní tloušťkou (X_1) a věkem (X_2). Předpokládáme lineární regresní model $y' = a + b_1x_1 + b_2x_2$. Pomocí metody nejmenších čtverců stanovte parametry tohoto regresního modelu a modelové hodnoty. Měřené hodnoty jsou v tabulce 10.3. Pro jednoduchost řešení a možnost znázornění matic se výpočet provede pouze pro 10 měření.

Výčetní tloušťka (cm)	Věk (roky)	Tloušťka kůry (cm)
X1	**X2**	**Y**
19.40	56	0.46
21.40	62	0.66
21.90	72	1.34
26.40	73	1.83
28.70	77	2.06
28.80	77	2.20
29.10	85	2.26
31.10	86	2.43
31.60	86	2.43
35.60	89	2.79

Tabulka 10.3 - Měřené hodnoty tloušťky kůry, výčetní tloušťky a věku.

Ze zadání vyplývá, že musíme použít vztah 10.37 pro výpočet parametrů **b** a 10.39 pro výpočet modelových (predikovaných) hodnot. Využijeme možnosti násobení matic, které poskytují moderní tabulkové kalkulátory.

Nejdříve vypočítáme výraz $\mathbf{b} = (\mathbf{X}^{\mathbf{T}}\ \mathbf{X})^{-1}\ \mathbf{X}^{\mathbf{T}}\ \mathbf{y}$. Je nutné si uvědomit, že musíme zachovat vzájemné postavení matic při násobení tak, jak je uvedeno ve vzorci (u matic rozli-

šujeme násobení zprava a násobení zleva). Transpozice matice znamená záměnu řádků a sloupců matice. Inverze matice je nalezení takové matice, jejíž součin s původní maticí dá jednotkovou matici (prvky hlavní diagonály jsou rovny jedné).

Vzhledem k tomu, že regresní model uvažuje i absolutní člen, přidáme k matici **X** jednotkový vektor, takže výsledná matice nezávisle proměnných bude mít podobu

$$\mathbf{X^T X} = \begin{pmatrix} 19{,}4 & 21{,}4 & 21{,}9 & 26{,}4 & 28{,}7 & 28{,}8 & 29{,}1 & 31{,}1 & 31{,}6 & 35{,}6 \\ 56 & 62 & 72 & 73 & 77 & 77 & 85 & 86 & 86 & 89 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 19{,}4 & 56 & 1 \\ 21{,}4 & 62 & 1 \\ 21{,}9 & 72 & 1 \\ 26{,}4 & 73 & 1 \\ 28{,}7 & 77 & 1 \\ 28{,}8 & 77 & 1 \\ 29{,}1 & 85 & 1 \\ 31{,}1 & 86 & 1 \\ 31{,}6 & 86 & 1 \\ 35{,}6 & 89 & 1 \end{pmatrix} \tag{10.41}$$

Výsledkem je čtvercová matice

$$\begin{matrix} 7743.96 & 21378.8 & 274 \\ 21378.8 & 59289 & 763 \\ 274 & 763 & 10 \end{matrix} \tag{10.42}$$

Z této matice určíme inverzní matici stejné velikosti $\mathbf{(X^T X)^{-1}}$

$$\begin{matrix} 0.036 & -0.016 & 0.222 \\ -0.016 & 0.008 & -0.169 \\ 0.222 & -0.169 & 6.917 \end{matrix} \tag{10.43}$$

a vynásobením této inverzní matice původní transponovanou maticí získáme výraz $\mathbf{(X^T X)^{-1} X^T}$

$$\begin{matrix} 0.034 & 0.011 & -0.129 & 0.016 & 0.035 & 0.039 & -0.076 & -0.021 & -0.003 & 0.093 \\ -0.034 & -0.018 & 0.053 & -0.010 & -0.015 & -0.017 & 0.042 & 0.018 & 0.010 & -0.029 \\ 1.756 & 1.185 & -0.396 & 0.436 & 0.271 & 0.293 & -0.994 & -0.718 & -0.607 & -0.225 \end{matrix} \tag{10.44}$$

Dalším krokem je vynásobení této matice závisle proměnnou **y** a výsledkem je vektor parametrů $\mathbf{b = (X^T X)^{-1} X^T y}$ v pořadí b_1, b_2, a

$$\begin{matrix} 0.069 \\ 0.039 \\ -3.037 \end{matrix} \tag{10.45}$$

Zjistili jsme tedy, že a = -3.037, $b_1 = 0.069$, $b_2 = 0.039$, a tedy regresní model má tvar

$$y = -3.037 + 0.069.x_1 + 0.039.x_2$$

Při výpočtu projekční matice (se kterou se ještě několikrát v dalším textu setkáme, zvláště v části věnované regresní diagnostice) využijeme matice 10.44, kterou vynásobíme zleva maticí **X**. Výsledkem je projekční matice **H (**zvýrazněny jsou diagonální prvky matice, protože mají zvláštní důležitost v detekci vlivných bodů i při dalších výpočtech):

$$\begin{array}{cccccccccc}
\mathbf{0.516} & 0.381 & 0.059 & 0.178 & 0.120 & 0.124 & -0.137 & -0.103 & -0.086 & -0.052 \\
0.381 & \mathbf{0.294} & 0.118 & 0.149 & 0.101 & 0.102 & -0.039 & -0.036 & -0.030 & -0.041 \\
0.059 & 0.118 & \mathbf{0.581} & 0.055 & -0.030 & -0.043 & 0.340 & 0.135 & 0.071 & -0.285 \\
0.178 & 0.149 & 0.055 & \mathbf{0.118} & 0.114 & 0.116 & 0.039 & 0.061 & 0.069 & 0.103 \\
0.120 & 0.101 & -0.030 & 0.114 & \mathbf{0.136} & 0.139 & 0.030 & 0.086 & 0.104 & 0.200 \\
0.124 & 0.102 & -0.043 & 0.116 & 0.139 & \mathbf{0.143} & 0.023 & 0.084 & 0.103 & 0.210 \\
-0.137 & -0.039 & 0.340 & 0.039 & 0.030 & 0.023 & \mathbf{0.333} & 0.223 & 0.184 & 0.005 \\
-0.103 & -0.036 & 0.135 & 0.061 & 0.086 & 0.084 & 0.223 & \mathbf{0.200} & 0.189 & 0.161 \\
-0.086 & -0.030 & 0.071 & 0.069 & 0.104 & 0.103 & 0.184 & 0.189 & \mathbf{0.188} & 0.208 \\
-0.052 & -0.041 & -0.285 & 0.103 & 0.200 & 0.210 & 0.005 & 0.161 & 0.208 & \mathbf{0.492}
\end{array} \qquad (10.46)$$

Pokud matici 10.46 vynásobíme vektorem **y**, získáme vektor predikce $\mathbf{y_P}$, tj. modelové (vyrovnané) hodnoty regresního modelu (obvykle označované)

$$\underbrace{\begin{pmatrix} 0.516 & \cdots & \cdots & -0.052 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ -0.052 & \cdots & \cdots & 0.492 \end{pmatrix}}_{\textbf{projekční matice}} \cdot \underbrace{\begin{pmatrix} 0.46 \\ \vdots \\ \vdots \\ 2.79 \end{pmatrix}}_{\substack{\textbf{závisle} \\ \textbf{proměnná}}} = \underbrace{\begin{pmatrix} 0.50 \\ 0.87 \\ 1.30 \\ 1.65 \\ 1.96 \\ 1.97 \\ 2.30 \\ 2.48 \\ 2.52 \\ 2.91 \end{pmatrix}}_{\textbf{vypočítané hodnoty modelu}} \qquad (10.47)$$

10.5.2.2 Předpoklady metody nejmenších čtverců

Metoda nejmenších čtverců má optimální vlastnosti za dodržení těchto předpokladů:

1) Regresní parametry β mohou teoreticky nabývat jakýchkoli hodnot (existují ovšem omezení daná povahou problému, který je regresním modelem řešen).
2) Regresní model je lineární v parametrech.
3) Matice nezávisle proměnných **X** má hodnost rovnou *m*. To znamená, že žádné dva její sloupce nejsou rovnoběžné (kolineární) vektory, tedy mezi nezávislými proměnnými nedochází k tzv. **multikolinearitě**.
4) Náhodné chyby mají nulovou střední hodnotu $E(\varepsilon_i) = 0$, konstantní a konečný rozptyl $E(\varepsilon_i^2) = 0$ a jsou nekorelované. Také podmíněný rozptyl $D(y/x) = \sigma^2$ je konstantní (tzv. podmínka **homoskedasticity**).

Pokud jsou tyto podmínky splněny, potom jsou odhady **b**, získané metodou nejmenších čtverců, **nejlepší nevychýlené lineární odhady** regresních parametrů (MELOUN - MILITKÝ 1994):

- ***nejlepší odhady*** **b** jsou proto, že jejich libovolná lineární kombinace má nejmenší rozptyl ze všech nevychýlených lineárních odhadů a také odhady rozptylů jednotlivých regresních koeficientů jsou minimální ze všech možných nevychýlených odhadů (mohou existovat vychýlené odhady s nižším rozptylem),

- ***nevychýlené odhady*** **b** jsou proto, že platí E(**β - b**) = 0, jinými slovy, střední hodnota vektoru odhadů E(**b**) je rovna vektoru regresních parametrů **β**.

Je nutné si uvědomit, že z určitého základního souboru můžeme provést teoreticky nekonečně mnoho výběrů stejného rozsahu a vždy vyjdou poněkud jiné hodnoty regresních koeficientů. Tedy i regresní koeficienty jsou náhodnou veličinou, pro kterou můžeme počítat běžné statistické charakteristiky, tedy i střední hodnotu nebo rozptyl, a také je můžeme testovat a počítat pro ně intervalové odhady.

Z podmínek MNČ si bližší vysvětlení zaslouží body 3. a 4. – především výklad pojmů multikolinearita a homoskedasticita.

Jednou ze základních podmínek řešení regresního modelu metodou nejmenších čtverců je to, že nezávislé (vysvětlující) proměnné nejsou nezávislé jen podle názvu, ale jsou skutečně vzájemně nezávislé. Tento předpoklad však nebývá často splněn. Jev, kdy v lineárním regresním modelu existuje závislost mezi vysvětlujícími proměnnými, se nazývá **multikolinearita**. Podrobné teoretické zdůvodnění podstaty tohoto jevu viz např. v MELOUN - MILITKÝ 1994. Tento jev způsobuje při řešení a interpretaci problémy dvojího druhu - statistické a numerické (výpočetní).

Mezi statistické problémy patří:

- **nelze odděleně sledovat skutečný vliv jednotlivých vysvětlujících vstupních proměnných na vysvětlovanou (závislou) proměnnou** – skutečné vztahy mezi nezávislými proměnnými a závislou proměnnou je v tomto případě často „maskován“ vztahy mezi „nezávislými“ proměnnými;
- **nestabilita odhadů regresních parametrů** - hodnota odhadů je velmi citlivá i na malé změny v datech (např. přidání bodu nebo malá chyba měření), což může vést např. k tomu, že odhady mají nesprávné znaménko, což znemožňuje jejich správnou věcnou interpretaci;
- **velké rozptyly odhadů regresních parametrů** - může nastat paradoxní situace, že model jako celek je vysoce významný, ale všechny jednotlivé regresní koeficienty nevýznamné (podrobněji v kapitole týkající se testování regresního modelu).

Mezi výpočetní problémy patří:

- multikolinearita způsobuje **špatnou podmíněnost matice $\mathbf{X}^T\mathbf{X}$**, což má za následek, že determinant této matice je nula nebo číslo blízké nule
- tyto skutečnosti způsobují **potíže při invertaci matice**, takže takovýto regresní model není jednoznačně řešitelný (singularita matice).

Mezi hlavní příčiny multikolinearity patří:

- **přeurčenost regresního modelu** - regresní model má zbytečně mnoho nezávisle proměnných, z nichž některé jsou lineární kombinací jiných, a tedy jsou v modelu zbytečné, nijak nepřispějí k určení hodnoty *y* ze známé hodnoty *x*. Ve statistických programech existují postupy, které jsou schopny určit správný počet nezávisle proměnných – např. kroková regrese, podrobný výklad těchto postupů je nad rámec tohoto učebního textu – přebytečné proměnné je nutné z modelu odstranit;
- **nevhodné rozmístění experimentálních bodů** - vznikají buď z neplánovitých nebo špatně postavených experimentů, kdy hodnoty vysvětlujících proměnných mají příliš malou variabilitu, takže i malá odchylka v měření může způsobit např. „obrácení“ regresní čáry (z kladné korelace se stane zá-

porná) tuto situaci ukazuje obrázek 10.11 , kde vlevo je schématicky zachycena situace, kdy se u dvou bodů nezávisle proměnné X, které jsou velmi blízko u sebe (tedy proměnná X má velmi malou variabilitu) vyskytla určitá experimentální chyba (naznačena šipkami). Tato chyba může být tak malá, že takto naměřené hodnoty jsou považovány za správné (správné hodnoty – černé tečky, hodnoty s chybou - černé čtverečky). Výsledkem je úplné „obrácení“ smyslu modelu (správný model – slabá plná čára, nesprávný model - silná plná čára). Na pravém obrázku je zachycen vliv chyb téže velikosti na body, které jsou v „rozumné“ vzdálenosti, tj. proměnná X má přiměřenou variabilitu. Zde chyba způsobí změnu směrnice, ale smysl modelu je zachován. Abychom u takto rozmístěných bodů „dosáhli“ obrácení smyslu modelu (naznačeno čárkovanou čarou), museli bychom se dopustit nepravděpodobně velké chyby (naznačena čárkovanou šipkou), kterou bychom jistě odhalili.

- **povaha modelu** - v některých typech modelů, např. polynomech, se vyskytuje multikolinearita prakticky vždy, což je v tomto případě dáno už strukturou modelu.

V této souvislosti je nutné zdůraznit, že multikolinearita nemusí „vadit“ vždy. Pokud při regresní analýze jde jen o „vyhlazení“ experimentálních dat a nikoli o postižení skutečných závislostí mezi proměnnými, zůstává problémem jen numerické hledisko. Pokud ovšem je naším cílem rozkrytí vazeb v regresním modelu a zjištění těch proměnných, které významně přispívají k objasnění variability závisle proměnné, potom je multikolinearita působí vážné potíže. Ovšem i v tomto případě je skutečným problémem pouze silná multikolinearita, kdy silné závislosti mezi vysvětlujícími proměnnými „přehluší“ skutečné vazby mezi vysvětlovanou a vysvětlujícími proměnnými. Závažnost multikolinearity se testuje speciálními metodami, které jsou obsaženy ve statistických programech.

Dalším problémem spojeným s výpočtem modelu pomocí MNČ je problém **konstantního rozptylu dat (homoskedasticity**). MNČ vyžaduje, aby hodnoty y měly v celém rozsahu hodnot x konstantní variabilitu (jako na obrázku 10.12 – měřené hodnoty jsou jako by mezi dvěma myšlenými rovnoběžkami). Pokud tomnu tak není (schématické znázornění je na obrázku 10.13), jedná se o nekonstantní rozptyl – **heteroskedasticitu**. Tento jev se vyznačuje tím, že rozptyl měření se pro různé hodnoty x výrazně mění a „mrak“ bodů získává tvar klínu. Příčinou heteroskedasticity bývá obvykle změna podmínek nebo nedodržení postupu měření, porucha přístroje, apod.

Diagnostické nástroje k určení míry multikolinearity a heteroskedasticity budou uvedeny v kapitolách týkajících se testování a regresní diagnostiky.

Další podmínky týkající se chyb a reziduí – normalita, nezávislost – se testují běžnými metodami (test normality, autokorelace).

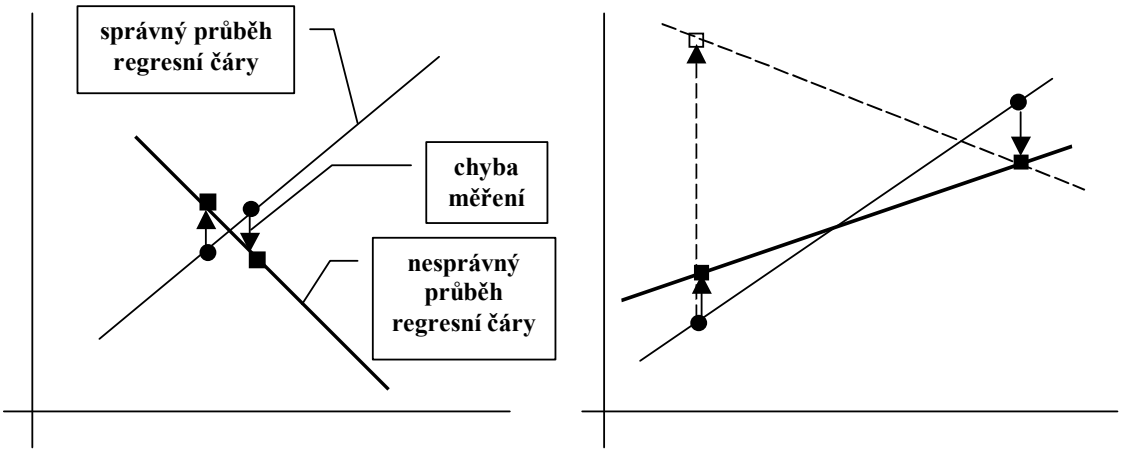
10.6 Intervalové odhady parametrů korelace a regrese

V případě, že pracujeme s výběry, jsou statistiky vypočítané pomocí korelační analýzy a MNČ (např. korelační koeficienty, regresní koeficienty, apod.) vlastně bodovými odhady příslušných parametrů základního souboru. Jak již bylo uvedeno, je tomu tak proto, že ze základního souboru můžeme teoreticky vytvořit nekonečně mnoho výběrů a jejich vypočítané statistiky se budou pro jednotlivé výběry poněkud

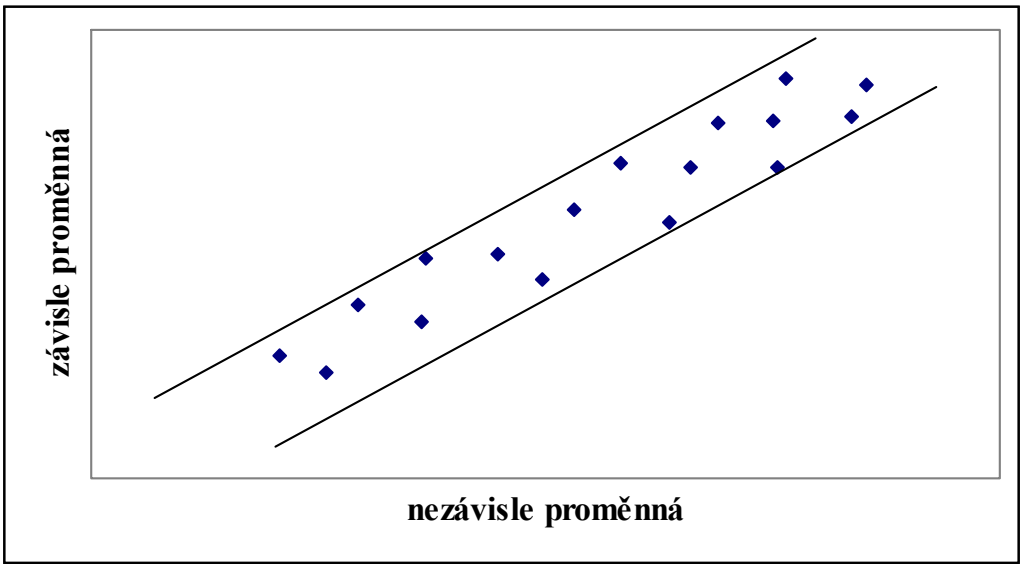
lišit. Je proto nutné tyto výběrové statistiky zobecnit pro základní soubor a tedy vypočítat intervalové odhady, ve kterých se budou posuzované parametry nacházet s předem zvolenou pravděpodobností. Z hlediska řešení konkrétních problémů mohou mít tyto intervalové odhady vyšší důležitost než samotné vypočítané parametry regresního modelu.

10.6.1 Intervalový odhad korelačního koeficientu

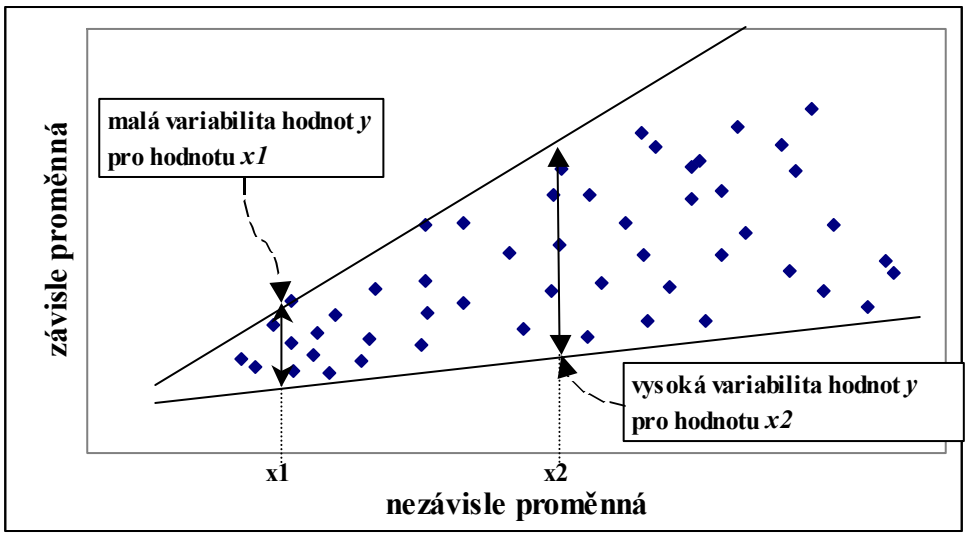
V této kapitole budeme označovat výběrový korelační koeficient R (vypočítaný přímo ze zadaných hodnot) a korelační koeficient základního souboru (nám neznámý) symbolem ρ.



Obrázek 10.11 – Vliv malé variability (vlevo) a přiměřené variability (vpravo) nezávisle proměnné na chybu regrese (podle MINAŘÍK 1995)



Obrázek 10.12 – Schématické znázornění dat s konstantní variabilitou (homoskedastická data)



Obrázek 10.13 - Schématické znázornění dat s nekonstantní variabilitou (heteroskedastická data)

Při konstrukci intervalového odhadu korelačního koeficientu vycházíme z poznatku, že rozdělení náhodné veličiny *R* není v běžných případech normální. Proto se pro tento odhad nepoužívají přímo hodnoty výběrového korelačního koeficientu, ale používáme Fisherovu transformaci

$$Z(R) = \operatorname{arctgh}(R) = 0.5 \ln \frac{1+R}{1-R} \qquad (10.48)$$

která má přibližně normální rozdělení se střední hodnotou E(Z) = Z(ρ) a rozptylem D(Z) = 1/(n-3).

Pomocí Fisherovy transformace se vypočítá transformovaný intervalový odhad

$$Z(\rho) = Z(R) \pm z_{1-\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{n-3}} \qquad (10.49)$$

kde $z_{1-\alpha/2}$ kvantil normovaného normálního rozdělení. Tyto transformované hranice se pomocí vztahu 10.48 retransformují na původní hodnoty R (transformace i retransformace se provádí buď pomocí tabulek – viz Tabulka 6 v I. dílu nebo pomocí funkcí Excelu FISHER(R), resp. FISHERINV(Z(R)).
Pro vyšší rozsah výběru (n > 50) je možné použít statistiku

$$\rho = R \pm t_{1-\frac{\alpha}{2},n-2} \cdot \sqrt{\frac{1-R^2}{n-2}} \qquad (10.50)$$

kde $t_{1-\alpha/2,n-2}$ je kvantil Studentova rozdělení pro (n-2) stupňů volnosti.

Pro velké rozsahy výběrů (n > 500) je možné použít vztahu

$$\rho = R \pm z_{1-\frac{\alpha}{2}} \cdot \frac{1-R^2}{\sqrt{n-1}} \qquad (10.51)$$

Tyto vztahy je možné použít i pro parciální korelační koeficienty s tím, že počet stupňů volnosti se stanoví výrazem f = n - k - 2, kde *k* je počet proměnných, které považujeme za konstantní a pro vícenásobný korelační koeficient s počtem stupňů volnosti (n - m).

Pro Spearmanův korelační koeficient se také používají vztahy 10.48 a 10.49 s tím, že někteří autoři (např. ZAR 1984)se upraví výraz pro výpočet rozptylu *Z* a výsledný vztah je

$$Z(\rho_s) = Z(R_s) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{1.06}{n-3}} \qquad (10.52)$$

Příklad 10.4:

Vypočítejte intervalové odhady korelačních koeficientů podle příkladu 10.1.

Odhad **Pearsonova korelačního koeficientu** (jeho vypočítaná hodnota je 0.9338) uděláme pomocí Fisherovy transformace

R = 0. 9338 $\Rightarrow$ Z(R) = 1.6873, z čehož plyne intervalový odhad podle vzorce 10.49

$$Z(\rho) = 1.6873 \pm 1.96 * \frac{1}{\sqrt{20-3}} = 1.6873 \pm 0.475 = \langle 1.2123;\ 2.1623 \rangle,$$

kde 1.96 je kvantil normovaného normálního rozdělení $u_{0.05}$. Hodnoty Fisherovy transformace se retransformují na původní hodnoty R a vyjde interval ⟨0.837; 0.974⟩. Fisherovu transformaci i retransformaci je možné provést podle statistických tabulek nebo pomocí funkcí Excelu (funkce FISHER pro převod na R $\rightarrow$ Z, resp. FISHERINV pro převod Z $\rightarrow$ R).

Odhad **Spearmanova korelačního koeficientu** se provede stejným způsobem a výsledkem je interval ⟨0.842; 0.975⟩. Pokud by se použil vzorec 10.52, výsledek se změní jen nepatrně na interval ⟨0.838; 0.975⟩.

10.6.2 Intervalové odhady regresních koeficientů

Interval spolehlivosti pro parametr β_J se stanoví

$$\beta_j = b_j \pm t_{\frac{\alpha}{2},n-m} \cdot \sqrt{D(b_j)} \qquad (10.53)$$

kde $D(b_j)$ je rozptyl parametru b_j vypočítaný podle vzorce.

$$D(b_j) = \sigma^2 \cdot c_{jj} \qquad (10.54)$$

kde je

cjj *j*-tý diagonální prvek matice $(\mathbf{X}^T \mathbf{X})^{-1}$

σ^2 odhad reziduálního rozptylu, který se vypočítá

$$\sigma^2 = \frac{\sum_{i=1}^{n} (y_i - y'_i)^2}{n-m} \qquad (10.55)$$

kde je

y_i měřená (experimentální) *i*-tá hodnota závisle proměnné

y'_i vypočítaná (modelová) *i*-tá hodnota závisle proměnné

m počet parametrů modelu (včetně absolutního členu, pokud je obsažen v modelu)

Pro nejběžnější model – přímku $y = a + bx$ – je možné použít vztah 10.53 s tím, že se jako odhad směrodatné odchylky parametrů použije pro absolutní člen a vztah (ŠMELKO 1991)

$$\sqrt{D(a)} = \frac{s_{yx}}{\sqrt{n-2}} \cdot \sqrt{1 + \frac{\bar{x}^2}{s_x^2}} \qquad (10.56)$$

a pro regresní parametr b

$$\sqrt{D(b)} = \frac{s_{xy}}{s_x \sqrt{n-2}} \qquad (10.57)$$

kde je

s_x směrodatná odchylka nezávislé (vysvětlující) proměnné
$\bar{x}$ aritmetický průměr nezávislé (vysvětlující) proměnné
s_{yx} směrodatná odchylka reziduí

Příklad 10.5:

Stanovte intervalové odhady regresních parametrů pro data z příkladu 10.1

Nejdříve musíme stanovit vhodný regresní model. Vzhledem k charakteru bodového pole tohoto dvourozměrného výběru (viz obrázek 10.7) se rozhodneme pro model přímky. Pomocí MNČ vypočítáme regresní koeficienty a = -2.752 a b = 0.676. Regresní model má tedy tvar y = -2.752 + 0.676 x, kde x je „délka listů“ a y je „šířka listů“.

Směrodatné odchylky parametrů stanovíme podle vztahu 10.54. Nejprve musíme vypočítat matici $\mathbf{X^T X}$ **(**matice **X** jsou tomto případě měřené hodnoty „délka listů“, ke kterým se musí přidat vektor jedniček kvůli výpočtu absolutního členu) postupem uvedeným v kapitole 10.5.2.1 a výsledná matice má tvar

$$\begin{matrix} 39649 & 865 \\ 865 & 20 \end{matrix}$$

dále k ní inverzní matici $\mathbf{(X^T X)^{-1}}$**,** jejíž diagonální prvky (c_{11} a c_{22}) jsou zvýrazněny

$$\begin{matrix} \mathbf{0.000447} & -0.01933 \\ -0.01933 & \mathbf{0.885912} \end{matrix}$$

Pomocí odmocniny ze vztahu 10.55 vypočítáme reziduální směrodatnou odchylku, který má hodnotu 2,9 a poté úpravou vztahu 10.54 ($\sigma \cdot \sqrt{c_{jj}}$) vypočítáme směrodatné odchylky obou parametrů $s_a = 2.729$ a $s_b = 0.0613$. Pomocí těchto údajů můžeme stanovit na základě vztahu 10.53 intervalové odhady regresních koeficientů

$$\beta_1 = -2.752 \pm 2.101 * 2.729 = -2.752 \pm 5.734$$
$$\beta_2 = 0.676 \pm 2.101 * 0.0613 = 0.676 \pm 0.129$$

kde 2.101 je kvantil Studentova rozdělení $t_{0,025;18}$.

Výsledkem je intervalový odhad absolutního členu ⟨-8.486, 2.982⟩ a regresního parametru ⟨0.548,0.805⟩. Interpretačně zajímavý je odhad absolutního členu. Ten totiž

obsahuje nulu (dolní hranice je záporná, horní hranice je kladná), což znamená, že v základním souboru nemůžeme vyloučit, že absolutní člen je nulový. To vede k závěru, že absolutní člen je v tomto modelu statisticky nevýznamný a může být z modelu vypuštěn. Konečný tvar modelu bude $y = bx$ (parametr b se musí znovu vypočítat).

10.6.3 Intervalový odhad regresního modelu

Podobně jako pro odhady parametrů **b** lze konstruovat interval spolehlivosti i pro regresní model (tj. pro vypočítanou hodnotu y'_i) v místě $\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{im})$ podle vztahu (MELOUN - MILITKÝ 1994)

$$\mathbf{x_i}^T\mathbf{b} = \mathbf{x_i}^T\mathbf{b} \pm t_{1-\frac{\alpha}{2},n-m} \cdot \sigma \cdot \sqrt{\mathbf{x_i}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x_i}} \qquad (10.58)$$

Pro model jednoduché korelace vyjádřené přímkou se dá vzorec přepsat do tvaru (ŠMELKO 1991)

$$\mu_{y'} = y'_i \pm t_{\frac{\alpha}{2},n-2} \cdot \frac{\sigma}{\sqrt{n-2}} \cdot \sqrt{1 + \frac{n(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (10.59)$$

10.6.4 Intervalový odhad měřených hodnot (pás spolehlivosti)

Kromě intervalového odhadu modelu je možné ještě vypočítat tzv. **pás spolehlivosti** měřených (empirických) hodnot, který udává rozpětí, ve kterém se budou v základním souboru nacházet hodnoty závisle (vysvětlované) proměnné se zvolenou pravděpodobností. Stanoví se podle vztahu (GROFÍK 1987)

$$y_{i(min,max)} = y'_i \pm t_{\frac{\alpha}{2};n-m} \cdot \sigma \qquad (10.60)$$

Příklad 10.6:

Stanovte intervalové odhady modelu a měřených hodnot pro data z příkladu 10.1.

Intervalový odhad modelu stanovíme podle vzorce 10.59 a pás spolehlivosti podle vzorce 10.60. Použijeme výsledků příkladu 10.5 ($t_{0.025;18} = 2.101$, $\sigma = 2.9$) a vypočítáme oba odhady. Číselné výsledky jsou v tabulce 10.4 a grafické znázornění na obrázku **10.14** .

Výsledky se dají interpretovat následujícím způsobem:

- vypočítaná (modelová) hodnota platí pro konkrétní výběr (v našem případě pro námi měřených 20 listů), tj. např. pro délku listu 24 mm (hodnota 1) bude vypočítaná šířka listu 13.5 mm;

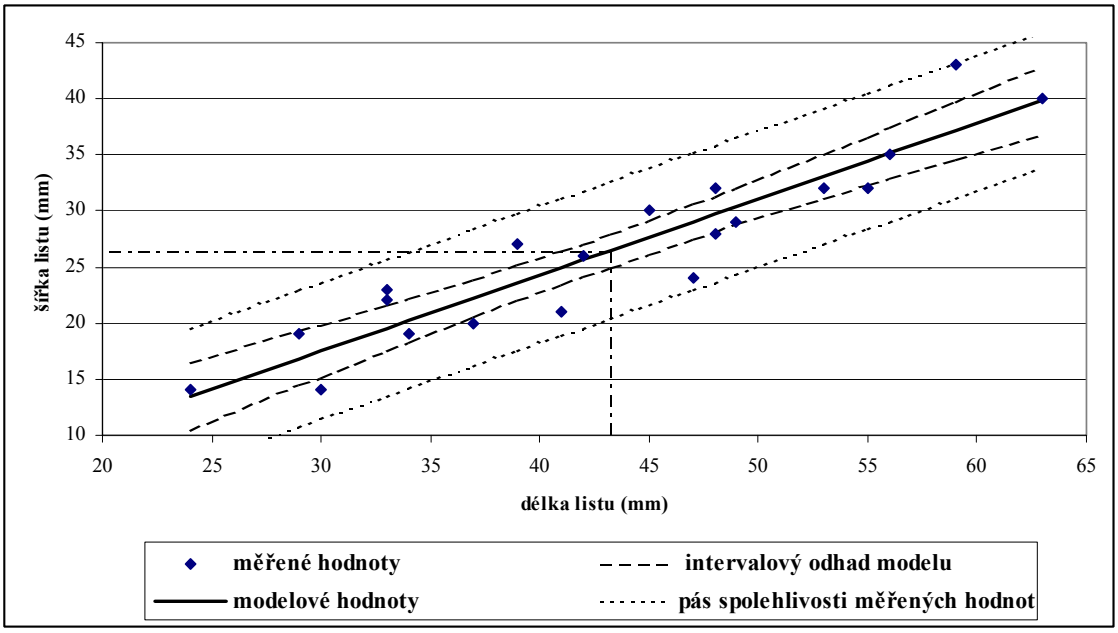
- intervalový odhad modelu platí pro základní soubor (tedy obecně pro všechny bukové listy) – jestliže bychom udělali jakýkoli výběr, tj. změřili libovolný počet jakýchkoliv bukových listů, tak bychom pro délku 24 mm dostali vypočítané hodnoty šířky listů v rozmezí 10.5 – 16.5 mm (s pravděpodobností 95 %);
- stejnou interpretaci má i pás spolehlivosti měřených hodnot - jestliže bychom udělali jakýkoli výběr, tj. změřili libovolný počet jakýchkoliv bukových listů, tak pro list dlouhý 24 mm bychom s pravděpodobností 95 % naměřili šířky v rozmezí 7.4 – 19.6 mm.

Číslo měření	Měřené (empirické) hodnoty (mm)		Modelové (vypočítané) hodnoty (mm)	Intervalový odhad modelových hodnot (mm)		Intervalový odhad (pás spolehlivosti) měřených hodnot (mm)	
	Délka listu	Šířka listu		Dolní hranice	Horní hranice	Dolní hranice	Horní hranice
1	24	14	13.5	10.5	16.5	7.4	19.6
2	29	19	16.9	14.5	19.3	10.8	23.0
3	30	14	17.5	15.2	19.8	11.4	23.6
4	33	22	19.6	17.6	21.6	13.5	25.7
5	33	23	19.6	17.6	21.6	13.5	25.7
6	34	19	20.2	18.3	22.2	14.2	26.3
7	37	20	22.3	20.6	23.9	16.2	28.4
8	39	27	23.6	22.1	25.2	17.5	29.7
9	41	21	25.0	23.5	26.4	18.9	31.1
10	42	26	25.7	24.2	27.1	19.6	31.7
11	45	30	27.7	26.2	29.1	21.6	33.8
12	47	24	29.0	27.5	30.6	22.9	35.1
13	48	28	29.7	28.1	31.3	23.6	35.8
14	48	32	29.7	28.1	31.3	23.6	35.8
15	49	29	30.4	28.8	32.0	24.3	36.5
16	53	32	33.1	31.1	35.0	27.0	39.2
17	55	32	34.4	32.3	36.6	28.4	40.5
18	56	35	35.1	32.9	37.4	29.0	41.2
19	59	43	37.2	34.6	39.7	31.1	43.2
20	63	40	39.9	36.8	42.9	33.8	45.9

Tabulka 10.4 – Intervalové odhady modelu a pás spolehlivosti měřených hodnot

Z grafu na obrázku **10.14** je vidět, že intervalový odhad modelu (čárkované čáry) není v celém průběhu měřených hodnot stejný, ale je nejužší pro bod, který je dán aritmetickým průměrem vysvětlující i vysvětlované proměnné (toto místo je označeno čerchovanou čárou, která na osách udává hodnoty obou průměrů). Je to způsobeno tím, že zde je čitatel zlomku pod odmocninou ve vzorci 10.59 nejmenší (protože je nejmenší rozdíl mezi měřenou hodnotou x a průměrem X). Směrem k „okrajům“, tj. dále od průměru, se intervalový odhad rozšiřuje, tedy pro stejnou spolehlivost (95 %) je širší.

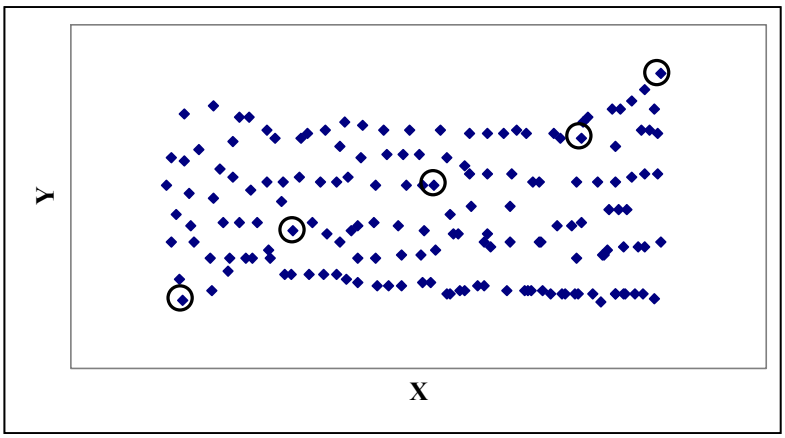
Naopak pás spolehlivosti empirických hodnot je konstantní a závisí pouze na velikosti výběru (prostřednictvím hodnoty t) a na variabilitě vysvětlované proměnné.



Obrázek 10.14 – Intervalový odhad modelu a měřených hodnot pro data příkladu 10.1

10.7 Testování statistických hypotéz v korelační a regresní analýze

Vzhledem k tomu, že obvykle pracujeme s výběry, je nutné veškeré vlastnosti základních souborů, ze kterých pochází studované výběry, testovat. Nejobvyklejší testy jsou testy významnosti modelu a regresních koeficientů. Tyto testy je nutné provést vždy, pokud stanovíme konkrétní tvar regresního modelu (tj. vypočítáme koeficienty všech regresních parametrů a korelační koeficient). Teoreticky se totiž může stát, že sledované veličiny jsou v základním souboru nezávislé, ale do výběru se náhodou dostanou hodnoty, které určitou závislost vykazují. Schématické znázornění takové situace je na obrázku 10.15 . Je proto nutné se ptát, jaká je pravděpodobnost, závislost dané díly najdeme jako důsledek náhody při výběru.



Obrázek 10.15 – Hypotetický základní soubor dat s korelačním koeficientem $\rho = 0$ i regresním koeficientem $\beta = 0$. Zakroužkované body jsou možným výběrem pěti pozorování, které vykazují statisticky významnou závislost.

10.7.1 Test významnosti korelačního koeficientu

Testujeme hypotézu

H_0: $\rho = 0$, *korelační koeficient základního souboru (ρ) je nulový, tj. mezi zkoumanými proměnnými není statisticky významná* ***lineární*** *korelace.*

Pro obecný případ mnohonásobného korelačního koeficientu použijeme testové kritérium

$$F_R = \frac{R^2(n-m)}{(1-R^2)(m-1)}, \tag{10.61}$$

kde je

R je vypočítaná hodnota mnohonásobného korelačního koeficientu

m počet parametrů modelu,

které má F-rozdělení s *(n-m)* a *(m-1)* stupni volnosti. Jestliže platí, že

$$F_R < F_{\alpha,n-m,m-1},$$

potom nezamítáme H_0.

Tento vzorec pro párový korelační koeficient přechází na tvar

$$t_R = \frac{R \cdot \sqrt{n-2}}{\sqrt{1-R^2}} \tag{10.62}$$

který má Studentovo rozdělení s *(n-2)* stupni volnosti. Platí-li $|t| > t_{\alpha,n-2}$, potom H_0 zamítáme.

Tento test je velmi citlivý na dodržení dvourozměrné normality. Pro urychlení konvergence náhodné veličiny R k normalitě můžeme použít Rubenovu transformaci, kdy veličinu t_R nahradíme veličinou (MELOUN - MILITKÝ 1994)

$$R(R) = \frac{\sqrt{n-2.5} \cdot R}{\sqrt{1-0.5R^2}} \tag{10.63}$$

která má i pro malé výběry normované normální rozdělení.

Testování podle vzorce 10.62 lze použít i pro parciální korelační koeficient s tím, že počet stupňů volnosti kritické hodnoty se upraví podle vztahu f = n - k - 2, kde *k* je počet proměnných, které považujeme za konstantní (stejně jako u intervalových odhadů). Podobně upravíme i výraz v čitateli vzorce pod odmocninou, tj. např. pro parciální korelační koeficient I. řádu bude zde n - 1- 2, tj. n – 3 apod.

10.7.2 Test významnosti regresního modelu jako celku

Je to test, který simultánně testuje významnost koeficientu determinace a všech regresních koeficientů vyjma absolutního členu.

H_0: $R^2 = 0$, $\boldsymbol{b} = 0$, *tj. regresní model je nevýznamný.*

Testové kritérium i kritická hodnota je shodné se vzorcem 10.61. Znamená to, že pokud je zamítnuta nulová hypotéza, tak regresní model **jako celek** (tj. lineární kombinace všech nezávislých proměnných) statisticky významně přispívá k odhadu závisle proměnné.

Kromě tohoto testu se často používá jako test významnosti modelu i **analýza rozptylu**. Používáme jednofaktorovou analýzu rozptylu (kde !faktorem“ je regresní model) v úpravě uvedené v tabulce 10.5 .

Využití analýzy rozptylu jako testu významnosti vychází se schématu uvedeného na obrázku 10.5 . Celková variabilita závisle proměnné se rozloží na část vysvětlenou modelem (analogie variability vysvětlené rozdílem mezi skupinami v běžné jednofaktorové analýze rozptylu) a na část nevysvětlenou modelem (analogie variability vysvětlené rozdíly hodnot uvnitř skupin).

Testové kritérium F se porovná s kritickou hodnotou $F_{\alpha;m-1;n-m}$. Pokud je $F > F_{\alpha;m-1;n-m}$, potom zamítáme nulovou hypotézu a přijímáme závěr, že regresní model je významný. Hodnoty F a F_R podle vzorce 10.61 vychází číselně stejně.

Zdroj variability	Součet čtverců odchylek	Počet stupňů volnosti	Průměrný čtverec odchylek (rozptyl)	Testové kritérium
regresní model	$S_{REG} = \sum_{i=1}^{n} (y'_i - \bar{y})^2$	$DF_{REG} = m - 1$	$M_{REG} = \frac{S_{REG}}{DF_{REG}}$	$F = \frac{M_{REG}}{M_R}$
reziduum (nevysvětleno regresním modelem)	$S_R = \sum_{i=1}^{n} (y_i - y'_i)^2$	$DF_R = n - m$	$M_R = \frac{S_R}{DF_R}$	
Celkový	$S_C = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$DF_C = n - 1$		

Tabulka 10.5 – Využití analýzy rozptylu jako testu významnosti regresního modelu

10.7.3 Test významnosti jednotlivých regresních koeficientů

Test uvedený v kapitole 10.7.2 testuje regresní model jako celek. Ovšem zvláště v modelech s větším počtem nezávisle proměnných je nutné testovat i významnost jednotlivých regresních koeficientů. Pokud se ukáže, že některý z nich není významný, je zpravidla možné ho z modelu vypustit bez ztráty významnosti modelu jako celku.

Můžeme také testovat hypotézu, že b_j se rovná určité hodnotě (nikoli pouze nule), protože obecně platí, že

$$t = \frac{b - \beta}{s_b} \qquad (10.64)$$

kde je

b	vypočítané hodnota parametru (odhad parametru)
β	hypotetická (testovaná) hodnota parametru
s_b	směrodatná odchylka parametru

Test pro *j-tý* regresní koeficient se provede následujícím způsobem (s využitím vztahů z kapitoly 10.6.2):

H_0: *$b_j = 0$, tj. j-tý regresní koeficient je nevýznamný.*

Použijeme testové kritérium

$$T_j = \frac{|b_j - \beta_j|}{\sigma\sqrt{c_{jj}}} \qquad (10.65)$$

kde je

b_j odhad (vypočítaná hodnota) *j*-tého regresního koeficientu

β_j stanovená hodnota *j*-tého regresního koeficientu (obvykle $\beta_j = 0$)

σ odhad reziduální směrodatné odchylky

c_{jj} *j-tý* diagonální prvek matice $\mathbf{(X^T X)^{-1}}$,

které má Studentovo t-rozdělení s *(n-m)* stupni volnosti.

Jmenovatel testového kritéria pro jednoduchou korelaci se může také stanovit podle vzorce 10.57.

Pokud platí, že $|T_j| > t_{\alpha/2,n-m}$, potom H_0 zamítáme a regresní koeficient považujeme za významný.

Příklad 10.7:

Testujte významnost regresního modelu i jednotlivých parametrů podle příkladu 10.3.

Pro data z příkladu 10.3 byl vypočítán mnohonásobný korelační koeficient R = 0.9844, z čehož plyne koeficient determinace $R^2 = 0.969$. Použijeme testové kritérium 10.61

$$F_R = \frac{0.969(10-3)}{(1-0.969)(3-1)} = 109.86$$

Kritická hodnota $F_{0.05,2,7} = 4.74$ je menší než F_R. Znamená to, že navržený regresní model je statisticky významný (ale to neznamená, že je navržen zcela optimálně, že je to nejlepší ze všech možných modelů). Zároveň to znamená, že i korelační koeficient je statisticky významný.

Test pomocí analýzy rozptylu je uveden v tabulce 10.6.

Zdroj variability	Součet čtverců odchylek	Počet stupňů volnosti	Průměrný čtverec odchylek (rozptyl)	Testové kritérium
Model	5.330	2	2.665	109.864
Rezidua	0.170	7	0.024	
Celkem	5.500	9		

Tabulka 10.6 – Výsledky analýzy rozptylu pro data příkladu 10.3

Výsledky obou testů potvrzují, že **model jako celek je významný**.

Testování jednotlivých regresních koeficientů se provede podle vzorce 10.65 s využitím výsledků příkladu 10.3, např. pro b_1:

$$T_1 = \frac{|0.069 - 0|}{0.15547 \cdot \sqrt{0.036}} = 2.34$$

Obdobně

$$T_2 = 2.80$$
$$T_0 = -7.43$$

Kritická hodnota $t_{0.025,7} = 2.365$, což znamená, že **koeficienty b_0 a b_2 jsou významné** (absolutní hodnoty jejich testových kritérií jsou vyšší než je kritická hodnota), **koeficient b_1** je (i když „těsně“) **nevýznamný**.

Pokud provedeme nový výpočet pro upravený regresní model $y = b_0 + b_2x_2$, získáme jiné regresní koeficienty $b_0 = 3.466$ a $b_2 = 0.0696$. Korelační koeficient poklesne jen nepatrně na 0.972. Pokud provedeme opakované testování významnosti, zjistíme, že model jako celek i jeho jednotlivé koeficienty jsou významné.

V této souvislosti je vhodné uvést možné kombinace výsledků F-testu (test významnosti celého modelu) a t-testů (testy významnosti pro jednotlivé regresní koeficienty) pro regresní modely a jejich hodnocení – přehled je v tabulce 10.7 .

Přesto, že ve většině případů tyto klasické testy významnosti plně postačují, je nutno upozornit, že se na ně nelze vždy „slepě“ spoléhat. Budeme to ilustrovat na následujícím příkladu.

Výsledek F testu	Výsledek t -testu	Hodnocení modelu
nevýznamný	všechny nevýznamné	posuzované veličiny jsou lineárně nezávislé nebo model je nevhodný (nevystihuje variabilitu závisle proměnné)
významný	všechny významné	vhodný (ale nemusí být optimálně navržen)
významný	některé nevýznamné	vhodný (je možné vypustit nevýznamné členy modelu)
významný	všechny nevýznamné	zvláštní případ způsobený multikolinearitou – je nutné upravit nebo zcela změnit model

Tabulka 10.7 - Hodnocení významnosti regresních modelů na základě F-testu a t-testu

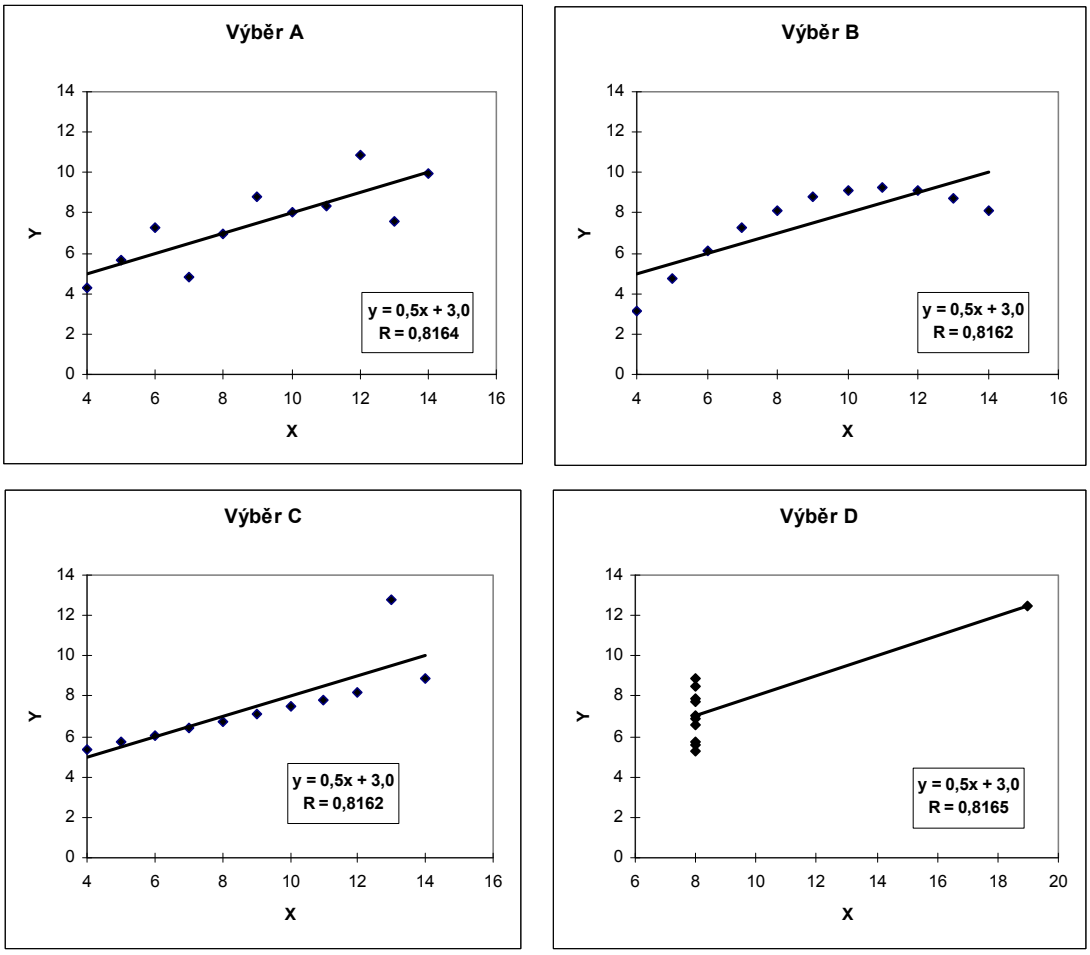
Příklad 10.8:

Pro následující simulovaná (tabulka 10.8) data podle Anscomba (podle MELOUN-MILITKÝ 1994) stanovte parametry regresního modelu $y = b_1 + b_2x$. Testujte významnost regresního modelu i jednotlivých parametrů.

Číslo bodu	X	Výběr A	Výběr B	Výběr C	Výběr D	
		Y	Y	Y	X	Y
1	4	4.26	3.10	5.39	8	6.58
2	5	5.68	4.74	5.73	8	5.76
3	6	7.24	6.13	6.08	8	7.71
4	7	4.82	7.26	6.42	8	8.84
5	8	6.95	8.14	6.77	8	8.47
6	9	8.81	8.77	7.11	8	7.04
7	10	8.04	9.14	7.46	8	5.25
8	11	8.33	9.26	7.81	8	5.56
9	12	10.84	9.13	8.15	8	7.91
10	13	7.58	8.10	12.74	8	6.89
11	14	9.96	8.10	8.84	19	12.50

Tabulka 10.8 – Simulovaná data podle Anscomba

Výsledky regresní analýzy těchto čtyř výběrů jsou překvapivé. Ačkoli se podle grafického zobrazení na obrázku všechny výběry zřetelně liší, platí pro ně stejné číselné výsledky: $b_1 = 3.0$, $b_2 = 0.5$, korelační koeficient je prakticky shodný a všechny parametry i model celkově se jeví na základě testů jako významné (hodnota F_R podle vzorce 10.61 se pohybuje v rozmezí 17.96 – 18.00 oproti kritické hodnotě $F_{0,05;1;9} = 5.12$, hodnota t podle vzorce 10.65 pro parametr b_1 je roven 2.67 a pro parametr $b_2 = 4.24$ oproti kritické hodnotě $t_{0,025;9} = 2.26$). To znamená, že podle hodnocení testů významnosti z tabulky by měl být model pro všechny výběry vhodný. Přitom z obrázku 10.16 je zřejmé, že přímka je vhodná pouze pro výběr A. Výběr B je možné lépe vystihnout křivočarou závislostí, výběr C, ačkoli je přímce nejblíže, je ovlivněn jedním vybočujícím bodem a výběr D je zcela zvláštní případ dat, který ilustruje, jakou „moc“ může mít v regresním modelu jediný bod.



Obrázek 10.16 – Grafické znázornění Anscombových dat a jejich modelu

Je nutné zdůraznit, že v tomto případě byla data úmyslně volena tak, aby neshoda modelu s daty byla „do očí bijící“. V mnoha případech jsou rozdíly daleko jemnější a volba optimálního modelu obtížnější. Některé vhodné metody posouzení vhodnosti modelu a jeho diagnostiky budou uvedeny v dalším textu.

10.7.4 Testy shody jednoho, dvou a více korelačních koeficientů

10.7.4.1 Test shody korelačního koeficientu se zadanou hodnotou (normou)

Testujeme nulovou hypotézu

H_0: *$\rho = \rho_0$, tj. korelační koeficient základního souboru ρ se rovná dané hodnotě (normě), tedy rozdíl mezi výběrovým korelačním koeficientem a hodnotou ρ_0 je jen náhodný.*

Vzhledem k tomu, že náhodná veličina *r* (korelační koeficient) nemá normální rozdělení, musíme i zde, podobně jako u intervalových odhadů, používat Fisherovu transformaci podle vztahu 10.48. Použijeme testové kritérium

$$Z_1 = \left| Z_r - Z_{\rho_0} \right| \cdot \sqrt{n-3} \qquad (10.66)$$

kde je

Z_r Fisherova transformace výběrového korelačního koeficientu

$Z_{\rho 0}$ Fisherova transformace normované (zadané) hodnoty korelačního koeficientu základního souboru

Testové kritérium Z_1 porovnáváme s kvantilem normovaného normálního rozdělení $z_{\alpha/2}$. Je-li testové kritérium vyšší než $z_{\alpha/2}$, zamítáme H_0.

10.7.4.2 Test shody dvou korelačních koeficientů

Testujeme nulovou hypotézu

H_0: *$\rho_1 = \rho_2$, výběrové korelační koeficienty r_1 a r_2 pocházejí ze základních souborů, jejichž korelační koeficienty jsou shodné, tedy rozdíl $r_1 - r_2$ je pouze náhodný.*

Je-li rozsah obou porovnávaných výběrů n_1 a n_2 různý, použijeme testové kritérium

$$Z_2 = \frac{\left| Z_{r_1} - Z_{r_2} \right|}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \qquad (10.67)$$

jsou-li rozsahy výběrů stejné, potom má testové kritérium tvar

$$Z_2 = \frac{\left| Z_{r_1} - Z_{r_2} \right|}{\sqrt{\frac{2}{n-3}}} \qquad (10.68)$$

Také zde se jako kritická hodnota používá kvantil normovaného normálního rozdělení $z_{\alpha/2}$. Je-li testové kritérium vyšší než $z_{\alpha/2}$, zamítáme H_0. V tom případě oba výběry pocházejí ze základních souborů, jejichž korelační koeficienty se liší.

V případě, že nulovou hypotézu nezamítneme, předpokládáme, že oba výběry pochází ze základních souborů se shodnými korelačními koeficienty a tehdy můžeme vypočítat společný korelační koeficient pro základní soubor vzniklý spojením obou porovnávaných souborů

$$z_w = \frac{(n_1 - 3)z_{r_1} + (n_2 - 3)z_{r_2}}{(n_1 - 3) + (n_2 - 3)} \qquad (10.69)$$

pokud mají oba původní výběry stejný počet prvků, potom se použije zjednodušený vzorec

$$z_w = \frac{z_{r_1} + z_{r_2}}{2} \qquad (10.70)$$

Hodnoty z_w vycházejí transformované, konečná hodnota r_w se získá odtransformováním podle příslušných tabulek nebo pomocí funkce FISHERINV(z_w) v Excelu.

Příklad 10.9:

V příkladu 10.1 byla pro závislost délky a šířky bukových listů stanovena hodnota $r_1 = 0.933$ pro výběr $n_1 = 20$. Poté byl proveden na jiné lokalitě výběr $n_2 = 26$ a byl stanoven $r_2 = 0.869$. Posuďte, zda těsnost závislosti na obou lokalitách je možné považovat za stejnou.

Použijeme testové kritérium 10.67, protože závislost bude možné považovat za stejnou, pokud se budou rovnat korelační koeficienty

$$Z_2 = \frac{|1.681 - 1{,}329|}{\sqrt{\frac{1}{20-3} + \frac{1}{26-3}}} = 1.101$$

Testové kritérium je menší než kritická hodnota $z_{\alpha/2} = 1.96$, tedy můžeme učinit závěr, že těsnost vztahu mezi délkou a šířkou bukových listů na obou lokalitách je stejná. Můžeme tedy vypočítat společný korelační koeficient podle vztahu 10.69

$$z_w = \frac{(20-3)1.681 + (26-3)1.329}{(20-3) + (26-3)} = 1.479 \Rightarrow \text{FISHERINV}(1.479) = r_w = 0.901$$

Společný korelační koeficient pro obě lokality je 0.901.

10.7.4.3 Test shody více korelačních koeficientů

Testujeme nulovou hypotézu

H_0: *$\rho_1 = \rho_2 = \ldots = \rho_k$, všechny porovnávané korelační koeficienty $r_1, r_2, \ldots, r_k$ pochází ze základních souborů se shodnými korelačními koeficienty, tedy rozdíly mezi nimi ($r_1 - r_2, r_1 - r_3, \ldots, r_{k-1} - r_k$) jsou pouze náhodné.*

Použijeme testové kritérium

$$\chi^2 = \sum_{i=1}^{k} (n_i - 3)Z_i^2 - \frac{\left[\sum_{i=1}^{k} (n_i - 3)Z_i\right]^2}{\sum_{i=1}^{k} (n_i - 3)} \qquad (10.71)$$

které porovnáme s kritickou hodnotou $\chi^2_{\alpha;k-1}$.

Pokud nezamítneme nulovou hypotézu, předpokládáme, že všechny výběry pochází ze základních souborů se společným korelačním koeficientem. Stejně jako v případě dvou korelačních koeficientů, i zde můžeme vypočítat společný korelační koeficient

$$z_w \frac{\sum_{i=1}^{k}(n_i - 3)Z_i}{\sum_{i=1}^{k}(n_i - 3)} \tag{10.72}$$

kde transformovanou hodnotu z_w převedeme na r_w pomocí tabulek nebo FISHERINV(z_w).

Pokud je nulová hypotéza zamítnuta, znamená to, že alespoň mezi dvěma korelačními koeficienty je statisticky významný rozdíl. Můžeme provést **mnohonásobná porovnání**, abychom zjistili, mezi kterými korelačními koeficienty tento rozdíl je. Použijeme k tomu modifikaci Tukeyho metody (viz kapitola 9 o analýze rozptylu). Testujeme nulovou hypotézu

H_0: *$\rho_A = \rho_B$, tj. výběry A a B pocházejí ze základních souborů, jejichž korelační koeficienty se rovnají.*

$$q = \frac{Z_A - Z_B}{SE} \tag{10.73}$$

kde je pro shodné velikosti obou porovnávaných výběrů (n)

$$SE = \sqrt{\frac{1}{n-3}} \tag{10.74}$$

a pro různé velikosti obou porovnávaných výběrů (n_1 a n_2)

$$SE = \sqrt{\frac{1}{2}\left(\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}\right)} \tag{10.75}$$

Testové kritérium q se porovnává s kritickou hodnotou studentizovaného rozpětí $q_{\alpha;\infty;k}$.

Příklad 10.10:

Testujte shodu tří korelačních koeficientů $r_1 = 0.52$ ($n_1 = 24$), $r_2 = 0.56$ ($n_2 = 29$) a $r_3 = 0.87$ ($n_3 = 32$).

Použijeme kritérium 10.71, kde transformované hodnoty podle Fisherovy transformace budou $Z_1 = 0.576$, $Z_2 = 0.633$ a $Z_3 = 1.333$. Poté vyjde hodnota $\chi^2 = 9.478$. Kritická hodnota $\chi^2_{0.05;2} = 5.991$, což znamená, že $\chi^2 > \chi^2_{0.05;2}$, tedy zamítáme nulovou hypotézu. Mezi porovnávanými korelačními koeficienty je alespoň jeden statisticky významný rozdíl. Který to je, zjistíme pomocí metody mnohonásobného porovnání prostřednictvím vzorců 10.73 a 10.75. Výsledky jsou v tabulce 10.9 .

Srovnání korelačních koeficientů výběrů B a A	Z_{B-A}	SE	Testové kritérium q	Kritická hodnota $q_{0.05}$	Výsledek testu
3 - 1	0.757	0.203	3.728	3.314	**Zamítáme H_0**
3 - 2	0.700	0.191	3.667	3.314	**Zamítáme H_0**
2 - 1	0.057	0.207	0.273	3.314	Nezamítáme H_0

Tabulka 10.9 – Výsledky mnohonásobného porovnání korelačních koeficientů

Výsledky je možné interpretovat tak, že korelační koeficient $r_3 = 0.87$ se statisticky významně liší od ostatních dvou koeficientů, které tvoří homogenní skupinu.

10.7.5 Testy shody regresních modelů

Častou statistickou úlohou je vyšetřit shodu regresních modelů. Nejčastější úlohy jsou tyto:

- porovnává se **jeden empirický regresní model** s „normovaným“ **(teoretickým) modelem**, tj. s danou závislostí (např. převzatou z literatury) a ověřuje se, zda empirický model teoretické závislosti vyhovuje;
- porovnávají se **dva nebo více empirických modelů mezi sebou** a ověřuje se, zda je možné přijmout tvrzení, že všechny porovnávané výběry pochází z jednoho základního souboru, kde platí jeden regresní model.

Obrázek 10.17 ukazuje možné varianty, v jakých parametrech se mohou lineární regresní modely lišit. Obrázek A ukazuje shodné modely – obě přímky se neliší ani v v absolutním členu (tj. úseku na ose Y, viz obrázek 10.9) ani ve směrnici, tj. sklonem přímky. Ostatní obrázky ukazují neshodné modely – lišící se buď úsekem nebo směrnicí nebo obojím.

Z obrázků vyplývá, že pro posouzení shody modelů je nutné testovat jak absolutní člen (a), tak i regresní koeficient (b). Jednotlivé testy si ukážeme na nejjednodušším modelu – na přímce.

10.7.5.1 Test shody empirického a teoretického modelu přímky

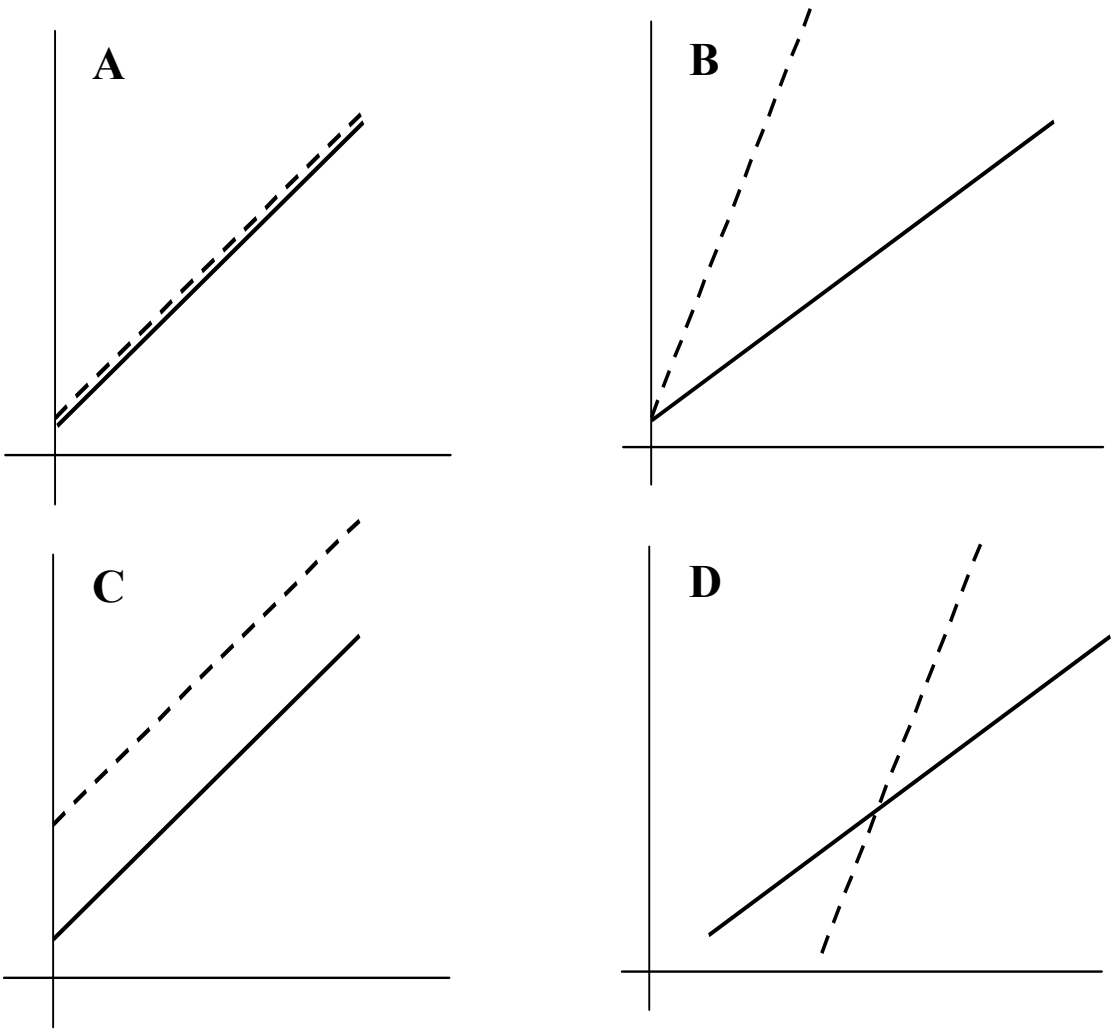
Testujeme nulovou hypotézu

H_0: *Empirický model $y' = a + bx$ pochází ze základního souboru, jehož model $y' = \alpha + \beta x$ je shodný s teoretickým modelem $y'_0 = \alpha_0 + \beta_0 x$, tj. platí $\alpha = \alpha_0$, $\beta = \beta_0$.*

Nejdříve budeme testovat regresní člen β, resp. jeho odhad b. Může se použít testovací kritérium známé již z kapitoly 10.7.3 o testování regresních parametrů

$$t = \frac{b - \beta_0}{s_b} \qquad (10.76)$$

kde s_b je směrodatná odchylka regresního (obecně testovaného) členu regresního modelu (stanovíme podle postupu v kapitole 10.7.3, které má kritickou hodnotu $t_{\alpha/2,n-2}$.



Obrázek 10.17 – Možné vztahy dvou regresních modelů – (A) shodné modely – shodují se v úseku i ve směrnici, (B) neshodné modely – shodují se v úseku, ale liší ve směrnici, (C) neshodné modely – liší se v úseku (systematické posunutí), shodují ve směrnici, (D) neshodné modely – liší se úsekem i směrnicí

Pokud je nulová hypotéza zamítnuta, již víme, že se nejedná o shodné modely, empirický a teoretický model se neshoduje minimálně ve směrnici (tedy případ B anebo D z obrázku 10.17). V tomto okamžiku obvykle již testování může skončit, pouze pokud je pro nás důležité, zda se modely shodují alespoň v úseku, můžeme testovat shodu absolutního členu. Pokud nulová hypotéza o regresním členu není zamítnuta (směrnice považujeme za shodné), musíme pokračovat testem absolutního členu, abychom zjistili, zda modely nejsou systematicky posunuty.
Testování absolutního členu provedeme stejně, tj. podle vzorce 10.76.
Nulová hypotéza v tomto případě říká, že oba absolutní členy, empirický i teoretický, si jsou rovny (v tom případě by přímky ležely na sobě, byly by shodné).

Příklad 10.11:

Stanovte parametry modelu přímky pro vztah mezi hustotou dřeva ρ (kg/m^3) a koeficientem objemového bobtnání αV (%). Posuďte, zda se tento empirický model shoduje s teoretickým modelem $\alpha V = 0.028 \cdot \rho$. Měřené hodnoty jsou v tabulce 10.10 .

Číslo měření	Hustota	Koeficient bobtnání	Číslo měření	Hustota	Koeficient bobtnání	Číslo měření	Hustota	Koeficient bobtnání
1	469.03	15.18	16	502.00	17.80	31	413.90	14.60
2	587.50	16.29	17	619.00	17.90	32	616.80	16.30
3	718.60	17.71	18	745.00	17.75	33	736.20	18.40
4	475.10	20.40	19	369.00	13.00	34	452.00	19.60
5	614.40	19.80	20	734.00	19.00	35	560.00	19.20
6	753.00	21.90	21	641.00	16.00	36	792.00	21.30
7	497.00	24.90	22	446.00	13.70	37	490.00	18.47
8	626.00	20.67	23	645.00	19.50	38	627.00	17.22
9	847.00	23.76	24	738.00	17.80	39	710.00	15.35
10	419.00	17.80	25	503.00	22.30	40	440.00	18.10
11	649.00	19.60	26	612.00	18.30	41	646.00	17.50
12	687.00	16.30	27	709.00	22.10	42	738.00	16.30
13	338.00	12.30	28	384.00	13.97	43	471.30	14.70
14	654.00	18.10	29	661.00	16.87	44	587.50	16.70
15	825.00	23.00	30	794.00	21.04	45	718.40	17.80

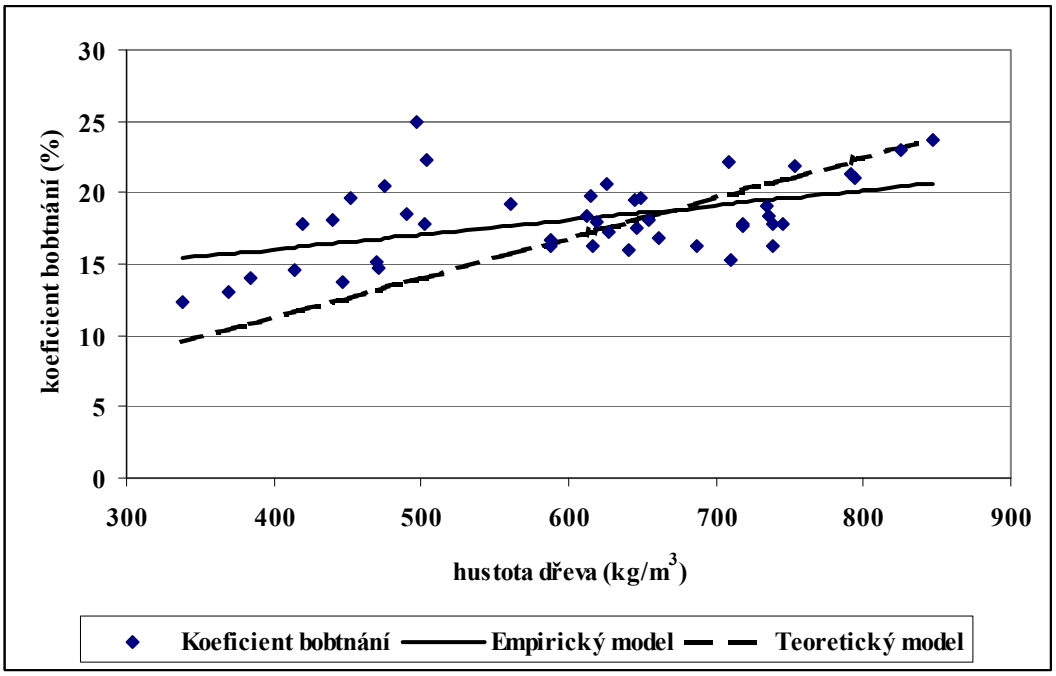
Tabulka 10.10 – Hodnoty hustoty dřeva a koeficientu bobtnání

Nejprve běžnými metodami stanovíme empirický model, pro zadaná data bude mít tvar $\alpha V' = 11.842 + 0.0104 \cdot \rho$. Teoretický model je dán tvarem $\alpha V = 0.028 \cdot \rho$. Musíme tedy testovat nulovou hypotézu , že hodnoty empirického (0.0104) a teoretického regresního koeficientu (0.028) se rovnají. Vypočítáme hodnotu s_b podle postupu v kapitole 10.7.3 a použijeme testové kritérium 10.76

$$t = \frac{|0.0104 - 0.028|}{0.002821} = 6.24$$

$$t_{0.025;43} = 2.02$$

Porovnáním testového kritéria a kritické hodnoty zjistíme, že 6.24 > 2.02, tedy nulovou hypotézu o shodě empirického a teoretického modelu zamítáme. Naměřené hodnoty neodpovídají teoretickému modelu. Dále již testovat nemusíme, protože jsme zodpověděli hlavní otázku – modely nejsou shodné. Závěr testu potvrzuje i grafické znázornění na obrázku 10.18 .



Obrázek 10.18 – Porovnání teoretického a empirického modelu

10.7.6 Test shody dvou lineárních modelů

Pro testování shody dvou empirických lineárních modelů se používá **Chowův test**. Vycházíme z testování shody regresních parametrů dvou lineárních modelů

$$\mathbf{y}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1$$
$$\mathbf{y}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2$$

kde je

$\mathbf{X}_1$ matice $n_1 \times m$ nezávisle proměnných prvního modelu
$\mathbf{X}_2$ matice $n_2 \times m$ nezávisle proměnných druhého modelu
$\mathbf{y}_1$ vektor $n_1 \times 1$ závisle proměnné prvního modelu
$\mathbf{y}_2$ vektor $n_2 \times 1$ závisle proměnné druhého modelu

Při tomto testu využijeme tzv. **složeného modelu**, tj. oba porovnávané výběry sloučíme do jednoho a také pro něj stanovíme parametry stejného modelu jako pro oba dílčí výběry.

Formulujeme nulovou hypotézu:

H_0: $\beta_1 = \beta_2$, *tj. regresní koeficienty obou modelů jsou shodné.*

Použijeme testové kritérium

$$F_C = \frac{(RSC_s - RSC_1 - RSC_2)(n - 2m)}{(RSC_1 + RSC_2)\cdot m} \qquad (10.77)$$

kde je

n celkový počet prvků obou výběrů, tj. $n_1 + n_2$
RSC_s reziduální součet čtverců složeného modelu
RSC_1 reziduální součet čtverců prvního modelu
RSC_2 reziduální součet čtverců druhého modelu

Reziduální součet čtverců se obecně vypočítá

$$RSC = \sum_{i=1}^{n} (y_i - y'_i)^2 \qquad (10.78)$$

Při hodnocení výsledku testu musíme brát v úvahu, zda reziduální rozptyly obou výběrů (podle vzorce 10.55) jsou shodné nebo nejsou, tj. $\sigma_1^2 = \sigma_2^2$ (nutno testovat F-testem pro rozptyly). Pokud jsou, použijeme F-rozdělení s *m* a *n-2m* stupni volnosti. Pokud platí, že $\sigma_1^2 \neq \sigma_2^2$, použijeme počet stupňů volnosti *m* a *r*, kde *r* vypočítáme

$$r = \frac{\left[(n_1 - m)\sigma_1^2 + (n_2 - m)\sigma_2^2\right]^2}{(n_1 - m)\sigma_1^4 + (n_2 - m)\sigma_2^4} \qquad (10.79)$$

Příklad 10.12:

Porovnejte dva modely závislosti mezi hustotou dřeva (kg/m^3) a koeficientem bobtnání (%). Stanovte, zda jsou oba modely shodné. Měřená data jsou v tabulce 10.12 .

Oba porovnávané modely jsou přímkové závislosti. Využijeme vztahu 10.77. K jeho výpočtu musíme znát reziduální sumy čtverců. Potřebné hodnoty udává tabulka 10.11 .

Model	a	b	RSC	n	Reziduální rozptyl
Model I	11.842	0.01040	266.982	45	6.209
Model II	10.235	0.01113	215.884	48	4.693
Složený model	10.999	0.01079	514.614	93	5.655

Tabulka 10.11 – Údaje potřebné k výpočtu testu shody dvou modelů

Dosadíme tyto hodnoty do testového kritéria

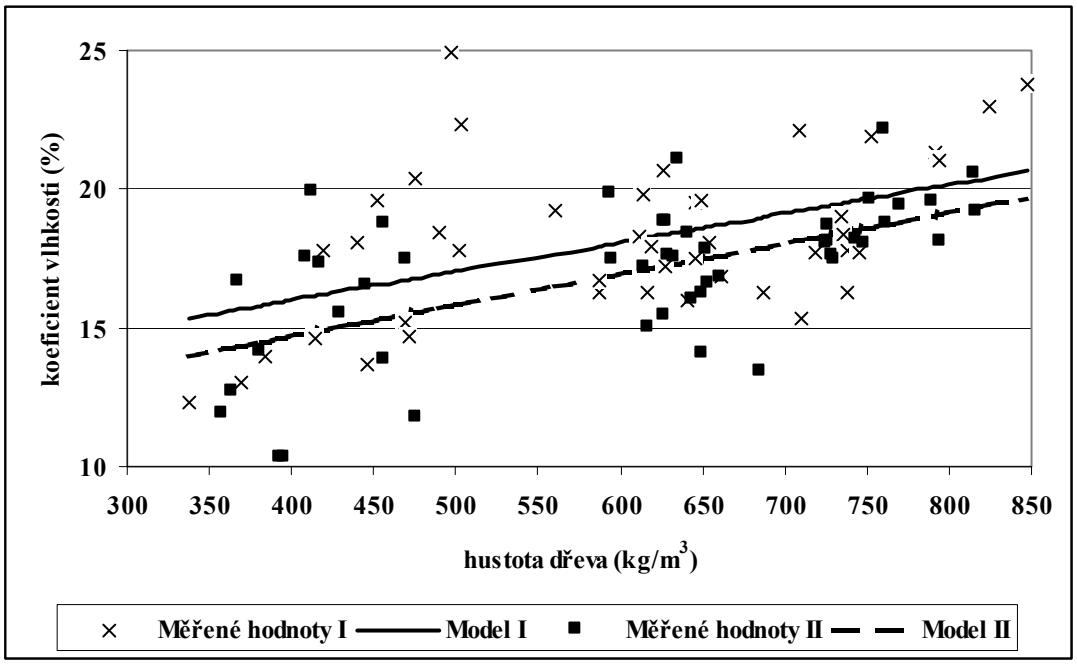
$$F_C = \frac{(514.614 - 266.982 - 215.884)\cdot(93 - 4)}{(266.982 + 215.884)\cdot 2} = 2.926$$

Výslednou hodnotu porovnáme s kritickou hodnotou $F_{0.05;2;89} = 3.099$. Znamená to, že testové kritérium je menší, tedy nezamítáme hypotézu o shodě obou modelů. Výše uvedenou kritickou hodnotu jsme mohli použít proto, že reziduální rozptyly obou modelů jsou stejné (potvrzeno F-testem).

Vzájemné vztahy obou modelů jsou zřetelné z obrázku **10.19** . Směrnice obou přímek jsou shodné, modely se liší systematickým posunutím (tedy absolutním členem). Rozdíl absolutních členů je na hladině významnosti $\alpha = 0.05$ považován ještě za náhodný.

Model I			Model II		
Číslo měření	Hustota dřeva (kg/m^3)	Koeficient bobtnání (%)	Číslo měření	Hustota dřeva (kg/m^3)	Koeficient bobtnání (%)
1	469.03	15.18	1	363.90	12.64
2	587.50	16.29	2	596.29	17.45
3	718.60	17.71	3	685.84	13.41
4	475.10	20.40	4	367.60	16.66
5	614.40	19.80	5	635.52	21.04
6	753.00	21.90	6	730.57	17.44
7	497.00	24.90	7	381.00	14.10
8	626.00	20.67	8	630.00	17.60
9	847.00	23.76	9	727.00	18.10
10	419.00	17.80	10	430.00	15.48
11	649.00	19.60	11	644.00	16.00
12	687.00	16.30	12	761.00	22.15
13	338.00	12.30	13	358.94	11.90
14	654.00	18.10	14	650.90	16.20
15	825.00	23.00	15	790.95	19.50
16	502.00	17.80	16	410.10	17.49
17	619.00	17.90	17	653.90	16.58
18	745.00	17.75	18	729.60	17.56
19	369.00	13.00	19	457.00	18.74
20	734.00	19.00	20	615.00	17.13
21	641.00	16.00	21	727.00	18.67
22	446.00	13.70	22	393.80	10.30
23	645.00	19.50	23	618.10	15.00
24	738.00	17.80	24	815.70	20.50
25	503.00	22.30	25	476.00	11.70
26	612.00	18.30	26	650.20	14.02
27	709.00	22.10	27	762.30	18.70
28	384.00	13.97	28	470.51	17.40
29	661.00	16.87	29	627.38	18.82
30	794.00	21.04	30	744.44	18.12
31	413.90	14.60	31	396.50	10.27
32	616.80	16.30	32	653.00	17.80
33	736.20	18.40	33	753.20	19.60
34	452.00	19.60	34	413.26	19.85
35	560.00	19.20	35	627.70	15.42
36	792.00	21.30	36	817.24	19.19
37	490.00	18.47	37	418.50	17.30
38	627.00	17.22	38	642.41	18.40
39	710.00	15.35	39	794.87	18.10
40	440.00	18.10	40	594.00	19.80
41	646.00	17.50	41	661.00	16.80
42	738.00	16.30	42	749.00	18.00
43	471.30	14.70	43	456.97	13.80
44	587.50	16.70	44	628.50	18.80
45	718.40	17.80	45	726.20	18.02
			46	446.60	16.51
			47	632.80	17.48
			48	770.97	19.34

Tabulka 10.12 – Měřené údaje pro porovnání dvou modelů vztahu mezi hustotou dřeva a koeficientem bobtnání



Obrázek 10.19 – Grafické porovnání dvou modelů

Existují i testy, které porovnávají více modelů (než dva) zároveň. Jejich výpočet je značně komplikovaný. Zájemci najdou podrobnosti např. v MELOUN-MILITKÝ 1994 a ZAR 1984.

10.7.7 Test vhodnosti lineárního modelu

V některých případech je nutné posoudit, zdali pro vystižení experimentálních dat je možné využít lineární model nebo je vhodnější použít nelineární. K tomuto účelu se požívá např. **test Uttsové**.

H_0: *navržený lineární regresní model je správný.*

Využívá se zde reziduální součet čtverců RSC pro zkoumaný model a RSC_1 pro model s využitím n_1 prvků ve statistice

$$F_U = \frac{(RSC - RSC_1)(n_1 - m)}{RSC_1(n - n_1)} \qquad (10.80)$$

která má F-rozdělení s n-n_1 a n_1 - m stupni volnosti. Doporučuje se volit $n_1 \approx n/2$ a zařadit mezi vybrané body ty, které mají nejmenší hodnoty diagonálních prvků projekční matice H_{ii} (leží nejblíže těžišti nezávisle proměnných). Pokud platí, že

$$F_U > F_{\alpha,n-n1,n1-m}$$

nelze považovat navržený lineární model za přijatelný.

Tento test je možné doplnit použitím charakteristik **určených k porovnání vhodnosti různých lineárních modelů**. Mezi jejich výhody patří zpravidla snadnější výpočet a jednoduchá interpretace. Mezi často užívané charakteristiky patří

- **střední kvadratická chyba predikce (MEP)**, která se vypočítá podle vztahu

$$MEP = \frac{1}{n}\sum_{i=1}^{n}\frac{e_i^2}{(1-H_{ii})^2} \qquad (10.81)$$

kde je

e_i^2 čtverec reziduí modelu

H_{ii} *i*-tý diagonální prvek projekční matice **H**

Čím je **MEP menší, tím je daný model vhodnější**.

- **Akaikovo informační kritérium (AIC),** které patří mezi nejznámější charakteristiky vhodnosti modelu

$$AIC = n \cdot \ln\left(\frac{RSC}{n}\right) + 2m \qquad (10.82)$$

I zde platí, že čím je **AIC menší, tím je model vhodnější**.

Pokud je to možné, není vhodné spoléhat pouze na jediný test nebo charakteristiku. Mohou nastat případy, kdy určitá statistika „selže", proto je vhodné porovnat více testů a jejich základě rozhodnout.

Na tomto místě je nutné zdůraznit, že výše uvedené statistiky a testy můžeme použít **jen pro porovnání těch modelů, které vyhovují svými vlastnostmi charakteru řešeného problému**.

Příklad 10.13:

Porovnejte vhodnost modelu přímky pro výběr A a výběr B z příkladu 10.8. *Pro výběr B také porovnejte vhodnost modelu paraboly* $y = \beta_1 + \beta_2 x + \beta_3 x^2$.

Řešení podrobně rozebereme na příkladu výběru A, pro ostatní výběry je řešení obdobné.

Pro test správnosti modelu využijeme test Uttsové:

- Vypočítáme parametry modelu přímky, z nichž nás pro výpočet testového kritéria bude hlavně zajímat RSC (reziduální suma čtverců), v tomto případě RSC = 13.763
- Využijeme diagonálních prvků projekční matice **H** (postup jejího výpočtu viz příklad 10.3) pro stanovení velikosti modelu n_1. Je nutné vybrat minimální diagonální prvky v rozsahu zhruba n/2. Pro výběr A vypadá matice **H** takto (diagonální prvky jsou zvýrazněny a pět nejmenších vybraných prvků je v rámečku):

0.100	0.082	0.127	0.091	0.109	0.136	0.064	0.045	0.118	0.073	0.055
0.082	**0.100**	0.055	0.091	0.073	0.045	0.118	0.136	0.064	0.109	0.127
0.127	0.055	**0.236**	0.091	0.164	0.273	-0.018	-0.091	0.200	0.018	-0.055
0.091	0.091	0.091	**0.091**	0.091	0.091	0.091	0.091	0.091	0.091	0.091
0.109	0.073	0.164	0.091	**0.127**	0.182	0.036	0.000	0.145	0.055	0.018
0.136	0.045	0.273	0.091	0.182	**0.318**	-0.045	-0.136	0.227	-0.000	-0.091
0.064	0.118	-0.018	0.091	0.036	-0.045	**0.173**	0.227	0.009	0.145	0.200
0.045	0.136	-0.091	0.091	0.000	-0.136	0.227	**0.318**	-0.045	0.182	0.273
0.118	0.064	0.200	0.091	0.145	0.227	0.009	-0.045	**0.173**	0.036	-0.018
0.073	0.109	0.018	0.091	0.055	-0.000	0.145	0.182	0.036	**0.127**	0.164
0.055	0.127	-0.055	0.091	0.018	-0.091	0.200	0.273	-0.018	0.164	**0.236**

Nejmenší diagonální prvky jsou $H_{1,1}$, $H_{2,2}$, $H_{4,4}$, $H_{5,5}$ a $H_{10,10}$. Znamená to, že pro model o rozsahu $n_1 = 5$ vybereme prvky 1, 2, 4, 5 a 10.

- Vypočítáme znovu regresní model pro tyto vybrané prvky, zde RSC = 3.54.
- Dosadíme do vztahu 10.80 (n = 11, m = 2, $n_1 = 5$):

$$F_U = \frac{13.763 - 3.54(5-2)}{3.54(11-5)} = 1.44$$

Vzhledem k tomu, že 1.44 < 8.94 ($F_{0.05,6,3}$), můžeme považovat model přímky za při- jatelný.

Pro výběr B obdobným způsobem vybereme stejné prvky (regresní parametry obou výběrů jsou stejné) a vypočítáme F_U = 29.98, což je větší hodnota než kvantil 8.94, takže model přímky je pro výběr B nevhodný.

Stejně postupujeme u modelu paraboly pro výběr B. Zde vybereme prvky 1, 2, 5, 7, 9 a 10 (tedy $n_1 = 6$, protože některé prvky H_{ii} byly shodné) a vypočítáme F_U = 5.98, což je menší než 9.01 ($F_{0.05,5,3}$), takže model paraboly je pro výběr B přijatelný.

Pokud pro porovnání použijeme MEP a AIC, dostaneme výsledky z následující tabulky

Výběr	Typ modelu	MEP	AIC
A	přímka	1.871	6.47
A	parabola	1.955	7.76
B	přímka	2.204	6.47
B	parabola	$3.11*10^{-6}$	-138.16

Tyto výsledky potvrzují, že pro výběr A je vhodnějším modelem přímka, pro výběr B parabola.

10.7.8 Test závažnosti multikolinearity

Problematika multikolinearity byla podrobněji rozebrána již v kapitole 10.5.2.2, která se týkala předpokladů MNČ. Uvedli jsme, že multikolinearita (která je v různé míře přítomna ve většině modelů mnohonásobné regrese) nemusí mít „škodlivé“ účinky vždy, ale až od určité míry jejího výskytu. Z tohoto důvodu byl vyvinut test, který „měří“ sílu multikolinearity - **Scottův test**, který je založen na testovém kritériu (MELOUN - MILITKÝ 1994)

$$M_T = \frac{\frac{F_R}{T_S} - 1}{\frac{F_R}{T_S} + 1} \qquad (10.83)$$

kde je

F_R testové kritérium významnosti regresního modelu (vztah 10.61)

T_S se stanoví podle vzorce

$$T_S = \frac{\sum_{j=1}^{m} T_j^2}{m-1} \qquad (10.84)$$

kde T_j je testové kritérium podle vzorce 10.65

Pokud je M_T

vyšší než 0.80	model je z hlediska multikolinearity **nevyhovující** a **je nutné provést jeho úpravu;**
0.33 - 0.80	model je z hlediska multikolinearity **nevhodný** a **je doporučeno provést jeho úpravu;**
do 0.33	model je z hlediska multikolinearity **vyhovující**, **úpravy nejsou potřebné.**

Jestliže test potvrdí silnou multikolinearitu, je možné vypustit některé proměnné (což není vždy vhodné nebo možné řešení) nebo je možné regresní model vypočítat metodou racionálních hodností místi MNČ (podrobněji viz MELOUN - MILITKÝ 1994).

Jiným kritériem popisujícím sílu multikolinearity, je **Variance Inflation Factor (VIF)**. Stanoví se jako diagonální prvky matice $\mathbf{(R)^{-1}}$, kde **R** je korelační matice vysvětlujících (nezávislých) proměnných. Postup je jednoduchý – vypočítáme korelační matici vysvětlujících proměnných, provedeme její inverzi (obojí lze provést např. v Excelu) a diagonální prvky této matice jsou přímo VIF hodnoty. Pokud jsou hodnoty VIF vyšší než 10, jedná se o nepřípustně silnou multikolinearitu.

Příklad 10.14:

Posuďte multikolinearitu pro data příkladu 10.2.

Příklad 10.2 byl zaměřen na výpočet parciálních korelačních koeficientů. Jeho výsledky indikovaly možnost výskytu multikolinearity (tedy vzájemné závislosti nezávislých proměnných). Máme posoudit, nakolik je multikolinearita v tomto případě závažná. Musíme stanovit regresní model (jeho konstrukce bude podrobněji rozebrána později v kapitole o regresní diagnostice). Model má celkem čtyři parametry (absolutní člen a tři regresní parametry pro výčetní tloušťku, výšku a délku koruny). Hodnoty testových kritérií T_j jsou následující: T_1 (pro absolutní člen) = -9.683, T_2 = 10.431, T_3 = 1.392 a T_4 = 1.497. Hodnota testového kritéria pro test významnosti modelu je F_R = 481.69. Z těchto údajů vypočítáme T_s podle vztahu 10.84 s výsledkem 68.918 a dosadíme do testového kritéria 10.83 a získáme hodnotu M_T = 0.749. Podle tabulky uvedené jako vyhodnocení se jedná o model nevhodný (blíží se hranici pro model nepřijatelný) a je doporučena jeho úprava.

Pokud bychom chtěli stanovit také VIF hodnoty, musíme nejdříve vypočítat korelační matici nezávislých proměnných a poté ji invertovat:

1	0.929868	0.90576
0.929868	1	0.934574
0.90576	0.934574	1

R

⇒

8.020048	-5.28252	-2.32733
-5.28252	**11.38007**	-5.85082
-2.32733	-5.85082	**8.576028**

$\mathbf{R^-}$

Diagonální prvky (VIF) jsou zvýrazněny tučně. Vidíme, že v jednom případě je hodnota 10 překročena, v ostatních případech se jí VIF značně blíží. Potvrzuje to tedy závěry Scottova kritéria, že v daném modelu je silná multikolinearita, která si nezbytně vyžaduje úpravu modelu.

10.8 Regresní diagnostika

MNČ plně vyhovuje pouze v případech, kdy jsou splněny její předpoklady podle kapitoly 10.5.2.2. Pokud tyto předpoklady nejsou splněny, potom MNČ nedává nejlepší nevychýlené odhady regresních parametrů. Problémy mohou nastat v kterékoli složce tzv. **regresního tripletu - data, model a metoda odhadu**.

Regresní diagnostika tedy zkoumá (MELOUN - MILITKÝ 1994):

- **kvalitu dat** pro navržený model
- **kvalitu modelu** pro daná data
- **splnění předpokladů MNČ**

10.8.1 Analýza reziduí

Analýza reziduí (odchylek naměřených a modelových hodnot skutečného modelu) je častou metodou analýzy regresního modelu. Vychází se z předpokladu, že rezidua e_i mají stejné vlastnosti jako chyby ε_i (které vyjadřují náhodnou složku teoretického, ideálního modelu). Tento předpoklad nebývá často splněn. Hlavní odchylky jsou v následujících vlastnostech:

- rezidua jsou korelovaná, i když chyby jsou nezávislé;
- rezidua mají nekonstantní rozptyl;
- neindikují správně vybočující body (bod s nejvyšším reziduem nemusí být vlivný);
- vykazují vyšší stupeň normality než chyb (tzv. efekt supernormality).

Proto je vhodné používat různé speciální typy reziduí (podrobněji např. MELOUN-MILITKÝ 1994).

K určitým účelům může být velmi názorná a vhodná grafická analýza reziduí. Použitívají se tři typy grafů:

Typ grafu	Osa X	Osa Y
I	pořadové číslo bodu i	reziduum e_i
II	j-*tá* nezávislá proměnná x_j	reziduum e_i
III	vypočítaná (modelová) hodnota y'_i	reziduum e_i

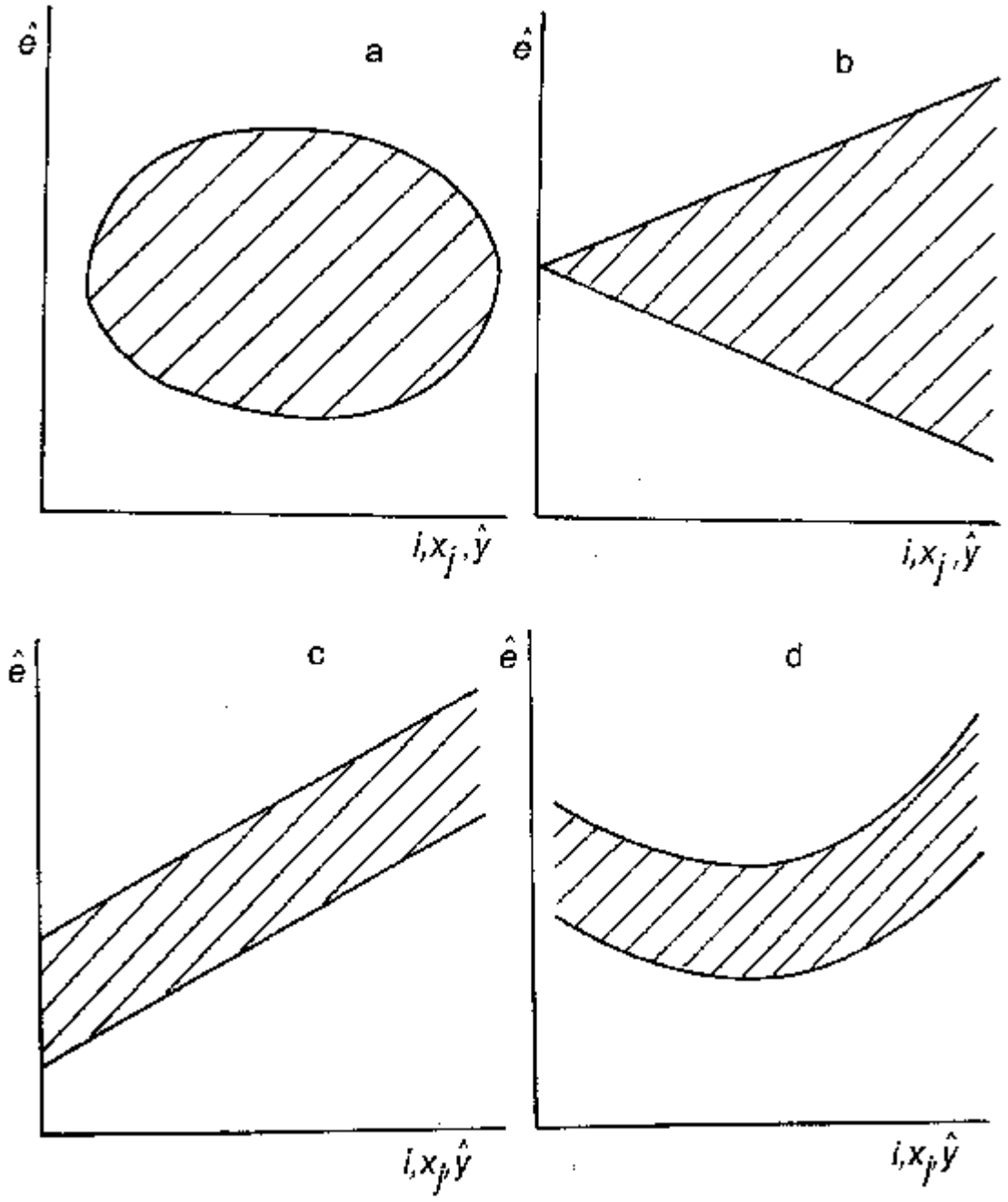
Základní typy obrazců grafů I – III jsou na obrázku 10.20 .

Základním tvarem všech tří typů je „**mrak“ bodů** (A), což je indikace „bezproblémového“ modelu.

Tvar **klínu** (B) ve všech třech typech indikuje heteroskedasticitu (nekonstantnost rozptylu) závisle proměnné (obvykle pomůže transformace, např. logaritmická, nebo použití modifikované MNČ).

Tvar **pásu** (C) u grafů I. typu indikuje chybu ve výpočtu nebo přítomnost vybočujících bodů, u typu grafu II. nepřítomnost proměnné x_j v modelu, u typu III je to upozornění na možnou chybu ve výpočtu nebo na chybějící absolutní člen.

Nelineární (D) tvar upozorňuje ve všech třech případech na nesprávně navržený model.



Obrázek 10.20 – Nejčastější tvary obrazce bodů v grafické analýze reziduí (podle MELOUN-MILITKÝ 1994)

10.8.2 Posouzení kvality dat

Při posouzení kvality dat se sleduje především výskyt tzv. **vlivných bodů** v závislosti na použitém modelu. Problematika vlivných bodů je velmi složitá, protože na jedné straně mohou velmi zkreslit odhady a zvětšit rozptyl parametrů tak, že model je

prakticky nepoužitelný, ale na druhé straně v určitých případech mohou zlepšit predikční schopnosti modelu. Vlivné body se v zásadě dělí do tří skupin:

- **hrubé chyby** - jsou způsobeny chybou měření nebo pozorování, dělí se na dvě skupiny:
  - *vybočující pozorování* - jsou způsobeny extrémní hodnotou měřené veličiny (projeví se na ose y);
  - *extrémy* - jsou způsobeny nevhodným nastavením vysvětlujících proměnných (projeví se extrémní hodnotou na ose x);
- **body s vysokým vlivem** (tzv. „zlaté body“) jsou speciálně vybrané body, které byly přesně změřeny a zpravidla zlepšují predikční schopnosti modelu;
- **zdánlivě vlivné body** - jsou způsobeny nevhodným modelem;

Je nutné podotknout, že v praktických úlohách ne vždy pracujeme s řízenými experimenty, a proto možnost nastavení vysvětlujících proměnných je malá. V těchto případech dělení na vybočující body a extrémy není podstatné, jedná se prostě o vybočující (podezřelé) hodnoty. **Jejich význam pro daný regresní model musí být velmi odpovědně posouzen a příslušné údaje z modelu vypuštěny pouze tehdy, je-li zcela zřejmé, že se jedná o závažné hrubé chyby měření**.

Vlivné body se určují různými metodami, z nichž uvedeme dvě základní:

- pomocí diagonálních prvků projekční matice H_{ii},
- pomocí speciálních grafických metod

10.8.2.1 Analýza prvků projekční matice

Diagonální prvky projekční matice H_{ii} (viz vzorec 10.40) obecně nabývají hodnot v rozmezí 0 - 1. Platí, že čím víc se prvky H_{ii} blíží jedné, tím je jejich vliv na predikci silnější a tím jsou vlivnější. Pro citlivější posouzení vlivných bodů se používá rozšířená projekční matice

$$H_{ii}^{*} = H_{ii} + \frac{e_i^2}{(n-m)\cdot\sigma^2} \qquad (10.85)$$

kde je e_i *i-té* reziduum a σ^2 reziduální rozptyl.

10.8.2.2 Grafy identifikace vlivných bodů

Z mnoha grafů identifikace vlivných bodů vybíráme dva nejjednodušší:

10.8.2.2.1 Graf predikovaných reziduí

Graf se zkonstruuje tak, že na osu X se vynesou predikovaná rezidua, na osu Y „klasická“ rezidua (tj. rozdíly experimentálních a vypočítaných hodnot).

Predikovaná rezidua se vypočítají

$$e_{Pi} = \frac{e_i}{1-H_{ii}} \qquad (10.86)$$

Interpretace grafu je velmi jednoduchá:

- pokud v datech nejsou žádné vybočující body, leží body grafu na přímce $y = x$
- pokud jsou v datech extrémy, potom tyto body leží výrazně mimo přímku $y = x$

- pokud jsou v datech vybočující hodnoty (na ose Y), leží sice na přímce, ale ve větší vzdálenosti od mraku ostatních bodů

Schéma grafu ukazuje obrázek 10.21 .

10.8.2.2.2 Williamsův graf

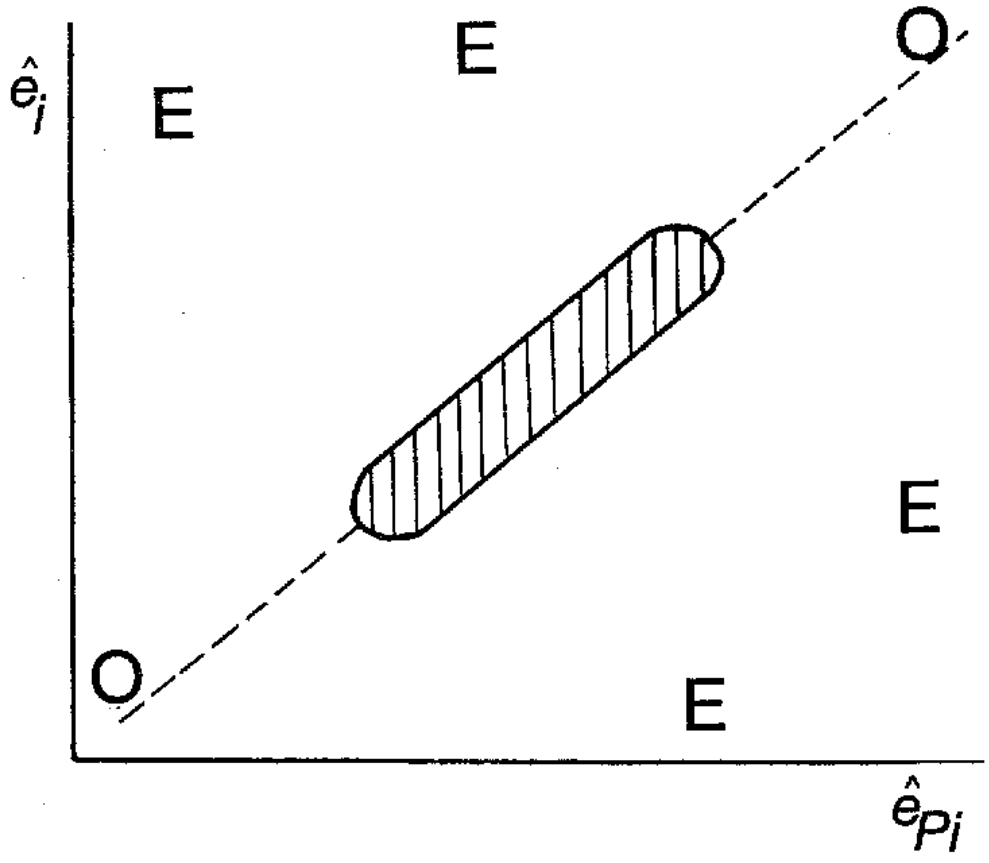
se sestrojí tak, že na osu X se vynášejí hodnoty H_{ii} (diagonální prvky projekční matice) a na osu Y Jackknife rezidua. Dále se zakreslí mezní linie pro vybočující body $y = t_{0.95;n-m-1}$ a pro extrémy $x = 2\ m / n$.

Jackknife rezidua se vypočítají

$$e_{Ji} = e_{Si} \cdot \sqrt{\frac{n-m-1}{n-m-e_{Si}}} \qquad (10.87)$$

kde e_{Si} jsou standardizovaná rezidua

$$e_{Si} = \frac{e_i}{\sigma\sqrt{1-H_{ii}}} \qquad (10.88)$$



Obrázek 10.21 - Schéma interpretace grafu predikovaných reziduí. O jsou vybočující měření, E jsou extrémy (podle MELOUN - MILITKÝ 1994)

Interpretace grafu je velmi jednoduchá:

- pokud v datech nejsou žádné vybočující body, leží body grafu uvnitř mezních linií;
- pokud jsou v datech extrémy, potom tyto body leží nad mezní linií y;
- pokud jsou v datech vybočující hodnoty, leží vpravo od mezní linie x;
- pokud jsou v datech takové body, které jsou jak vybočujícími hodnotami, tak i extrémy, leží tyto body šikmo vpravo nahoru od průsečíku mezních linií.

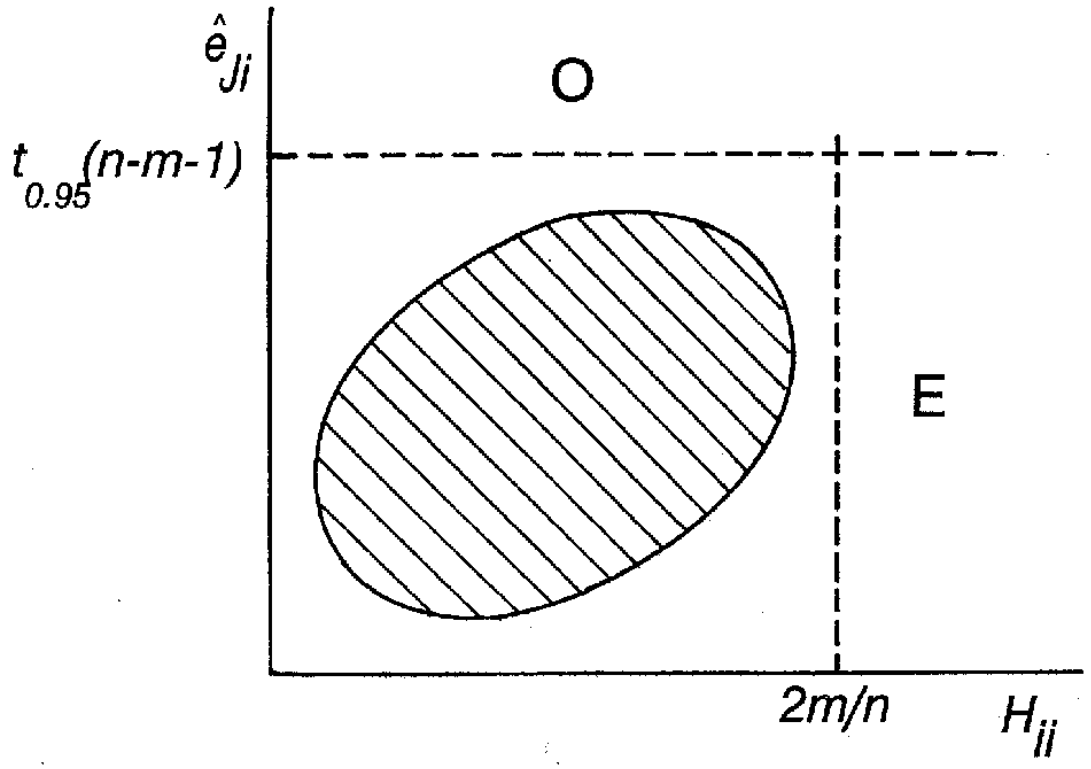
Schéma grafu je na obrázku 10.22 .

Příklad 10.15:

Pro Výběr C z příkladu 10.8 proveďte identifikaci vlivných bodů.

Již ze zadání tohoto výběru je zřejmé, že vlivným bodem je bod č. 10. Použijeme metodu rozšířené projekční matice a diagnostických grafů a posoudíme, nakolik jsou schopny vlivný bod detekovat.

Z tabulky 10.13 vyplývá, že rozšířená diagonální matice podle vzorce 10.85 svou hodnotou 1 indikuje výrazně vybočující bod (z tabulky je vidět, že původní diagonální prvek není zdaleka tak citlivý).



Obrázek 10.22 - Schéma konstrukce a interpretace Williamsova grafu. O jsou vybočující měření, E jsou extrémy (podle MELOUN - MILITKÝ 1994)

Číslo bodu	Diagonální prvky původní projekční matice	Diagonální prvky rozšířené projekční matice
1	0.3182	0.3292
2	0.2364	0.2402
3	0.1727	0.1732
4	0.1273	0.1277
5	0.1000	0.1039
6	0.0909	0.1020
7	0.1000	0.1212
8	0.1273	0.1618
9	0.1727	0.2252
10	0.2364	**1.0000**
11	0.3182	0.4158

Tabulka 10.13- Hodnoty diagonálních prvků projekční matice pro Výběr C

Dále použijeme graf predikovaných reziduí a Williamsův graf. Na obrázku 10.23 vidíme, že oba grafy indikují silný vliv bodu č. 10.

10.8.3 Posouzení kvality navrženého regresního modelu

V případě **jedné nezávisle proměnné** je situace zpravidla jednoduchá - stačí sestrojit tzv. **rozptylový graf**, tj. vynést hodnoty závisle proměnné proti nezávisle proměnné a podle výsledného mraku bodů posoudit vhodnost navrženého modelu.

V případě **více vysvětlujících proměnných** je problém složitější. Do rozhodování vstupují různé interakce mezi vysvětlujícími proměnnými (např. multikolinearita) a zde mohou být prosté rozptylové grafy zavádějící. V takových případech se používají speciální metody, z nichž uvedeme jen několik nejdůležitějších.

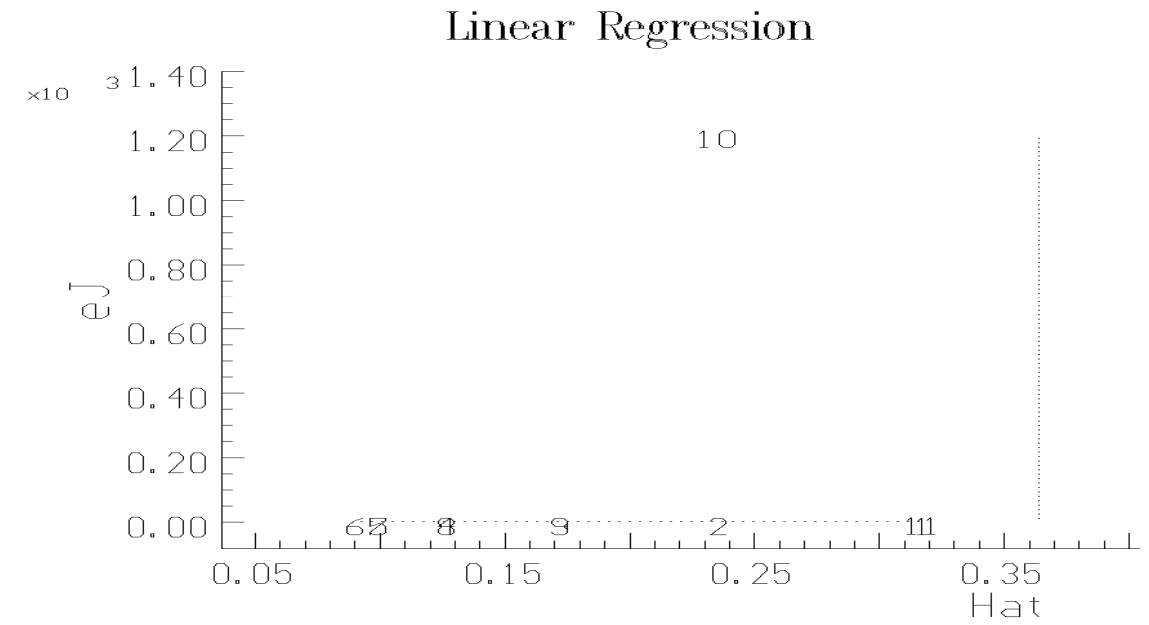
Poměrně jednoduchým diagnostickým prostředkem může být graf reziduí (osa Y) proti vysvětlované (závislé) proměnné (osa X). Jestliže model je nevhodný, potom rezidua v grafu tvoří nelineární obrazec (zpravidla tvaru U).

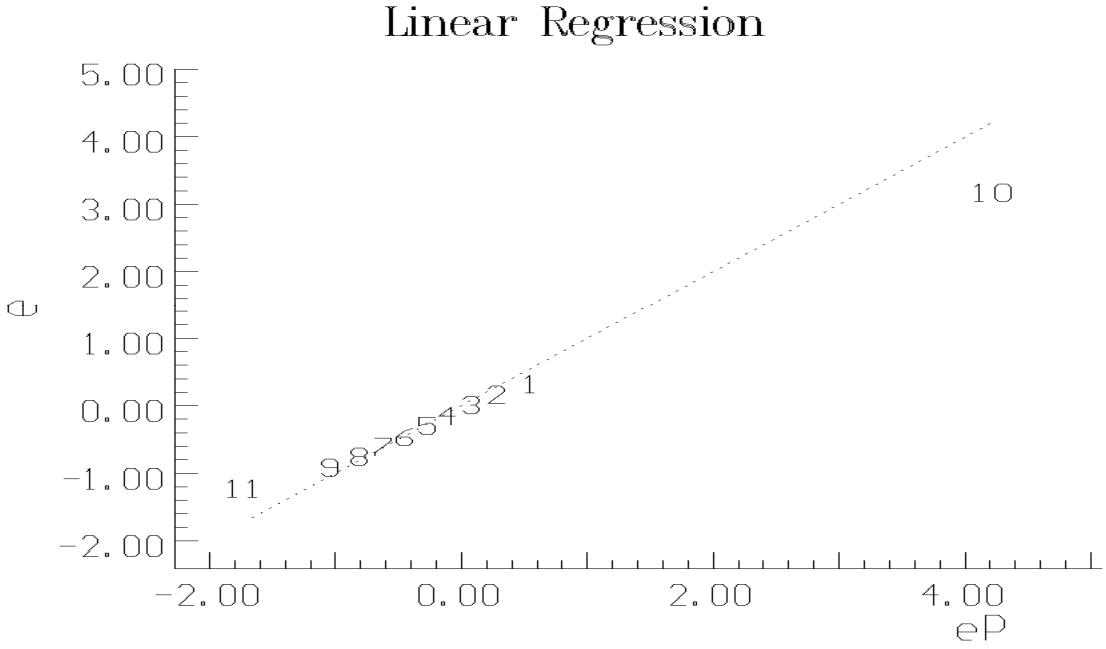
V případě potřeby detailnějšího rozboru regresního modelu se používají, kromě jiných, dva typy grafů (MELOUN - MILITKÝ 1994):

- **parciální regresní grafy,**
- **parciální reziduální grafy**

10.8.3.1 Parciální regresní grafy

Jedná se o jeden ze základních diagnostických grafů, protože kromě posouzení kvality regresního modelu v určitých případech umožňují i indikaci dalších podstatných vlastností.





Obrázek 10.23 - Williamsův graf (nahoře) a graf predikovaných reziduí (dole) pro Výběr C. Čísla značí pořadová čísla jednotlivých hodnot. Odloučenost bodu č. 10 od ostatních indikuje, že je vlivný

Parciální regresní graf vyjadřuje **závislost mezi vysvětlovanou proměnno**u (tedy vektorem **y**) a **jednou vysvětlující proměnnou x_j při statisticky neměnném vlivu ostatních vysvětlujících proměnných**, které tvoří matici $\mathbf{X_{(j)}}$ (tento symbol označuje matici vysvětlujících proměnných s vynechanou *j*-tou proměnnou). Je to tedy určitá grafická obdoba parciálního korelačního koeficientu u korelačních modelů.

Podrobné teoretické odvození parciálního regresního grafu viz (MELOUN - MILITKÝ 1994). Zde se budeme zabývat pouze jeho sestrojením a interpretací.

Parciální regresní graf se sestrojí následujícím způsobem:

- určíme vysvětlující proměnnou $\mathbf{x_j}$, kterou budeme analyzovat,
- provedeme regresi, kde $\mathbf{x_j}$ bude vysvětlovaná (závisle) proměnná proti zbylým vysvětlujícím proměnným $\mathbf{X_{(j)}}$. Rezidua tohoto regresního modelu nazveme $\mathbf{v_j}$ a budou tvořit hodnoty na ose X parciálního regresního grafu,
- provedeme regresi vysvětlované (závislé) proměnné **y** na nezávisle proměnných $\mathbf{X_{(j)}}$. Rezidua tohoto regresního modelu nazveme $\mathbf{u_j}$ a budou tvořit hodnoty na ose Y parciálního regresního grafu.

Interpretace parciálního regresního grafu je následující:

- pokud body parciálního regresního grafu leží na přímce s nulovým úsekem (absolutním členem), potom existuje skutečná lineární závislost mezi y a $\mathbf{x_j}$
- směrnice přímky proložené body parciálního regresního grafu číselně odpovídá příslušnému regresnímu koeficientu b_j původního (posuzovaného) regresního modelu
- korelační koeficient mezi $\mathbf{u_j}$ a $\mathbf{v_j}$ odpovídá parciálnímu korelačnímu koeficientu $R_{yx_j(\mathbf{x_{(j)}})}$
- rezidua regresní přímky mezi $\mathbf{u_j}$ a $\mathbf{v_j}$ odpovídají reziduím původního modelu

10.8.3.2 Parciální reziduální grafy

Je to analogie parciálního regresního grafu, kdy graf zobrazuje přímo závislost parciálních reziduí **s** na $\mathbf{x_j}$.

V grafu se znázorňují dvě složky:

- deterministická komponenta C, kde $c_{ij} = (x_{ij} - \bar{x}_j) \cdot b_j$
- vlastní parciální reziduum s, kde $s_i = c_{ij} + e_i$

Pro parciální reziduální grafy platí:

- pokud je příslušná $\mathbf{x_j}$ vhodně do modelu zařazena, potom je závislost s na $\mathbf{x_j}$ lineární s nulovým absolutním členem, přičemž směrnice této regrese je číselně rovna b_j
- rezidua této regresní přímky se rovnají reziduím původního modelu

Parciální reziduální grafy se používají především ke stanovení správnosti zařazení určité proměnné do modelu a k indikaci případných nelinearit v případě nesprávně navrženého modelu.

10.8.4 Ověření předpokladů MNČ

MNČ je nejběžnější metodou výpočtu regresních parametrů a za předpokladu dodržení podmínek uvedených v kapitole 10.5.2.2 dává jejich nejlepší nevychýlené

odhady. Pokud tyto předpoklady nejsou dodrženy, odhady získané pomocí klasické MNČ nejsou zcela korektní. Zde si uvedeme pouze základní numerické a grafické metody k odhalení různých porušení předpokladů MNČ V takovýchto případech je zpravidla nutné použít různým způsobem upravené MNČ. Vzhledem k tomu, že možných modifikací MNČ je celá řada a jejich použití závisí na typu odchylky od klasické MNČ, jejich podrobný rozbor přesahuje rozsah tohoto textu. Podrobnosti včetně řešených příkladů uvádí např. MELOUN - MILITKÝ 1994. V této kapitole se pouze zmíníme o dvou častých komplikacích, se kterými se u regresních modelů můžeme setkat - s heteroskedasticitou a autokorelací chyb **ε**.

10.8.4.1 Heteroskedasticita

Heteroskedasticita (nekonstantnost rozptylu) se u měřených dat vyskytuje poměrně často. Za předpokladu relativní konstantní přesnosti měření bývá rozptyl rostoucí funkcí velikosti proměnné y. V tomto případě se identifikuje diagnostickým grafem závislosti e_i^2 (kvadráty reziduí) na y'_i (predikovaných – vypočítaných - hodnotách). V případě heteroskedasticity tohoto typu vzniká obrazec s výrazným trendem (lineárním nebo nelineárním).

V mnoha případech se vychází z představy, že rozptyl naměřené hodnoty y_i je určitou funkcí proměnné $\mathbf{x_i}\,\beta$ (např. exponenciální). V tomto případě se používá **Cookův - Weisbergův test**

$$S_f = \frac{\left[\sum_{i=1}^{n}(y'_i - \bar{y}')^2 e_i^2\right]^2}{2\cdot\sigma^4 \sum_{i=1}^{n}(y'_i - \bar{y}')^2}, \qquad (10.89)$$

kde $\bar{y}'$ je aritmetický průměr predikovaných hodnot. Pokud v datech není heteroskedasticita, potom platí, že $S_f < \chi^2(1)$ (kvantil chi-kvadrát rozdělení s jedním stupněm volnosti).

Problematika identifikace, stanovení typu heteroskedasticity a následného výpočtu parametrů regresního modelu je složitá a není ji zde možné podrobně rozvádět (viz např. MELOUN - MILITKÝ 1994) . Nejjednodušší metodou, jak vypočítat parametry regresního modelu pro data zatížená heteroskedasticitou, je metoda **vážených nejmenších čtverců**. Určení vah modelu se provádí různými způsoby. Jednoduchým, ale v mnoha případech dostatečným způsobem, je užití převrácených hodnot závisle proměnné, tj.$1/y_i$.

10.8.4.2 Autokorelace

Autokorelace vzniká u dat, která mají charakter časových řad. Jedná se vlastně o závislost rezidua s předchozími rezidui. Podle délky posunutí hovoříme např. o autokorelaci I. řádu (závislost e_i na e_{i-1}), II. řádu (závislost e_i na e_{i-2}) apod. Nejvýznamnější a nejčastější je autokorelace I. řádu. Graficky se dá odhalit jako závislost e_i na e_{i-1} - pokud je v grafu výrazná lineární závislost, je to důkaz autokorelace reziduí. Je nutné upozornit, že u malých výběrů dochází často k tomu, že rezidua jsou korelovaná i tehdy, jestliže chyby ε korelované nejsou. Proto se doporučuje používat tzv. rekur-

zívní rezidua. Autokorelaci lze také testovat některými testy, např. Waldovým nebo Durbinovým - Watsonovým testem - viz (MELOUN - MILITKÝ 1994).

10.8.4.3 Normalita chyb

K ověření normality se může použít testů uvedených v kap. 3., z grafických technik se nejčastěji používají rankitové grafy. Kromě těchto technik se u regresních modelů používá **Jarque-Berrův test** (viz MELOUN - MILITKÝ 1994).

10.8.5 Stanovení vhodného regresního modelu na příkladu

Příklad 10.16:

Pro data z příkladu 10.2 *využijte technik korelační a regresní analýzy a regresní diagnostiky a navrhněte optimální regresní model s využitím metody nejmenších čtverců.*

Pro závislost objemu na výčetní tloušťce, výšce a délce zelené koruny budeme předpokládat lineární regresní model

$$v = b_0 + b_1 d + b_2 h + b_3 k.$$

Při použití MNČ dostaneme následující podobu regresního modelu:

Parametr	Odhad parametru	t-kritérium	Významnost parametru ($t_{0.025,47} = 2.013$)
b_0	-0.090 28	- 9.683	významný
b_1	0.010 57	10.431	významný
b_2	0.002 27	1.392	nevýznamný
b_3	0.002 08	1.497	nevýznamný

Výsledky korelační analýzy z příkladu 10.2 nám daly tyto výsledky:

Charakteristika korelace	Hodnota
Vícenásobný korelační koeficient	0.984 45
Vícenásobný koeficient determinace	0.969 15
Parciální korelační koeficient II. řádu objem - tloušťka	0.838 36
Parciální korelační koeficient II. řádu objem - výška	0.201 09
Parciální korelační koeficient II. řádu objem - délka zelené koruny	0.215 58

Následující tabulka uvádí výsledky dalších důležitých testů:

Testovaná vlastnost	Testové kritérium	Vypočítaná hodnota testového kritéria	Kritická hodnota testu	Výsledek testu
významnost modelu	test podle vztahu 10.61	481.69	2.807	model je významný
multikolinearita	Scottův test	0.75		navržený model není korektní
heteroskedasticita	Cook - Weisbergův test	3.89	3.84	rezidua vykazují heteroskedasticitu

Z dalších testů (které zde pro úsporu místa neuvádíme) by vyplynulo, že nezávislost a normalita reziduí je dodržena.

Jaké hodnocení modelu můžeme z těchto podkladů udělat? Jak výsledky testů významnosti parametrů modelu, tak i parciální korelační koeficienty ukazují, že proměnné *výška* a *délka zelené koruny* nejsou v daném modelu významné. Dalšími „velkými" problémy jsou multikolinearita (svědčí o lineární závislosti vysvětlujících proměnných) a heteroskedasticita (svědčí o nekonstantnosti rozptylu).

Je tedy zřejmé, že dvě vysvětlující proměnné by bylo vhodné z modelu vyloučit. Tím se celý model výrazně zjednoduší a také bude odstraněn problém multikolinearity.

Pro ilustraci si ukážeme také grafické techniky posuzování modelu - parciální regresní grafy. Obrázky 10.24 a 10.25 zobrazují parciální regresní grafy postupně vzhledem k proměnné tloušťka, výška a délka zelené koruny. Je vidět, že pouze první proměnná - tloušťka - vykazuje přibližně lineární trend a tedy je vhodným členem tohoto modelu. Ostatní proměnné vytvářejí mrak bodů, takže pro ně tento graf lineární model nedoporučuje.

Zjednodušený regresní model má tuto podobu:

Parametr	**Odhad parametru**	**t-kritérium**	**Významnost parametru ($t_{0.025,49} = 2.011$)**
b_0	-0.086 49	- 22.214	významný
b_1	0.013 57	34.994	významný

s následujícími charakteristikami korelace:

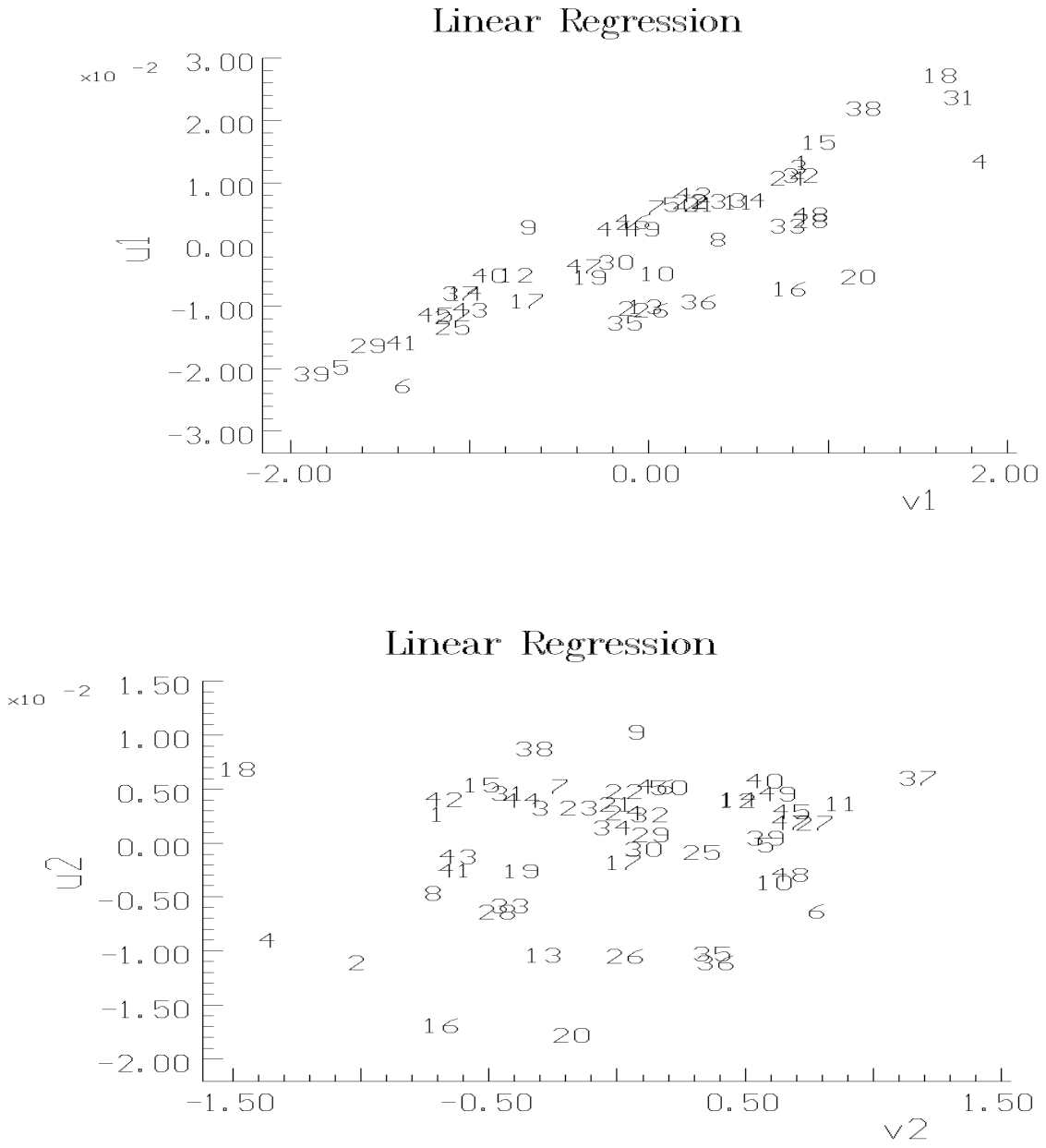
Charakteristika korelace	**Hodnota**
Vícenásobný korelační koeficient	0.980 96
Vícenásobný koeficient determinace	0.962 28

Další testy poskytly tyto závěry:

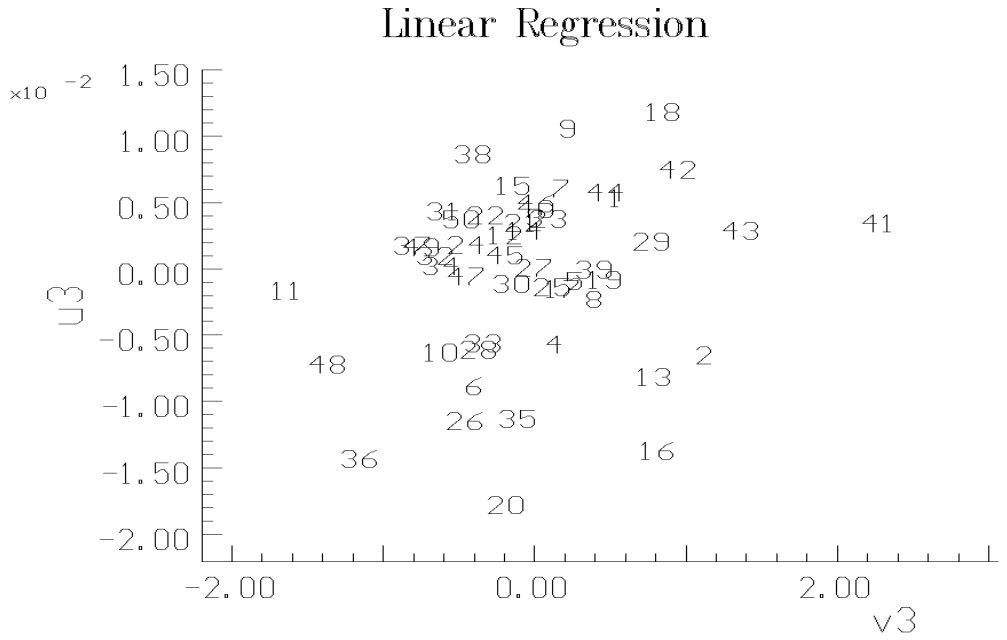
Testové kritérium	**Testovaná vlastnost**	**Vypočítaná hodnota testového kritéria**	**Kritická hodnota testu**	**Výsledek testu**
test 10.61	významnost modelu	1224.6	4.043	model je významný
Cook - Weisbergův test	heteroskedasticita	6.597	3.842	rezidua vykazují heteroskedasticitu

Z výsledků korelační a regresní analýzy je zřejmé, že i přes zjednodušení modelu korelační koeficient poklesl pouze nepatrně, což potvrzuje fakt, že ostatní proměnné neměly statisticky významný vliv. Stále přetrvává problém heteroskedasticity. Tento problém je možné vyřešit např. metodou **vážené MNČ** (podrobnosti viz MELOUN-MILITKÝ 1994), kde použijeme váhu **1/y**. Výpočet pomocí vážené MNČ je nutné

provádět pomocí specializovaného softwaru – statistických programů (např. AD-STAT).



Obrázek 10.24 - Parciální regresní grafy: nahoře pro tloušťku, dole pro výšku. Interpretace viz v textu.



Obrázek 10.25 - Parciální regresní graf pro proměnnou „délka koruny". Interpretace viz v textu.

Regresní model vypočítaný metodou vážených nejmenších čtverců poskytl tyto výsledky:

Parametr	Odhad parametru	t-kritérium	Významnost parametru ($t_{0.025,49}$= 2.011)
b_0	-0.069 65	- 23.902	významný
b_1	0.011 66	30.332	významný

s následujícími charakteristikami korelace:

Charakteristika korelace	Hodnota
Vícenásobný korelační koeficient	0.974 89
Vícenásobný koeficient determinace	0.950 41

Další testy poskytly tyto závěry:

Testovaná vlastnost	Testové kritérium	Vypočítaná hodnota testového kritéria	Kritická hodnota testu	Výsledek testu
významnost modelu	test 10.61	920	4.043	model je významný
heteroskedasticita	Cook - Weisbergův test	1.114	3.842	rezidua vykazují homoskedasticitu

Ukázalo se, že jednoduchá metoda vážené MNČ dostatečně odstranila heteroskedasticitu v datech a v rámci daných možností může být tento model považován za nejlepší.

O vhodnosti modelu svědčí i hodnoty MEP a AIC kritéria:

Model	MEP	AIC
původní	$4.8303 \cdot 10^{-5}$	- 498.97
zjednodušený (klasická MNČ)	$5.3958 \cdot 10^{-5}$	- 492.92
zjednodušený (vážená MNČ)	$2.6098 \cdot 10^{-5}$	- 529.31

Vzhledem k tomu, že u obou kritérií platí, že čím menší hodnota, tím vhodnější model, i zde se potvrzuje, že poslední varianta je nejvhodnější.

Závěrem je nutné podotknout, že platnost tohoto modelu se omezuje na použití MNČ. Vzhledem k silnější multikolinearitě by se zde nabízelo použití tzv. metody racionálních hodností, které by vedlo k jiné podobě modelu a v dané situaci by zřejmě bylo vhodnější. Podrobnosti o této výpočetní metodě včetně příkladů použití viz např. (MELOUN - MILITKÝ 1994).

10.9 Nelineární regrese

Při modelování mnoha reálných systémů nevystačíme s lineárními regresními modely, neboť popisované závislosti mají průběh, které je možné popsat pouze složitějšími regresními modely nelineárního typu.

Jako příklad nám může sloužit růstová křivka – funkce vyjadřující závislost růstu (tj. změny nějaké růstové veličiny, např. výšky organismu) na věku. Pokud bychom použili jednoduchý lineární model, např. přímku, zjistili bychom, že se nedá smysluplně interpretovat a naprosto neodpovídá realitě: musili bychom připustit, že růst živého organismu probíhá stále stejně rychle, nikdy nekončí a roste nade všechny meze. To je samozřejmě nesmysl. Proto je v takovém případě nutné použít speciální tzv. růstovou funkci, která splňuje požadavky kladené na správní modelování růstu (např. má asymptotu – tj. růst je shora omezen, funkce má typický tvar protáhlého písmene S, má inflexní bod, ve kterém dosahuje rychlost růstu maxima, apod.). Matematický tvar růstové funkce se ovšem nedá vyjádřit jednoduchým lineárním modelem, musí se použít model nelineární.

Formálně považujeme za nelineární takové regresní modely, jejichž **parametry nejsou ve vzájemném lineárním postavení**. Jako příklad mohou sloužit modely $y = a \cdot x^b$, $y = a \cdot e^{bx}$ nebo třeba Korfova růstová funkce $y = A \cdot e^{\left[\frac{k}{(1-n)t^{n-1}}\right]}$. *A, a, b, n k* jsou parametry nelineárních funkcí, které musíme stanovit..

Výpočet parametrů těchto modelů je značně komplikovaný, je daleko složitější než u lineárních modelů. V podstatě také používáme MNČ a minimalizujeme součet reziduálních čtverců, ale problém je v tom, tato minimalizační funkce, tzv. účelová,

nemá jednoznačně definované minimum, může mít minim několik (kromě tzv. globálního, tj. skutečného minima pro celou funkci může mít ještě několik lokálních minim pro určité úseky funkce). Minimum hledáme pomocí numerických metod, které pracují iteračně: začnou s prvním odhadem parametrů (který musí zpravidla zadat uživatel), vypočítají první svůj odhad parametrů, tento odhad vezmou za základ nového výpočtu, provedou druhý odhad, a tímto způsobem pokračují tak dlouho, dokud nejsou splněny podmínky ukončující výpočet (to může být např. zadaná nepatrná změna součtu čtverců – pokud další výpočet nezlepší odhady parametrů, tj. součet čtverců se dále podstatněji nezmenšuje, výpočet je možné ukončit). Hlavním problémem je to, že při ukončení výpočtu nevíme, zda jsme opravdu v globálním (optimální řešení) nebo jen v lokálním minimu. Výpočetních algoritmů je celá řada (derivační, nederivační, speciální) a každý má své výhody a nevýhody (podrobněji k těmto metodám i k teorii nelineární regrese obecně viz MELOUN-MILITKÝ 1994). Obecně jsou tyto metody velmi citlivé na počáteční odhady parametrů. Některé z nich zcela selžou, pokud jsou tyto odhady zadány hodně „daleko“ od jejich skutečných hodnot. Tato situace je velmi častá, protože v mnohých případech neznáme ani přibližně „jak by to asi mohlo vyjít“ a odhady parametrů zadáváme v podstatě náhodně. Parametry mají v nelineární regresi velký význam a ve většině případů mají přesný fyzikální (reálný) smysl (na rozdíl od lineární regrese, kde to jsou mnohdy jen numerické koeficienty bez reálné interpretace). Kvalitní algoritmy se již s touto situací umějí vyrovnat lépe a dojdou k přijatelnému řešení i z velmi vzdálených odhadů. Z výše uvedených skutečností vyplývá, že výpočet parametrů nelineárních modelů je záležitostí kvalitních statistických programů.

Jednou z možností, jak určité nelineární funkce spočítat bez nutnosti použití iteračních algoritmů, je **linearizace**. Principem je následující postup:

- pomocí vhodné transformace se nelineární model převede na model lineární (např. substitucí, logaritmováním,apod.);
- běžnou MNČ se vypočítají parametry lineárního modelu;
- parametry lineárního modelu se převedou (retransformují) na původní nelineární model.

Je nutno zdůraznit, že linearizace zhoršuje některé statistické vlastnosti odhadů parametrů, proto se používá jen těch případech, kdy není k dispozici kvalitní program na výpočet nelineární regrese.

Mírou těsnosti závislosti u nelineárních modelů je **index korelace,** který se vypočítá

$$I_{yx} = \sqrt{\frac{S^2_{y'}}{S^2_y}} = \sqrt{1 - \frac{S^2_{yx}}{S^2_y}} \qquad (10.90)$$

kde je

$S^2_{y'}$ část celkového rozptylu vysvětlená regresním modelem podle vzorce

$$S^2_{y'} = \frac{\sum_{i=1}^{n} (y'_i - \bar{y})^2}{n}$$

S^2_y celkový rozptyl podle vztahu

$$S^2_{y'} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

S^2_{yx} část celkového rozptylu nevysvětlená regresním modelem (reziduální rozptyl) podle vzorce

$$S^2_{yx} = \frac{\sum_{i=1}^{n}(y_i - y'_i)^2}{n}$$

Interpretace indexu korelace je stejná jako v případě korelačního koeficientu, pouze neplatí rovnost při přehození proměnných, tedy **platí $I_{yx} \neq I_{xy}$.** Druhá mocnina indexu korelace se nazývá **index determinace** a stejně jako koeficient determinace vyjadřuje, jaká část celkového rozptylu je vysvětlena regresním modelem.

Příklad 10.17:

Stanovte parametry Michajlovovy růstové funkce pro zadané hodnoty výšky pomocí linearizace i pomocí statistického programu. Měřené hodnoty jsou v tabulce 10.14 .

Věk	Výška stromu (m)	Růstová funkce (m)
10	3.7	3.0
15	6.4	6.7
20	8.9	10.0
25	11.2	12.8
30	13.3	15.0
35	15.2	16.8
40	16.9	18.3
45	18.4	19.5
50	19.8	20.6
55	21.1	21.5
60	22.3	22.3
65	23.4	23.0
70	24.4	23.6
75	25.3	24.1
80	26.1	24.6
85	26.9	25.0
90	27.6	25.4
95	28.3	25.8
100	29.0	26.1

Tabulka 10.14 – Zadané hodnoty výšky (vlevo) a vypočítané hodnoty modelu – Michajlovovy růstové funkce (vpravo)

Nejprve provedeme výpočet linearizací. Michajlovova růstová funkce má tvar

$$y' = a \cdot e^{\frac{k}{t}}$$

který je možné snadno převést na lineární tvar $Y = A + B \cdot X$ logaritmováním

$$\ln y = \ln a + k \cdot (1/t) \cdot \ln e.$$

kde $\ln y = Y$; $\ln a = A$; $k = B$; $1/t = X$ a $\ln e = 1$

Znamená to, že do lineární regrese nevstupují původní hodnoty „věk“ a „výška“, ale transformované hodnoty: jako nezávisle proměnná to bude *1/t* a jako závisle proměnná *ln(h)*, kde h je výška.

Běžnou MNČ se vypočítají koeficienty A = 3.502 a B = -23.909. Tyto koeficienty se musí retransformovat na koeficienty původní nelineární rovnice:

$$\ln a = A \Rightarrow a = e^A$$
$$k = B$$

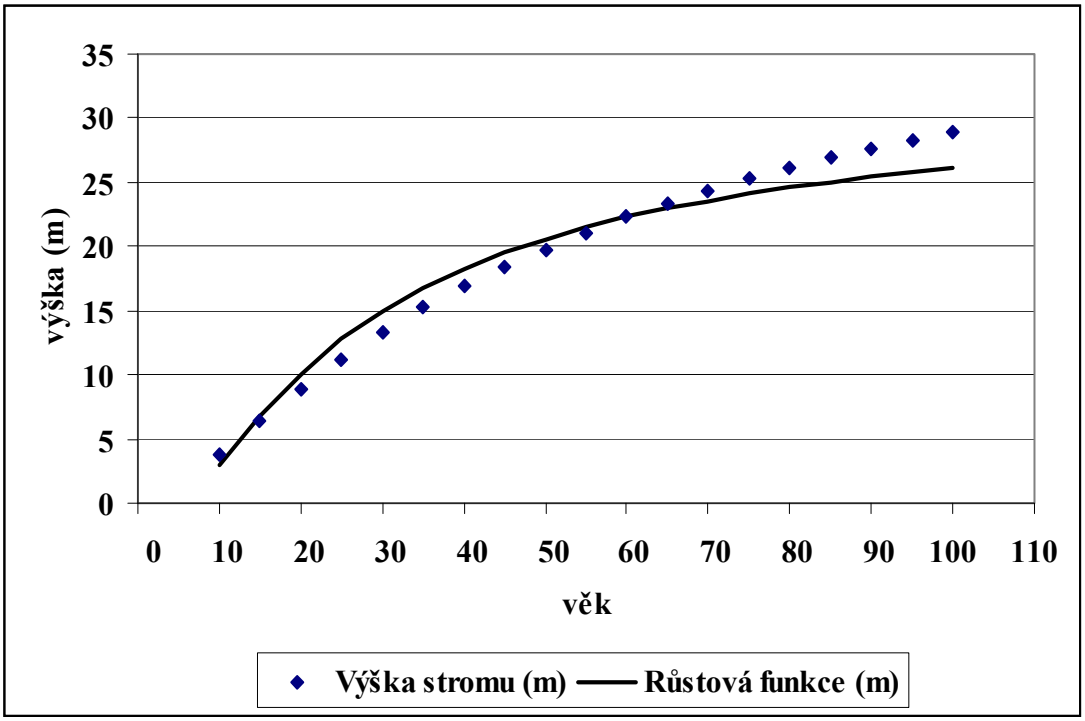
Výsledné hodnoty koeficientů tedy jsou $a = e^{3.502} = 33.179$ a $k = B = -23.909$. Tyto hodnoty se dosadí do **původní** (nelineární) rovnice růstové funkce a vypočítají se modelové hodnoty (jsou uvedeny v pravém

sloupci tabulky 10.14 jako „růstová funkce“). Pomocí vztahu 10.90 se stanoví míra těsnosti závislosti I_{yx} pomocí výpočtu

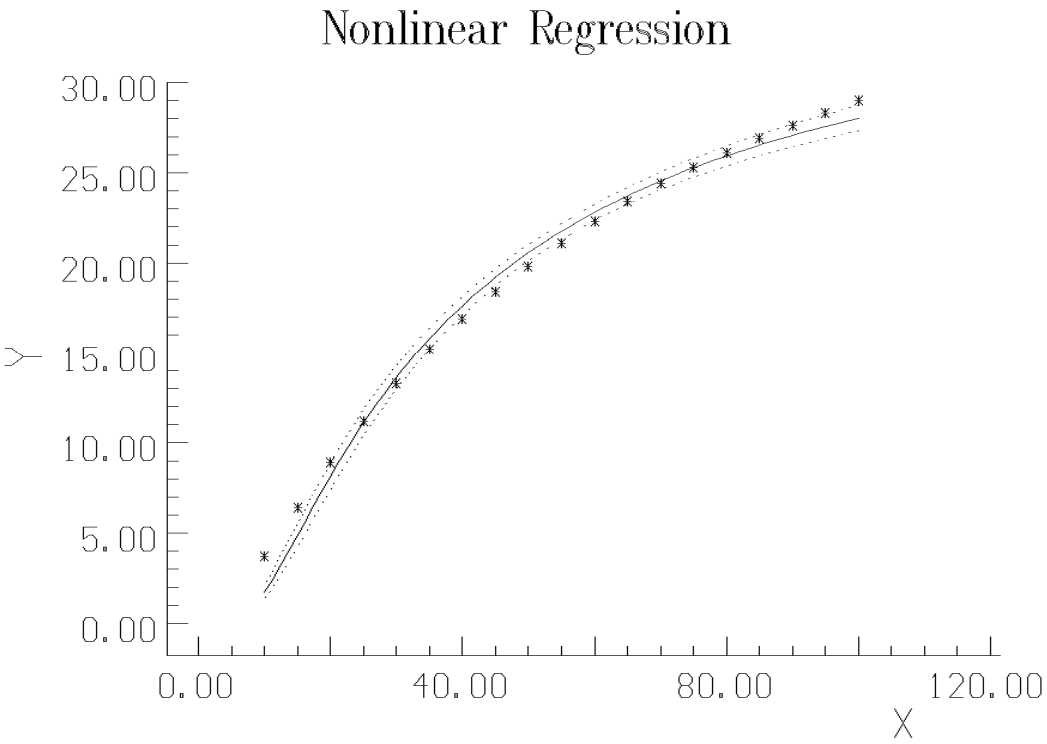
$$I_{yx} = \sqrt{1 - \frac{40.322}{1082.092}} = 0.981$$

Grafické znázornění výsledné růstové funkce je na obrázku 10.26 . Je zřejmé, že výpočet pomocí linearizace dosáhl kvalitní výsledek s vysokou mírou shody s naměřenými daty.

Pokud použijeme statistický program (v tomto případě ADSTAT), musíme zadat tvar modelu (podle zadané syntaxe P1*EXP(P2/X1)) a počáteční odhady parametrů. Pokud známe reálný význam koeficientů, je to snazší, protože jsem schopni odhadnout, v jakých mezích se hodnota může pohybovat. V našem případě je koeficient *a* asymptota funkce, tj. maximálně teoreticky dosažitelná hodnota výšky. Zadáme tedy číslo vyšší než nejvýše naměřená hodnota, např. 35 m, druhý koeficient *k* je koeficient ovlivňující tvar křivky a obvykle vychází jako záporné číslo řádově v desítkách, zadáme např. –25. Po spuštění výpočtu se po několika iteracích objeví výsledek – *a* = 38.151 a *k* = -30.878 a míra těsnosti závislosti je 0.989. Pokud nejsou známy odhady koeficientů, kvalitní algoritmus si poradí i tímto problémem, např. jestliže byly v tomto příkladu zadány oba první odhady rovny 1, program došel ke stejnému výsledku. Výsledek proložení pomocí statistického programu je na obrázku 10.27 .



Obrázek 10.26 – Růstová funkce vypočítaná pomocí linearizace



Obrázek 10.27 – Růstová funkce vypočítaná pomocí statistického programu

11 Použitá a doporučená literatura (pro I. i II.díl)

ANDĚL, J., 1978: Matematická statistika. Praha, SNTL -Alfa .

BENEDÍK, J., 1989: Biostatistika. Brno, UJEP, 233 s.

CIPRA, T., 1986: Analýza časových řad s aplikacemi v ekonomii. Praha, SNTL-Alfa

CYHELSKÝ, L., NOVÁK,I., 1967: Statistika. Praha, SNTL, 288 s.

ČERMÁK, V., 1968: Statistika. Praha, SNTL, 208 s.

DRÁPELA, K., ZACH, J., 1995: Dendrometrie (dendrochronologie). Skriptum MZLU Brno, 152 s.

DRÁPELA, K. ZACH, J., 1996: Biometrie (biostatistika) – vybrané části, Skriptum MZLU Brno, 153 s.

GROFÍK, R. a kol., 1987: Štatistika. Bratislava, Príroda, 520 s.

HALD, A., 1956: Matematičeskaja statistika s techničeskimi priloženijami. Moskva, Izdavatělstvo inostrannoj litertury, 664 s.

HÁTLE, J., LIKEŠ, J., 1972: Základy počtu pravděpodobnosti a matematické statistiky, Praha, SNTL, 464 s.

HAVRÁNEK, T. 1993: Statistika pro biologické a lékařské vědy. Academia, Praha.

HEBÁK, P., KAHOUNOVÁ, J., 1988: Počet pravděpodobnosti v příkladech. SNTL, Praha, 312 s.

CHAMBERS. J.M. a kol., 1983: Graphical Methods for Data Analysis. Belmont, Duxburry Press.

CHATFIELD, C., 1984: The Analysis of Time Series. An Introduction. London, Chapman and Hall, 286 s.

KENDALL, M. G., STUART, A. 1966: The Advanced Theory of Statistics. New York.

KUBÁČEK, L., PÁZMAN, A., 1979: Štatistické metódy v meraní. Bratislava, Veda, 148 s.

LAAR, A., 1979: Biometrische Methoden in der Forstwissenschaft. München, 633 s.

LEPORSKÝ, A., 1953: Statistické metody. Učební texty vysokých škol. Lesnická fakulta VŠZ Brno, SPN, Praha

MEAD, R. 1988: The design of experiments. Statistical pronciples for practical application. Cambridge Uneversity Press, Cambridge.

MELOUN, M., MILITKÝ, J., 1994: Statistické zpracování experimentálních dat. Praha, Plus, 839 s.

MICHÁLEK a kol., 1982: Biometrika. Praha, SPN, 404 s.

MINAŘÍK, B., 1995: Statistika I pro ekonomy a manažery. Skriptum MZLU Brno, 160 s.

MINAŘÍK, B.,1996: Statistika II pro ekonomy a manažery. Skriptum MZLU Brno, 144 s.

MINAŘÍK, B., 1996: Statistika III. Skriptum MZLU Brno, 156 s.

MONTGROMERY, D.C. 1991: Design and Analysis of Experiments. John Wiley and Sons, New York.

MORRISON, D.F. 1984: Multivariate Statistical Methods. McGraw-Hill Co.

MYSLIVEC, V., 1957: Statistické metody zemědělského a lesnického výzkumnictví. Praha, SZN

REISENAUER, R., 1970: Metody matematické statistiky a jejich aplikace v technice. Praha, SNTL, 240 s.

SACHS, L., 1972: Statistische Auswertugsmethoden. Berlin, Heidelberg, New York, Springer - Verlag, 506 s.

SIOTANI, M., HAYAKAWA, T., FUJIKOSHI, Y. 1985: Modern Multivariate Statistical Analysis. A Graduate Course and Handbook. American Science Press, Columbia.

ŠMELKO, Š. 1991: Štatistické metódy v lesníctve. Skriptum VŠLD Zvolen, 276 s.

ŠMELKO, Š., WOLF, J., 1977: Štatistické metódy v lesníctve. Bratislava, Príroda, 330 s.

ŠTULAJTER, F.,1989: Odhady v náhodných procesoch. Bratislava, ALFA, 288 s.

TUKEY, J. W., 1977: Exploratory Data Analysis. Adison-Wesley, 670 s.

ÜBERLA, K., 1974: Faktorová analýza. Bratislava, ALFA.

ZACH, J., 1990 A: Statistické metody - cvičení. Skriptum VŠZ Brno, 74 s.

ZACH, J., 1990 B: Statistické metody - vybrané části. Skriptum VŠZ Brno, 74 s.

ZACH, J., 1993: Statistické metody. Skriptum VŠZ Brno, 165 s.

ZACH, J., DRÁPELA, K., SIMON, J., 1994: Dendrometrie (cvičení). Skriptum VŠZ Brno, 167 s.

ZAR, J.H., 1984: Biostatistical Analysis, Prentice-Hall Int., New Jersey, 718 s.

Obsah II. dílu