

1

Úvod

1.1. Náhodný výběr, historie matematické statistiky

Hlavním úkolem matematické statistiky je zpracovat a vyhodnotit data z náhodného výběru. Tato data samozřejmě vykazují náhodnou variabilitu a to i v případě dat pořízených v připraveném pokusu se stálou kontrolou experimentálních podmínek. Obecně lze považovat výsledky měření – experimentální data – za realizaci náhodné veličiny.

Např. životnost akumulátoru anebo doba do další poruchy prvního bloku jaderné elektrárny Temelín je náhodná veličina. Podobně čas hoření svíčky anebo doba života úsporné žárovky je náhodná veličina. Uvádí-li výrobce životnost 50 hodin resp. 10 000 hodin, nelze samozřejmě očekávat, že svíčka resp. žárovka bude svítit přesně po tuto dobu. Udávané hodnoty ale byly nejspíše změřeny na omezeném počtu vzorků a jsou tedy ve své podstatě pozorovanými hodnotami náhodné veličiny. Zkoušky životnosti se samozřejmě neprovádí proto, aby se získaly údaje pro několik vybraných výrobků, ale aby se získaly informace o celé sérii výrobků. Tyto zkoušky mohou být také prováděny pro posouzení alternativních technologií. A v těchto situacích vzniká problém matematicko-statistického rozboru dat.

Úlohy z matematické statistiky mají z logického hlediska ve srovnání s úlohami z počtu pravděpodobnosti zcela opačný charakter. Pro úlohy z počtu pravděpodobnosti je typické při znalosti modelu chování náhodné veličiny (zákona rozdělení) usuzovat na pravděpodobnost určitého projevu chování v konkrétní situaci (výsledku náhodného pokusu) a jde tedy v zásadě o deduktivní myšlenkový pochod. Naproti tomu v matematické statistice se usuzuje z konkrétních výsledků (náhodného výběru) na obecný model chování náhodné veličiny (charakteristiky rozdělení); jde o induktivní myšlenkový pochod a mluvíme o statistické indukci.

Základní typy statistické indukce — bodový odhad, interval spolehlivosti, testování statistických hypotéz — budou předmětem našeho studia.

Cílem statistických výpočtů je využití počtu pravděpodobnosti k ohodnocení přesnosti a spolehlivosti získaných výsledků, např. ke stanovení hranic, které chyba výsledku s vysokou pravděpodobností nepřekročí, k výpočtu rizika, že chyba bude větší než určitá přípustná mez, k výpočtu rizika, že rozhodnutí učiněné na základě výsledků experimentu bude chybné, atd. K úlohám matematické statistiky dále patří i stanovení počtu pozorování potřebného k tomu, aby zmíněné rizika chyb byla udržena na přijatelné úrovni.

To co dělá teorii matematické statistiky obtížnou jsou úvahy založené na induktivním myšlení. Pro mnoho studentů je tato inference přijatelná až po obrovském studijním úsilí a proto také často studia statistiky zanechávají. Cílem této práce je zejména přiblížit matematickou statistiku každému laskavému čtenáři, pomoci získat nadhled a usnadnit mu bližší přístup k teoretičtějším publikacím.

Nepostradatelným nástrojem při zpracování experimentálních dat je teorie pravděpodobnosti. Ke studiu matematické statistiky je nutná znalost teorie pravděpodobnosti, zejména distribuční funkce, jednotlivá rozdělení náhodné veličiny, atd.

S příkladem jednoduchého statistického uvažování se můžeme setkat v detektivním románu *Felidae* od autora Akifa Pirincci. Detektiv kocour Francis je svědkem vražd v okolí. Má k dispozici seznam elektronicky evidovaných příslušníků jeho druhu a zjišťuje přibližné číslo 800 zmizelých. Odhaduje, že 250 se odstěhovalo a 100 se odebralo na věčnost v důsledku

stářím podmíněné slabosti či nemoci. Počet 100 odhadne z průměrné délky života devět až patnáct let. Krádeže čistokrevných zvířat nebo oběti dopravních nehod odhadne velkoryse na deset procent. Dochází k závěru, že u 350 z 800 zmizelých nebylo zavražděno. A počítá na str. 181 dál.

Pokud kat provozoval své řemeslo s neměnnou pravidelností, znamená to, že ročně poslal do věčných lovišť 64,28, měsíčně 5,35 a týdně 1,33 felidae. Z pohledu statistiky každých pět dní jednoho z nás poslal předstoupit před Stvořitele. Tyto úvahy se však rozcházejí s realitou posledních dvou až tří týdnů. I kdybychom vzali v úvahu určité nepřesnosti, zdálo se, že vrah momentálně řadí s téměř dvojnásobnou razancí a v intervalu dvou až tří dnů.

Zde by se samozřejmě nabízelo modelovat počet nevysvětlených zmizení Poissonovým rozdělením. Lehce můžeme spočítat pravděpodobnost deseti zmizelých za poslední měsíc $P(X = 10) = \frac{5,35^{10}}{10!} e^{-5,35} = 0,0251$.

1.2. Historie matematické statistiky

Důvodů pro relativně pozdní vznik statistiky je ještě více než u počtu pravděpodobností.

„Dobrý křesťan by si měl dát pozor na matematiky a všechny ty, kteří marně věští. Vždy existuje nebezpečí, že matematici uzavřeli smlouvu s ďáblem, aby očernili ducha a spoutali člověka do okovů pekla.“ Sv. Augustýn.

Vznik matematické statistiky je úzce spjat s nahromaděním množství dat v oblasti astronomie a demografických výzkumů v 18. století.

K autorů, kteří stáli u zrodu matematické statistiky patří

- Arbuthnott, který v roce 1712 zkoumá jaká je pravděpodobnost, že se v Londýně během 82 po sobě následujících roků narodí více chlapců než dívek.
- James Bernoulli, kterému posmrtně vychází v roce 1713 spis *Ars Conjectandi*. V práci chybí poslední kapitola, ale z textu předchozích se lze domnívat, že se v ní chtěl zabývat statistikou.
- Daniel Bernoulli, který v roce 1735 zkoumá dráhu 24 komet.
- Tobias Mayer, který v roce 1750 vytváří metodu průměrů pro řešení soustav rovnic a metodu lunárních vzdáleností pro určování zeměpisné polohy na základě polohy Měsíce.
- Johann Heinrich Lambert, který v roce 1772 vydává spis *Remarks about mortality, death lists, births and marriages*. Navrhuje také jeden z prvních algoritmů pro aproximaci dat přímkou.

V publikacích o matematické statistice jsou zmiňovány tzv. tři revoluce, kterým se budeme věnovat podrobněji v dalších kapitolách.

První revoluce je spojena se jménem Laplace, který v roce 1774 navrhuje metodu nejmenších absolutních odchylek.

Druhou revoluci přináší do statistiky metoda nejmenších čtverců pro aproximaci dat kterou použije Gauss v roce 1809. Takřka současně tuto metodu vytváří Legendre a Adrain.

Za třetí revoluci je považován Fisherův test, založený na χ -kvadrátu, z roku 1922.

Ke vzniku první a druhé revoluce přispívají data z oblasti měření Země a z astronomických měření. Třetí revoluce je již spjata se studii různých oblastí lidské činnosti. Např. William Sealy Gosset (autor Studentova rozdělení, 1935) zpracovává data v pivovaru Guinness s cílem vyrobit co nejlahodnější nápoj.

1.3. Cíle popisné statistiky a její historie

Informace obsažené ve velkém počtu dat se jeví lidskému pozorovateli jako nepřehledné. Proto se popisná statistika snaží tuto informaci zhustit do snadněji vnímatelné formy různých tabulek, grafů, číselných a funkcionálních charakteristik.

W.S. Jevons^{*)} komentuje své časové diagramy, v nichž sleduje změny cen základních i méně běžných produktů v závislosti na „komerčních bouřích“ typu objevení australského zlata v roce 1849 takto: „Jejich smyslem není ani odkaz ke konkrétním číslům, která lze lépe zjistit z odpovídajících tabulek, jako předvést očím obecné výsledky vyplývající z velkého množství čísel, jež nemohou být zachyceny jinak než graficky. Mé diagramy ukazují i ty nejmenší detaily tabulek, ale předčí i výpočty středních hodnot, protože oko či mysl samy zaznamenají obecný trend číselných souborů. Pouze tato reprezentace může být základem politicko-ekonomických debat a přesto většina statistických zdůvodnění závisí na pár číslech více či méně náhodně vybraných.“

Základní myšlenky popisné statistiky sice ke svému vyjádření používají jen elementárních matematických prostředků, ale jsou na jedné straně východiskem k poznání hromadných jevů reálného světa, na druhé straně motivací nejdůležitějších pojmů v počtu pravděpodobnosti a v matematické statistice.

1.3.1. Mapy a diagramy

První mapy a diagramy se statistickými údaji se objevují v 17. století, k rozmachu statistické grafiky dochází až koncem 18. století a je dílem francouzských stavebních inženýrů okolo Gasparda Monge. Grafika nachází uplatnění ve společenských studiích, v epidemiologii, v biologii a grafy se začínají objevovat i ve školních učebnicích. Samotné slovo *graf* je poměrně nové — objevilo se až v koncem 19. století — předtím se používalo převážně slov *mapa* a *diagram*.

Pro zajímavost se podíváme na první mapu, která vznikla 6200 př. Kr. Nachází se v muzeu ve městě Konya v Turecku a jde část fresky nalezené v Catal Hüyük.



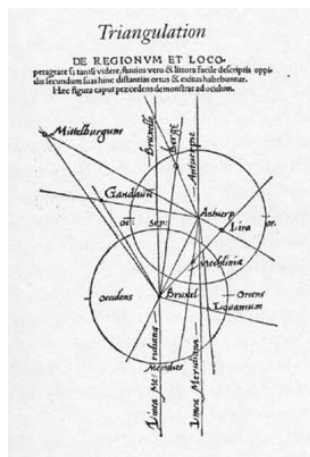
Zpracování velkého počtu dat si vyžádala Ptolemaiova mapa světa z roku 150, která vychází z Poseidoniova měření délky poledníku.

^{*)}R. D. Block (edit.) Papers and correspondence of William Stanley Jevons, vols. 1–7, Macmillan, London 1972–1981, vol. 2, 450. Dopis R. Huttonovi z 1. 9. 1862.



ଫିପ୍. ଏ ଫିଲ୍ ଓ ଫିଲ୍
 ଫିପ୍. ଏ ଫିଲ୍ ଓ ଫିଲ୍
 ଫିପ୍. ଏ ଫିଲ୍ ଓ ଫିଲ୍

V r. 1533 znáročuje belgičan Regnier Gemma-Frisius (1508-1555) způsob jak pomocí triangulace zjistit polohu pomocí měření úhlů.

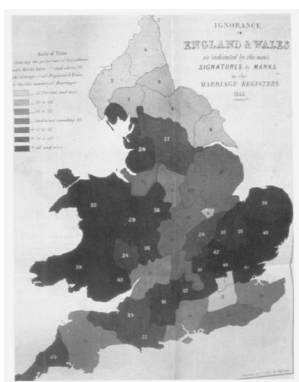


Dalším vývojovým momentem je zakreslování doplňujících charakteristik; E. Halley roku 1701 publikuje mapu se zakreslenými isogonálami spojujícími místa se stejnou magnetickou deklinací.

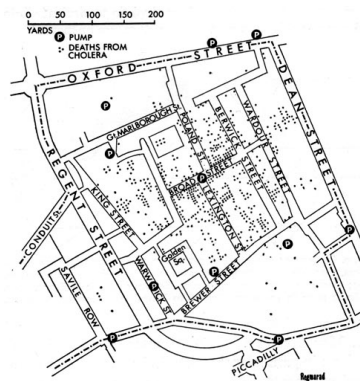


Tím začíná obor tématické kartografie, v níž jsou do map vedle územního členění zanášena data vztahující se k obyvatelstvu, obchodu, dopravě i k historickým událostem.

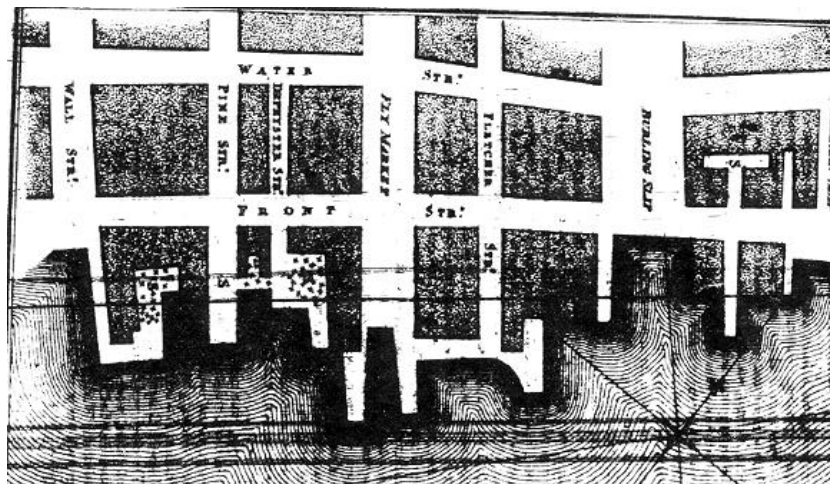
Na jejím počátku jsou mapy analfabetismu ve Francii (P. Ch. F. Dupin: *Carte de la France éclairée et de la France obscure*, 1819) a v Anglii (J. Fletcher: *Distribution of ignorance in England*, 1834) založené na průzkumu matrik (záznamy sňatků analfabetů mají značky místo podpisů).



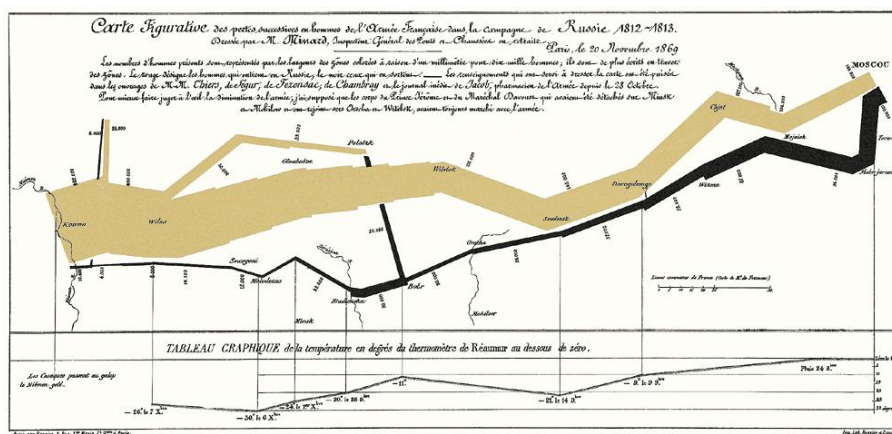
Zřejmě sem patří také slavný plán Londýna vytvořený Johnem Snowem v roce 1854 za účelem objasnění příčiny cholery. Zakreslením poloh studní a bydlíšť nemocných se podařilo lokalizovat nakaženou studnu a zjistit způsob šíření nákazy, který do té doby nebyl bezpečně znám.



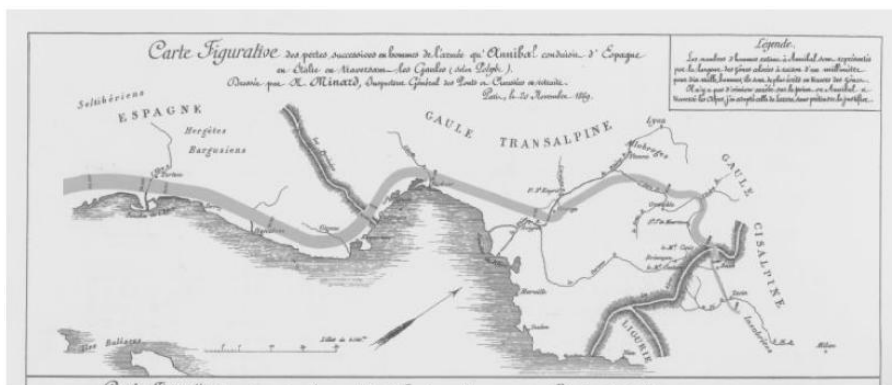
Prvenství v mapování šíření epidemie však patří Valentinu Seamanovi, který publikoval podobnou mapu jako J. Snow v souvislosti s epidemií žluté horečky v New Yorku v roce 1795 a několik map výskytu cholery bylo publikováno v Anglii již v první polovině 19. století.



Nepřekonaným vrcholem co do emocionální působnosti je „nejslavnější“ mapa všech dob, Napoleonovo tažení na Moskvu Charlese Josepha Minarda z roku 1869.



Obdobně Minard znázornil i Hanibalovo tažení do Itálie.



V současné době jsou nejběžnějším produktem tématické kartografie mapy zachycující okamžitý stav počasí, publikované v denním tisku a v televizi.

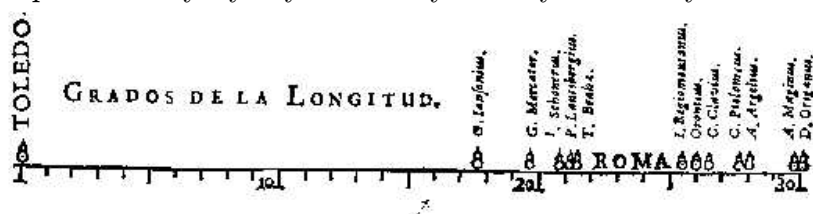
1.3.2. *Počátky statistické grafiky*

Statistická grafika prezentuje nejrůznější data v závislosti na zvoleném parametru, jímž bývá velmi často čas. Samotné slovo graf do angličtiny zavedl J. J. Sylvester v roce 1878 v souvislosti s konstatováním podobnosti mezi schématy molekulárních vazeb a grafickou reprezentací algebraických invariantů. Zhruba v téže době definuje graf Charles S. Peirce jako „plošný diagram sestávající z bodů či jejich ekvivalentů a jejich spojnice na omezené ploše“. Potřeba takové definice ukazuje, do jaké míry byly grafy ještě koncem XIX. století málo běžným informačním prostředkem.

Jejich počáteční rozvoj byl do značné míry ovlivněn, ne-li podmíněn, několika vynálezy umožňujícími grafické zaznamenávání kontinuálně probíhajících fyzikálních procesů. Prvním z nich je Christopherem Wrenem vynalezený zapisovač počasí (weather-clock) zaznamenávající teplotu a směr větru v polárních souřadnicích, dalším Wattův indikátor tlaku v parním stroji.

Za zakladatele statistické grafiky je obecně považován William Playfaire. Ve svých grafikách, které nazýval „čárovou aritmetikou“ (lineal arithmetics), využíval převážně kartézské souřadné soustavy, v níž znázorňoval závislosti jedné i více zvolených veličin na vybraném parametru, jímž byl nezřídka čas.

Měl však řadu předchůdců. Prvním z nich je Michael F. van Langren publikující v roce 1644 srovnání rozdílů zeměpisných délek Říma a Toleda. Údaj známý v jeho současnosti je srovnán s vesměs pozitivně vychýlenými odhady získanými z různých historických map.



Údaj známý v jeho současnosti (v obrázku označen šipkou) je srovnán s vesměs pozitivně vychýlenými odhady získanými z různých historických map. A právě tímto grafem se zabývá v románu *Ostrov věčejšího dne* Umberto Eco.

A tento rukopis obsahuje zprávu nějakého Vlása, který se opravdu vyzná, protože upozorňuje španělského krále, že nikdy nedošlo ke shodě při stanovení vzdálenosti z Říma do Toleda per los errores tan enormes, come se conoce per esta línea, que muestra la diferencia de las distancias, a tak dále. Umístíme-li první poledník do Toleda (Španělé se považují za pupek světa), pak podle Kremera by Řím byl o dvaadvacet, podle Regiomontana neboli Mullera málem o pětadvacet, podle Clavia o sedmadvacet a podle vynikajícího Ptolemaia o dvacet osm a podle Origana o třicet stupňů dále na východ. A k takovéto hromadě chyb došlo pouze při měření vzdáleností mezi Římem a Toledem.

Odhadneme-li střední hodnotu aritmetickým průměrem, dostáváme

$$\bar{X} = \frac{22 + 25 + 27 + 28 + 30}{5} = 26,4^\circ.$$

Disperzi odhadneme veličinou

$$s^2 = \frac{1}{5} \sum_{i=1}^{5-1} (X_i - \bar{X})^2 = \frac{1}{4} 37,2 = (3,05^\circ)^2.$$

Interval spolehlivosti pro střední hodnotu je

$$(\bar{X} - \frac{St_{n-1,1-\alpha/2}}{\sqrt{n}}, \bar{X} + \frac{St_{n-1,1-\alpha/2}}{\sqrt{n}}),$$

pro $\alpha = 0,05$ dostáváme

$$(26,4 - \frac{3,51 \cdot 2,57}{\sqrt{5}}, 26,4 + \frac{3,51 \cdot 2,57}{\sqrt{5}}) = (22,9; 29,9).$$

Skutečná hodnota ale je pouze 16° , Eco se tak může našim odhadům vysmívat, stejně jako se vysmívá kartografům, když píše: „Měl-li jeden z nich pravdu, ostatní se zmílili o vzdálenost mezi Londýnem a Zemí královny ze Sáby.“

To je samozřejmě nadsázka. Položme si ale otázku, zda tuto nadsázku (chyba měření je větší nebo rovna rozdílu zeměpisných délek mezi Londýnem a Saúdskou Arábií, t.j. 30°) jsme schopni odhalit pomocí nerovností používaných v matematické statistice.

1. Čebyševova nerovnost

$$P(|X - \mu| > \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

nám, použijeme-li odhad disperze S^2 , určí

$$P(|X - \mu| \geq 30) \leq \frac{3,05}{30^2} = 0,0034,$$

což je pravděpodobnost malá ale nikoliv zanedbatelná.

2. Markovova nerovnost

$$P(X \geq \varepsilon) \leq \frac{\mu}{\varepsilon}$$

nám, použijeme-li odhad střední hodnoty \bar{X} , určí

$$P(X \geq 26,4 + 30) \leq \frac{26,4}{56,4} \doteq 0,468,$$

což je až neuvěřitelně velká pravděpodobnost.

S uvedenými nerovnostmi tedy nadsázku Umberta Eca odhalit můžeme, i když některé získané odhady pravděpodobností mohou vzbudit posměch. Je důležité si uvědomit, že uvedené nerovnosti jsou velmi hrubé.

Samozřejmě můžeme (za předpokladu normality měření kartografů) určit pravděpodobnost mnohem přesněji. Použijeme-li určené výběrové momenty, dostáváme s využitím statistických tabulek

$$P(X \geq 30) = 1 - F_X(30) = 1 - \int_{x \rightarrow -\infty}^{30} \frac{1}{\sqrt{2\pi}s} e^{-\frac{(x-\bar{X})^2}{s^2}} = 1 - 0,99865 = 0,00135,$$

což činí náš jev velmi nepravděpodobným.

Důležitým cílem v matematické statistice je získat nevychýlený odhad. Bohužel zkonstruovaný odhad vychází z náhodného výběru, který je zatížen systematickou chybou. Všichni jmenovaní kartografové při vytváření svých map vycházeli z Ptolemaiovy mapy. Přitom je důležité vědět, že Ptolemaios při konstrukci mapy uvažoval o třetinu menší obvod Země než je ve skutečnost. Tento omyl je skutečnou příčinou vychýlení citovaných měření vzdáleností mezi Toledem a Římem. Měření kartografů sice byla nezávislá ale vychýlená (biasovaná).

Připomeňme si nyní definici nestranného odhadu, biasu a nejlepšího nestranného odhadu.

Výběrová funkce $T = T(X_1, \dots, X_n)$ se nazývá:
 – **nestranný (nevychýlený) odhad** parametrické funkce $\tau(\theta)$, jestliže platí

$$E_\theta[T(X)] = \tau(\theta), \forall \theta$$

Rozdíl

$$b(\theta) = E_\theta[T(X)] - \tau(\theta)$$

se nazývá bias (vychýlení) odhadu $T(X)$

– **stejněměrně nejlepší nestranný** parametrické funkce $\tau(\theta)$, je-li $T(X)$ nestranný odhad $\tau(X)$ a platí

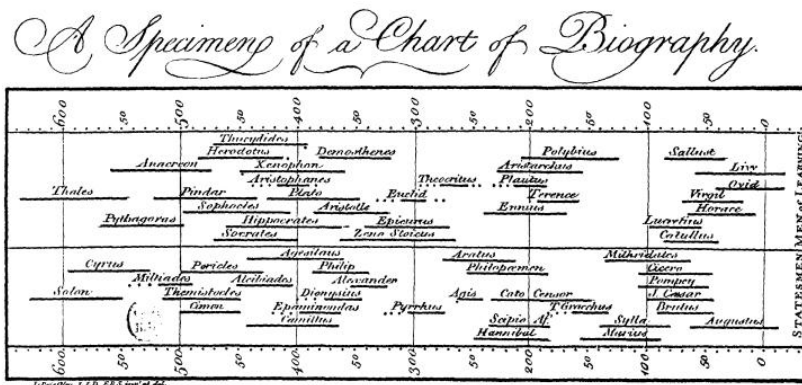
$$\text{var}_\theta[T(X)] \leq \text{var}_\theta[T^*(X)], \forall \theta$$

kde $T^*(X)$ je libovolný jiný nestranný odhad $\tau(\theta)$, tj. $T(X)$ má mezi všemi nestrannými odhady $\tau(\theta)$ nejmenší rozptyl.

V našem případě tedy bias je $16^\circ - 26,4^\circ = -10,4^\circ$.

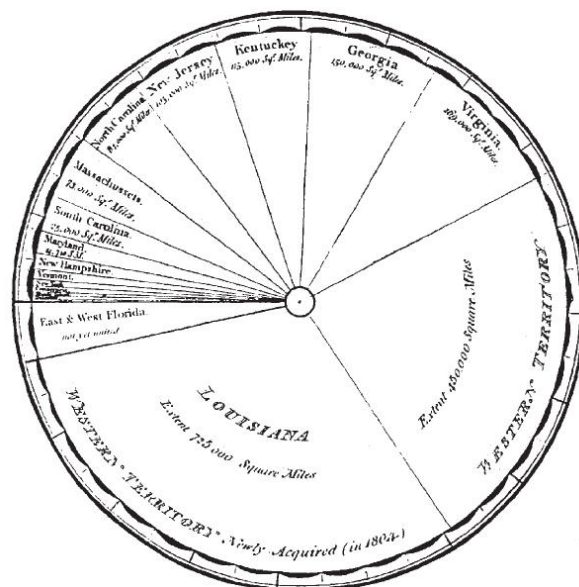
Při čtení této publikace se předpokládá znalost pravděpodobnosti a matematické statistiky v rozsahu bakalářského studia. Pokud čtenář pojmům na předcházejících dvou stranám nerozuměl, je třeba, aby si základní kurz dostudoval.

Dalším známým ranným grafem je Životopisná mapa (Chart of Biography, 1765), jejíž autorem je neobyčejně plodný vynálezce, vědec, teolog a politik Joseph Priestley. Opět se jedná v podstatě o jednorozměrný diagram prezentující životní rozpětí 2000 významných osobností žijících v letech 1200 př. Kr. až 1750.



Priestley na čtyřech stránkách komentáře přesvědčuje čtenáře, že toto znázornění času je možné a účelné. Zatímco dnes je tento přístup považován za zcela přirozený, v polovině XVIII. století tomu bylo jinak. Kartézské souřadnice byly obecně přijaty jako systém vhodný pro znázornění prostoru, v němž existuje pozorovatelný materiální svět (nezavedl je však Descartes, ale Leonardo da Vinci kolem roku 1500 pro analýzu rychlosti padání objektů). Historický čas však byl považován za jev subjektivní, vázaný na schopnost myšlení a sám Descartes zdůrazňoval „nezbytnost úplného abstrahování od analogií s hmotou při studiu zákonitostí Mysli“.

Dalším příkladem je kruhový diagram rozlohy amerických států zkonstruovaný Williamem Playfairem.

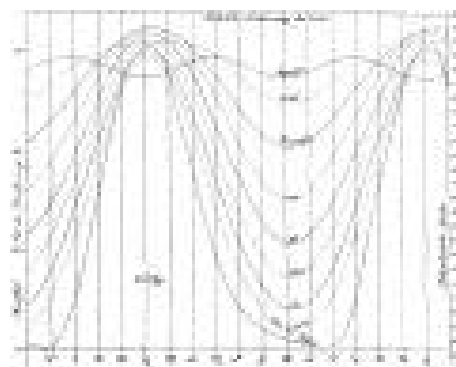
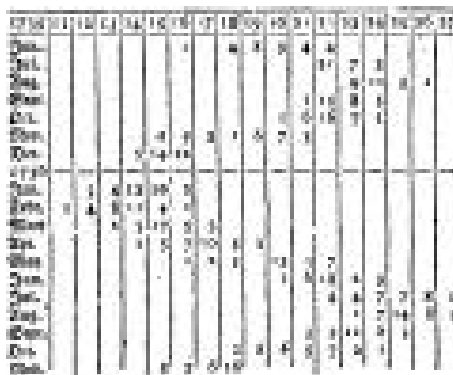


STATISTICAL REPRESENTATION of the UNITED STATES of AMERICA.

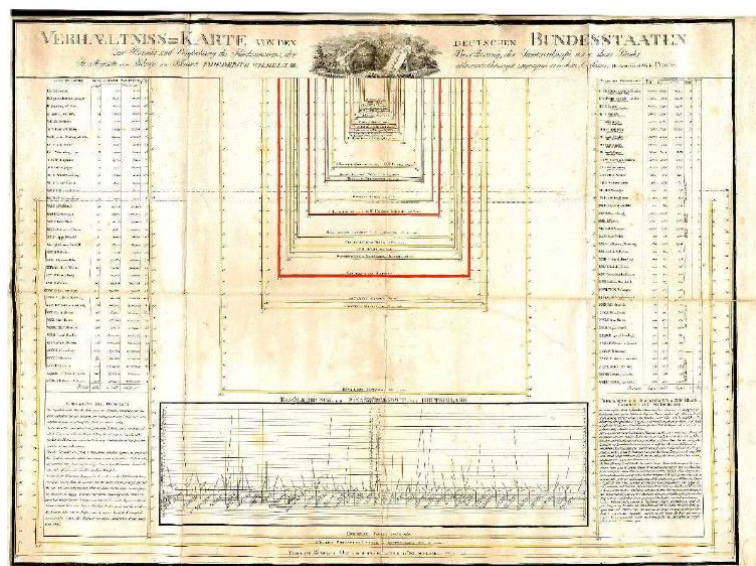
by W. PLAYFAIR.

The Newly invented Method is intended to show the Proportions between the different parts of the Union.
Total Extent 1,228,000 Square Miles or 3,224 Millions of Acres.

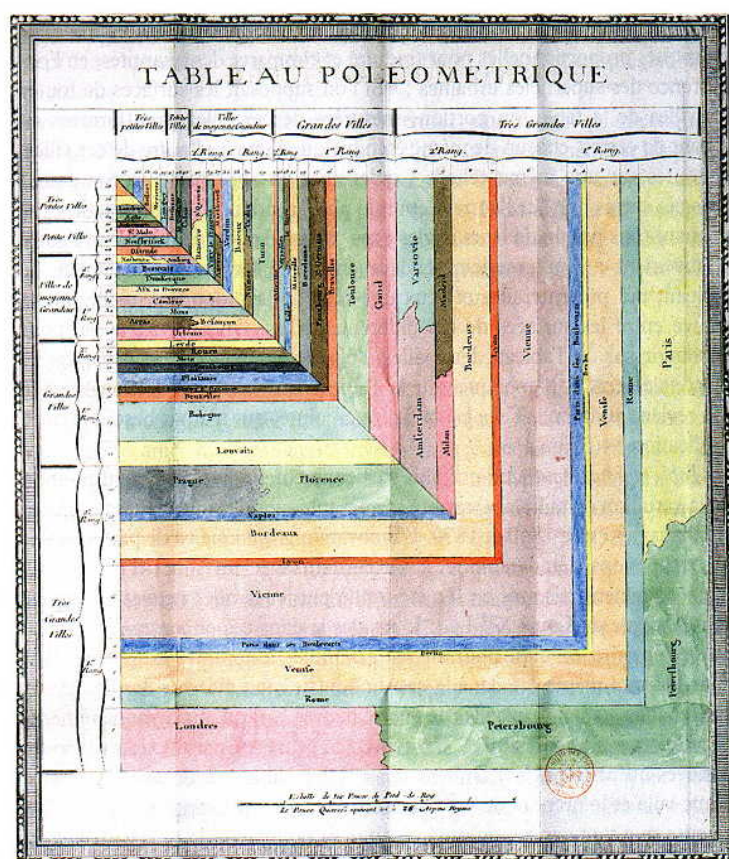
Jedním z prvních uživatelů grafického zobrazení dat byl také alsaský přírodovědec Johann Heinrich Lambert, jehož hlavním zájmem byla fotometrie a fyzikální či astronomická měření. Byl patrně první, kdo vytvořil „číselný graf“ vhodným rozmístěním číselných hodnot v rovině.



S dalším propagátorem grafických metod se vrací problematika sociálních a politických věd. August Friedrich Wilhelm Crome byl profesorem politických věd v Gießen a je známý jednak svými knihami (např. *Über die Größe und Bevölkerung der europäischen Staaten* z roku 1785), jednak řadou pamfletů, v nichž vedl vášnivé politické diskuse a své názory často dokazoval graficky zpracovanými statistickými údaji. Pomocí diagramů různých typů porovnával situaci v jednotlivých státech, např. velikost států znázorňuje pomocí pravidelných obrazců (čtverců, obdélníků či kruhů) o plochách úměrných rozlohám států, takže optický dojem není zkreslen komplikovaným průběhem hranic.



Autorem prvního takového grafu byl však Charles de Fourcroy; v práci l'Essay d'une table poléographique z roku 1782 srovnává rozlohy evropských měst čtvercovým diagramem. Slavná je také jeho mapa Produkten-Karte von Europa z roku 1782, znázorňující vedle měst a přístavů také přírodní a průmyslovou produkci v jednotlivých zemích.

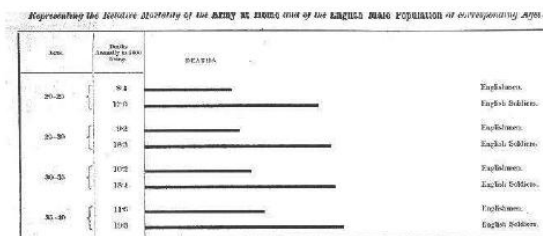
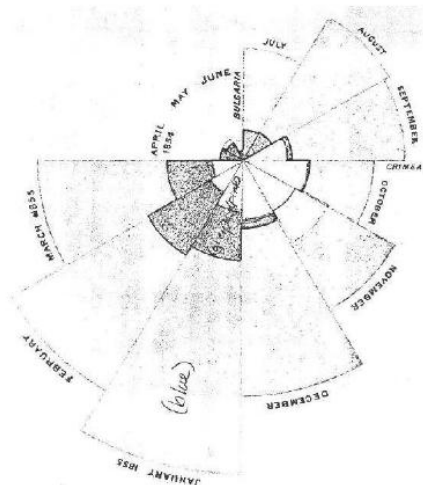


1.3.3. Statistická grafika v 19. století

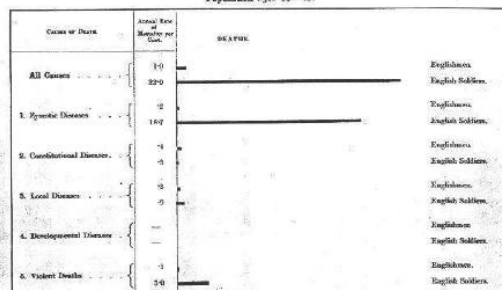
Poté, co se grafické zobrazování začalo v širší míře používat, vyskytla se potřeba technických prostředků, které by usnadňovaly jeho realizaci a šíření. V Anglii v roce 1794 začíná Dr. Buxton vyrábět rastrovaný papír, v Německu v roce 1798 pražský rodák Aloys Senefelder vynalézá litografickou techniku pro tisk map a diagramů (své výsledky shrnuje v knize *Vollständiges Lehrbuch der Steindruckerei*, 1818). Ve Francii v roce 1843 Léon Lalanne začíná používat sférické souřadnice a v roce 1846 zavádí logaritmickou stupnici na obě pravoúhlé osy. Semilogaritmickou stupnici používá jako první pro své diagramy W. S. Jevons v roce 1863.

Playfairovy grafy byly patrně inspirací pro anglickou statističku Florence Nightingaleovou. Přihlásila se jako dobrovolná zdravotní sestra v době krymské války, sestavovala časové tabulky úmrtí pacientů podle příčin a jimi dokazovala nedostatečnost nemocniční hygieny v polních podmínkách. V prvním provedení byly počty úmrtí úměrné úsekům poloměrů výsečí a tedy zkreslené, poté si uvědomila svou chybu a jako první zavedla radiální graf. Vedle podrobné zprávy pro vojenské kruhy vydala stručný souhrn svých výsledků také jako malou brožurku (*Mortality of the British Army*, 1858) s cílem ovlivnit veřejné mínění.

Radiální graf F. Nightingaleové (1858) znázorňuje příčiny úmrtí vojáků (počet úmrtí je úměrný ploše) v krymské válce (1854–55). Vnitřní malé světlé výseče zachycují po jednotlivých měsících úmrtí na zranění, velké světlé výseče úmrtí na nakažlivé choroby vyvolané nedostatečnou hygienou a vnitřní malé tmavé výseče libovolné jiné příčiny. Sloupcové diagramy porovnávají procentuální úmrtnost v různých věkových kategoriích (horní diagram) a podle příčin (spodní diagram) u běžných anglických mužů a u vojáků (vždy spodní sloupec v páru).



Representing the Relative Mortality, from different Causes, of the Army in the East in Hospital and of the English Male Population aged 15-45.



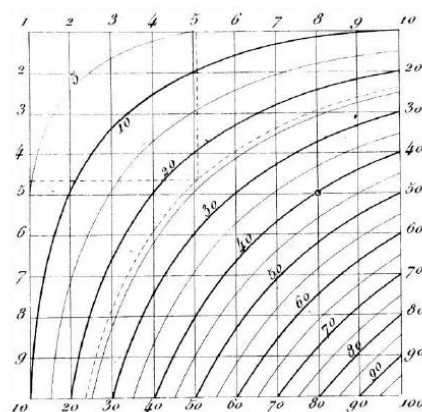
Ať již její grafické zpracování přesvědčilo velení armády či veřejnost, která uplatnila svůj vliv, hygieně v nemocnicích začala být věnována podstatně větší pozornost, a to nejen v armádě. Po návratu do Anglie měla F. Nightingaleová značný (údajně dodnes přetrvávající) podíl na celkovém zlepšení nemocniční péče, již věnovala veškerou svou pozornost po zbytek života. Její radiální grafy bývají v literatuře nazývány kohoutími hřebínky (coxcombs), jedná se však o jeden z historických omylů; kohoutím hřebínkem nazvala F. Nightingalová

v průvodním dopise z 25. 12. 1857 k výše zmíněné brožurce prezidentovi Královské armádní komise Sidney Herbertovi právě tuto brožurku, nikoliv svůj radiální graf.

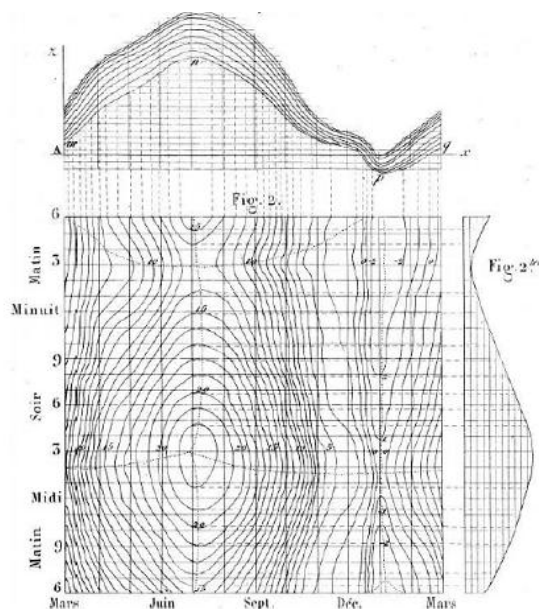
Od začátku 19. století se střediskem vývoje grafického znázorňování dat stává Francie. Jejich technické využití a rozvoj jsou svázány s odvozem městských odpadků, který byl aktuální již v 18. století a vynucoval si stále rozsáhlejší stavbu silnic. Maximální efektivností této problematiky se zabývaly dva přední francouzské vzdělávací ústavy: vojenská École de Génie v Mezières (s těžištěm v likvidaci pevnostního odpadu) a École des Ponts et Chaussées v Paříži zaměřená civilně. Profesorem na první škole byl Gaspard Monge, zakladatel deskriptivní geometrie, a právě z jeho žáků a následovníků se rekrutovali významní propagátoři grafického zobrazování. Na druhé z uvedených škol zase vyučoval již zmíněný Ch. J. Minard. Záměr pokrýt celou Francii vyhovující sítí silnic hvězdovitě vycházejících z Paříže se stává aktuální kolem roku 1842. Při jeho realizaci opět přichází ke slovu grafická kartografie, zvláště díky Ch. J. Minardovi, který se snažil prosadit decentralizovanější dopravní síť, jejíž výhodnost demonstroval čarami s tloušťkou úměrnou přepravním nárokům; tato forma grafického znázornění vyvrcholila posléze jeho Napoleonovým tažením. Výstavba dopravní sítě však byla svěřena centrální státní organizaci Corps des Ponts et Chaussées řízené Victorem Legrandem; její charakter vyjadřoval hovorový název „Legrandova hvězda“ a byla spojena s obrovskými přesuny půdy díky přísným požadavkům na povolené maximální stoupání a minimální poloměry křivosti. Již v letech 1835 a 1837 byly vypracovány tabulky pro výpočet nezbytných přesunů zeminy, platily však pouze pro jeden pevný profil silničního uložení.

Grafické konverze výpočetních tabulek se ujal Léon Lalanne. Vyšel při tom z tzv. pytagorejské tabulky typu 10×10 , kterou Louis-Ézechiel Pouchet (v souvislosti se snahami francouzské vlády přejít na decimální soustavu jednotek) v roce 1795 doplnil isočarami (hyperbolami) $xy = 5k$, $k = 1, 2, \dots, 19$

1	2	3	4	5	6	7	8	9	10
2	4	6	8	10	12	14	16	18	20
3	6	9	12	15	18	21	24	27	30
4	8	12	16	20	24	28	32	36	40
5	10	15	20	25	30	35	40	45	50
6	12	18	24	30	36	42	48	54	60
7	14	21	28	35	42	49	56	63	70
8	16	24	32	40	48	56	64	72	80
9	18	27	36	45	54	63	72	81	90
10	20	30	40	50	60	70	80	90	100

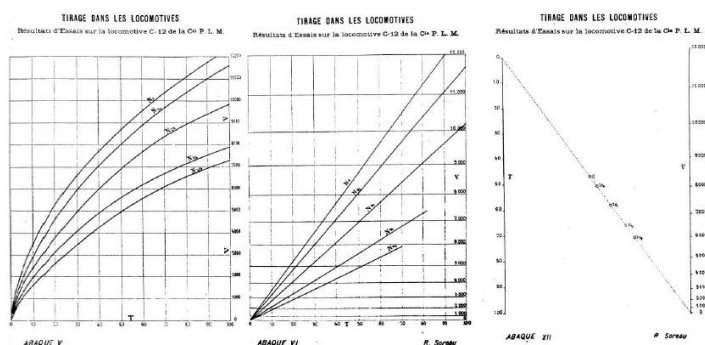


Tabulka se sice obecně neprosadila, byla však používána pro inženýrské výpočty k převodu různých měr, např. při kalibraci děl. Lalanne nejdříve upozornil, že čáry $xy = \text{konst.}$ můžeme chápat jako ortogonální projekce čar konstantní výšky na 3D ploše $z = xy$ a pro demonstraci této myšlenky vytvořil projekci isotherm v 3D grafu typu (měsíc \times hodina \times teplota) s projekcí do roviny (měsíc \times teplota) a řezem rovinou (hodina \times teplota) — Mongeova škola se nedala nezapřít.



Druhou inovací bylo zavedení logaritmických souřadnic (Pouchetovy hyperboly se pak staly přímkami) a v roce 1846 již Lalanne publikuje grafickou tabulku s lineárními závislostmi půdních přenosů pro dvoukolejnou železnici.

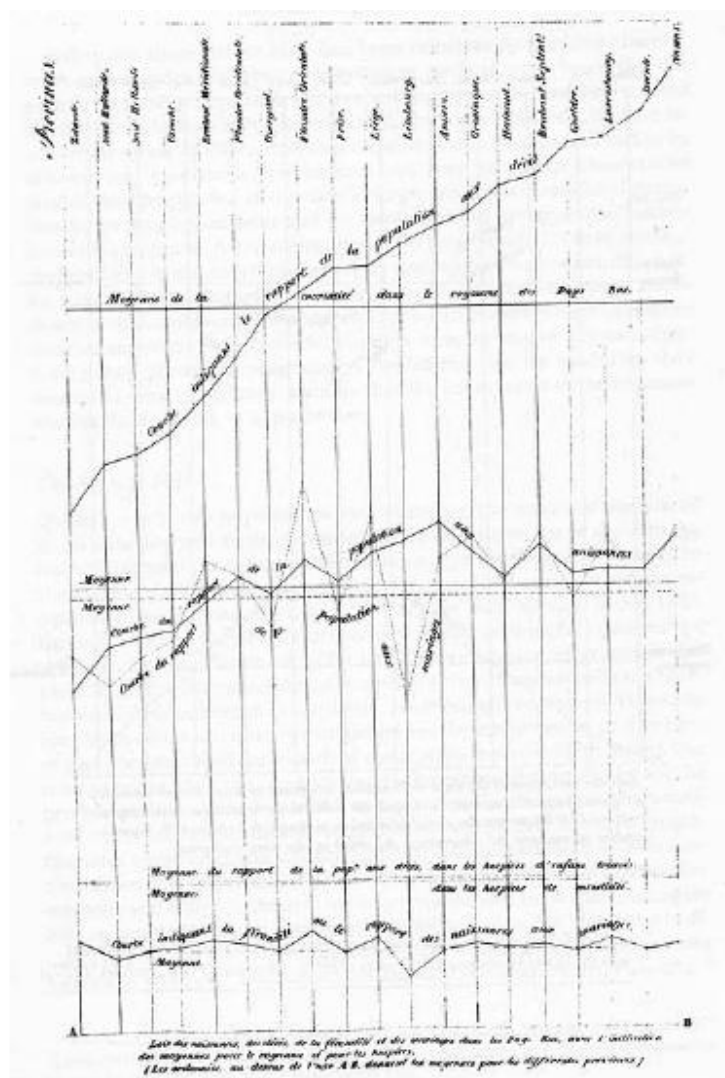
Vývoj dovršuje v roce 1884 Maurice d'Ocagne vytvořením nomogramu. Pravoúhlé osy nahrazuje osami rovnoběžnými a využívá principu duality z projektivní geometrie, podle něž lze body zobrazit jako přímky a přímky jako body. Soubor přímek z Lalanneova grafu pak přechází v přímku jedinou.



Zásluhy L. A. J. Queteleta o rozvoj statistiky v sociální oblasti jsou dostatečně známe: jeho názory jsou různě vykládány, interpretovány i kritizovány, jeho podíl na vzniku statistických společností v evropských státech i v Americe je však nesporný, stejně jako inspirativní vliv na celou řadu statistických aktivit. Z Queteletových grafických prací si všimneme aspoň jednoho okruhu studií včetně okolností, za nichž vznikly.

Sčítání lidu je velmi nákladná akce, a když se v porevoluční Francii o ní začalo uvažovat, přišel P. S. Laplace s návrhem určité formy výběrového šetření. Doporučil využít přesně vedených matrik narozených dětí v celé zemi a celkový počet obyvatel N_O určit ze vztahu $N_O = r_D N_D$, kde N_D je počet všech narozených dětí za nějaké období a $r_D = n_O/n_D$ je pečlivě stanovený poměr počtu obyvatel a narozených dětí ve vybraných „reprezentativních“ oblastech, rovnoměrně rozložených po celé ploše státu a s pozorností k jednotlivým skupinám obyvatel“.

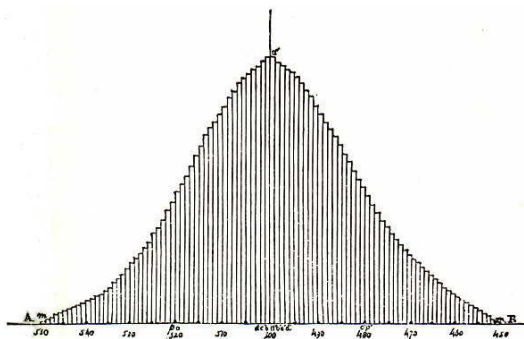
Quetelet byl nejprve (v roce 1824) nakloněn použití této metody i v Belgii a Nizozemí, avšak v roce 1829 podává návrh na kompletní sčítání. Byl totiž zřejmě ovlivněn pamětním spisem, který mu poslal baron de Keverberg v roce 1827 a v němž zpochybňuje možnost dostatečně vhodného výběru podoblastí pro odhad poměru r_D , protože relace mezi n_O a n_D závisí nesnadno definovatelným způsobem na množství lokálních proměnných. Patrně inspirován de Keverbergovým spisem, provedl Quetelet v 19 oblastech Belgie, Holandska a Lucemburku vlastní výběrové odhady následujících veličin: počtu obyvatel n_O , počtu narozených dětí n_D , počtu uzavřených manželství n_S a počtu úmrtí n_M , z nichž pro každou oblast odhadl poměry $r_M = n_O/n_M$, $r_S = n_O/n_S$, $r_F = n_D/n_S$ a $r_D = n_O/n_D$ a oblasti srovnal za sebou tak, aby r_M bylo monotónní rostoucí. Výsledky jsou shrnuty ve známém Queteletově diagramu, který ukazuje poměrně velké rozdíly mezi hodnotami poměrů v jednotlivých oblastech a dále naznačuje, že mezi nimi je jen stěží nějaká korelace.



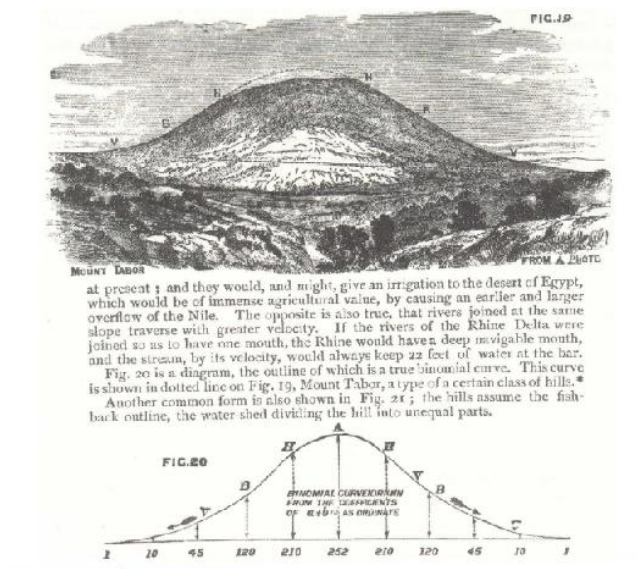
Odtud tedy vyplynula Queteletova ztráta důvěry v Laplaceův návrh výběrového sčítání.

Další Queteletova grafická práce se vztahuje k jeho koncepci „průměrného člověka“, jehož psychické i fyzické vlastnosti mají normální rozdělení (Quetelet však používal termíny křivka možností, rozdělení možností, binomická křivka). Přesvědčení, že každý homogenní

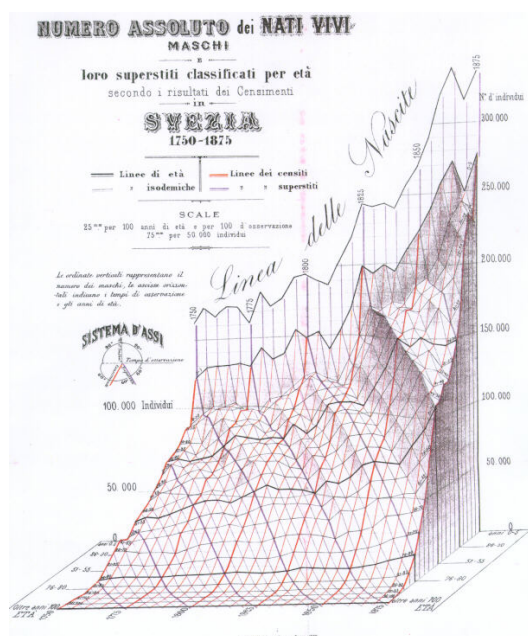
soubor údajů musí mít normální rozdělení, považoval Quetelet za řešení de Kevenbergovy námitky o nemožnosti posoudit, zda data vytvářejí homogenní soubor či nikoliv. V řadě prací srovnával zjištěná data s normálním rozdělením, jež však nepoužíval v Gaussově integrálním tvaru, ale vycházel z binomického rozdělení $Bi(999, 1/2)$.



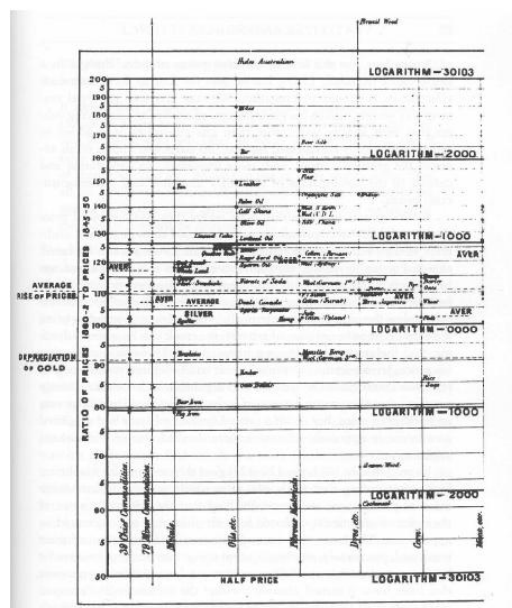
Obecnou popularitu normálního rozdělení dokumentuje článek A. Tylora z roku 1875, v němž autor povýšil křivku normálního rozdělení na universální geologický standard (binomická křivka nebo-li denudační křivka) tvaru hor. Odchyłky od ní jsou něj důkazem lokální eroze demonstrováné na příkladu biblické hory Tábor.



Luigi Perozzo vstoupil do historie grafického zobrazování prvním 3D grafem, který nazval stereogramem a jenž využívá axonometrického promítání navrženého Gustavem Zeunerem v knize *Abhandlungen aus der mathematischen Statistik*, Leipzig (1969). 3D grafy byly často využívány pro znázornění vícerozměrných distribučních funkcí a hustot pravděpodobnosti.



W. S. Jevons se v roce 1863 začal zabývat problémem kvantitativního popisu cenových změn vyvolaných událostmi obecného dosahu, konkrétně např. objevením australského a kalifornského zlata v roce 1849, jež mělo za následek dlouhodobý pokles ceny zlata. Ze sledovaných 118 produktů jich 84 zdražilo, ostatní zlevnily. Všechny změny Jevons zanesl do souborného semilogaritmického grafu a stanovil jejich geometrické průměry.

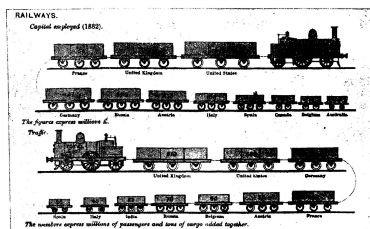


Jevonsův semilogaritmický diagram (1863) cenových změn po objevu australského a kalifornského zlata. Vyneseny jsou poměry průměrných cen v objevu zlata ovlivněných letech 1860 až 1862 k průměrným cenám ve srovnávacím období 1845 až 1850. Na levých dvou svislých přímkách jsou vyneseny všechny hlavní (39 položek) a vedlejší (89 položek) produkty a vyznačeny jim odpovídající průměrné změny, dále průměrné relativní zvýšení (cca 11 %) a jemu odpovídající relativní pokles ceny zlata (cca 9 %). V jednotlivých sloupcích

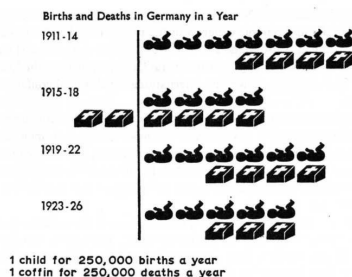
jsou většinou příbuzné produkty, jako železo, oceli a stříbro, různé druhy olejů, textilní látky, obilí atd.

Tento typ výpočtu cenových změn je od té doby široce používán. Nebyl zdaleka první, avšak prosadil se ze dvou důvodů. Předně pro výrazně asymetrické rozdělení relativních cenových změn je geometrický průměr vhodnější, než do té doby používaný průměr aritmetický, jednak se ukázalo, že je vhodné sledovat velmi široký výběr produktů, což Jevonsovi předchůdci nedělali. Jevons pro svůj postup měl ovšem jen intuitivní důvody; zmiňoval např. alternativní možnost sledovat množství zboží, které lze po skokové změně zakoupit za stejnou cenu, cožby vedlo k průměru harmonickému, a svůj geometrický průměr vydával za střední cestu mezi oběma alternativami.

První piktogram znázorňuje Michael George Mulhall (1836-1900).



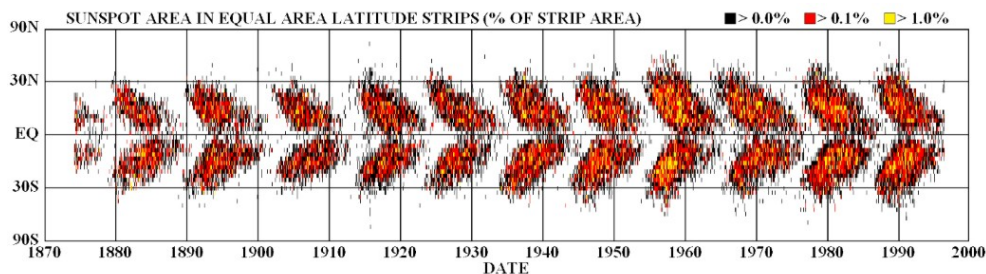
Na dalším obrázku vidíme piktogram z 20. století.



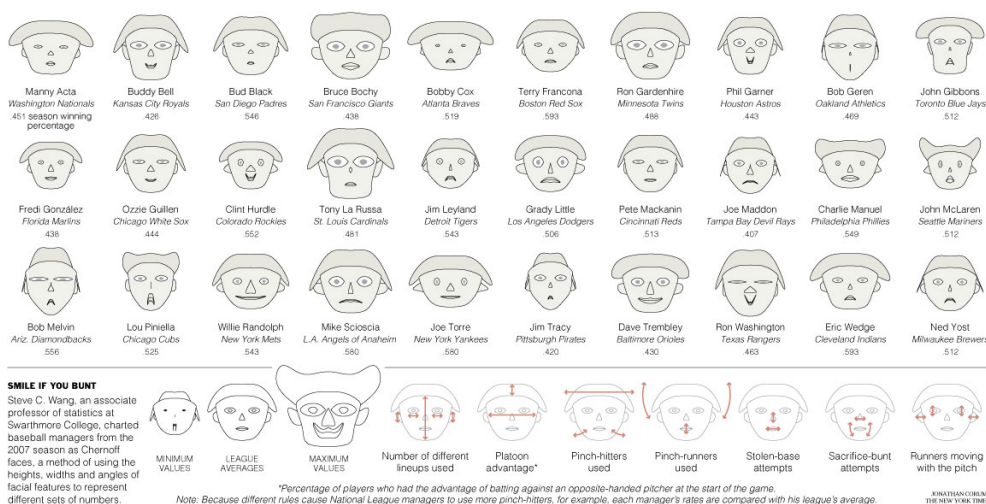
Koncem 19. století se začíná rozvíjet intenzivní zkoumání v oblasti lékařství a biologie v Anglii, do značné míry spojené s osobou Francise Galtona. Jeho zájem o aplikaci statistických přístupů včetně jejich grafické prezentace byl zcela mimořádný a vedl ke vzniku tzv. biometrické školy, jejímiž představiteli vedle Galtona byli Karl Pearson, Francis Weldon, Udna Yule a samozřejmě Ronald Aylmer Fisher. Grafická prezentace se v jejich pracích stala běžným prostředkem do té míry, že si dnes bez ní statistiku dovedeme jen stěží představit. Zhruba do konce 19. století je možné vývoj grafického zobrazování alespoň v hrubých rysech sledovat v příspěvcích rozsahu srovnatelného s tímto textem.

20. století, zejména jeho druhá polovina ovlivněná rozvojem počítačové techniky a vstupem grafiky do všech medií, představuje pravý grafický výbuch. Jeho rozsah lze snad alespoň přibližně ocenit z odhadu E. Tufta, prezentovaného v osmdesátých letech minulého století: počet grafů vytvořených za rok se pohybuje mezi $9 \cdot 10^{11}$ až $2 \cdot 10^{12}$; při počtu lidí v řádu 10^{10} připadá tedy 100 grafů na osobu za rok. Toto množství zdaleka neznamená, že grafy jsou kvalitní a že data jsou prezentována optimálním způsobem. Právě naopak; převážná většina grafů má za úkol zaujmout, obrátit pozornost k určité obchodní nebo politické problematice, „nakazit“ konzumenta názorem či záměrem svých autorů. Současná teorie grafického zobrazování dává přednost těm nejjednodušším formám snadno dešifrovatelné bodové a čárové reprezentace.

Ze složitějších grafů se podíváme na motýlí diagram, který navrhl v r. 1904 Edward Walter Maunder (1851-1928) a který znázorňuje sluneční aktivitu.



Zobrazit vícerozměrná data se můžeme pokusit pomocí Chernofových obličejů. Na obrázku šířka, výška, tvar úst, obočí a další parametry obličeje reprezentují různé měřené veličiny.



2

Jednorozměrný statistický soubor

Ke zhuštění údajů o kvantitativním statistickém znaku (respektive o nespojitém znaku s velkým počtem variant) máme dvě cesty:

- přes intervalové rozdělení četností,
- přes konstrukci měř (polohy, variability, šikmosti a špičatosti), které mohou být založeny na tzv. kvantilech.

2.1. Intervalové rozdělení četností

Na příkladu budeme demonstrovat první přístup, který nám umožní zkonstruovat histogram.

Mějme k dispozici jednorozměrný statistický soubor s jedním znakem — údaje o příjmech 51 osob:

```
6400 7975 8434 9059 9808
9876 10015 10349 10560 10805
10825 10884 10923 11143 11429
12000 12420 12956 13075 13213
13243 13630 13804 13809 13899
14390 14406 14569 14890 14906
14928 15243 15300 15908 15963
16075 16343 16451 16983 17473
17825 17875 17962 18563 19849
20016 20102 20377 22012 22865
25416
```

Pro dělení n pozorování do k tříd se doporučuje orientačně využít některé z následujících pravidel:

- a) $k \leq 5 \log n$,
- b) $k \sim \sqrt{n}$,
- c) (Sturgesovo pravidlo) $k \sim 1 + 3,3 \log n$.

V našem případě pro $n = 51$ dostáváme

ad a) $5 \cdot \log 51 = 5 \cdot 1,70757 \doteq 8,5$,

ad b) $\sqrt{51} \doteq 7,14$,

ad c) $1 + 3,3 \log 51 \doteq 6,635$.

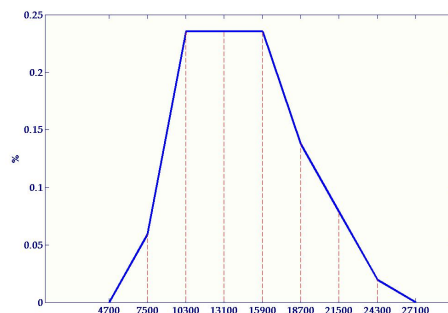
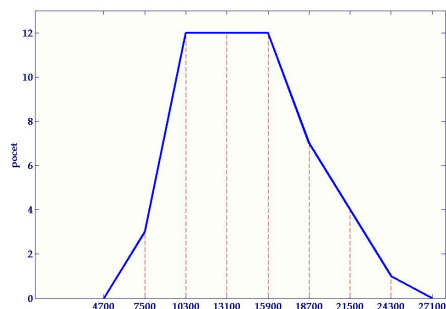
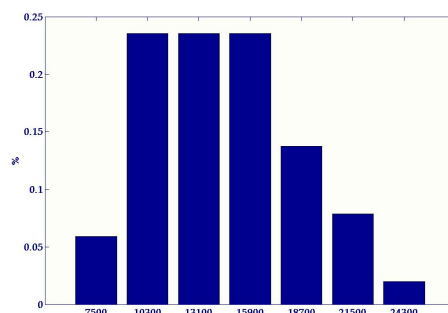
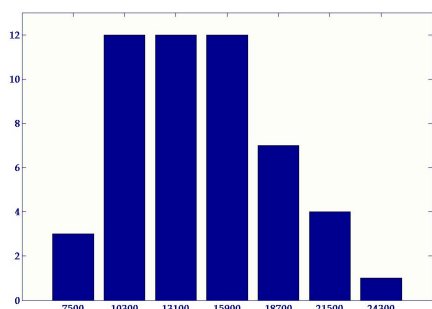
Variační obor $19016 = 25416 - 6400$ rozdělíme na 7 stejných intervalů o délce $19016/7 \doteq 2717 \doteq 2800$. Pozor: zaokrouhlujeme vždy nahoru (laskavý čtenář jistě domyslí proč)!

Náš statistický soubor rozdělíme do tříd a určíme jejich četnosti.

č.	interval	střed intervalu	četnost		kumul.četnost	
			abs.	rel.	abs.	rel.
1.	6100–8900	7500	3	0.059	3	0.059
2.	8900–11700	10300	12	0.235	15	0.294
3.	11700–14500	13100	12	0.235	27	0.529
4.	14500–17300	15900	12	0.235	39	0.764
5.	17300–20100	18700	7	0.137	46	0.901
6.	20100–22900	21500	4	0.079	50	0.980
7.	22900–25700	24300	1	0.020	51	1.000
součet	—	—	51	1,000	—	—

Na takto zhuštěném statistickém souboru budeme demonstrovat užití nástrojů statistické grafiky.

Nejprve vykreslíme polygon četností.



2.2. Kvantily

Kvantilem resp. fraktilem (lat. fractus = zlomený) nazýváme hodnotu kvantitativního statistického znaku (nejčastěji spojitého) určenou tak, že hodnoty, které jsou menší (stejně) než ona, tvoří určitou stanovenou část rozsahu souboru, např. 1 %, 5 %, 20 %, 25 %, 40 %, 50 %, 95 %.

Příslušné kvantily nějakého kvantitativního znaku můžeme značit $\tilde{x}_{0.5}$, \tilde{x}_1 , \tilde{x}_2 , \tilde{x}_{10} , \tilde{x}_{20} , \tilde{x}_{25} , \tilde{x}_{40} , \tilde{x}_{50} , \tilde{x}_{95} apod.

- přes intervalové rozdělení četností,
- přes konstrukci měr (polohy, variability, šikmosti a špičatosti), které mohou být založeny na tzv. kvantilech.

Nejpoužívanější kvantily:

- \tilde{x}_{50} — člen souboru na dvě stejně velké poloviny, tedy 50%-ní kvantil, nazývá se medián neboli prostřední hodnota (lat. medius = prostřední)
- kvantily menší než \tilde{x}_{50} se nazývají dolní,
- kvantily větší než \tilde{x}_{50} se nazývají horní,
- kvartily (lat. quartus = čtvrtý),
- decily (lat. decilus = desátý),
- kvintily (lat. quintus = pátý),
- percentily, které rozdělují uspořádanou řadu hodnot znaku na 100 stejně četných částí.

Nalezení některých kvantilů v našich datech

- medián je 26. člen řady seřazených hodnot znaku
- dolní kvartil \tilde{x}_{25} bude člen, jehož pořadové číslo je $\frac{n+1}{4} = \frac{52}{4} = 13$, což je 10923,
- horní kvartil \tilde{x}_{75} bude člen, jehož pořadové číslo je $3\frac{n+1}{4} = 3\frac{52}{4} = 39$, což je 16983.

Kdyby měla řada 136 členů, pak by $\frac{n+1}{4} = \frac{137}{4} = 34,25$ a za dolní kvartil by bylo nutné stanovit některou hodnotu ležící mezi hodnotami s pořadovým číslem 34 a 35.

Nabízí se 2 řešení:

a) průměr 34 a 35 hodnoty.

b) 34. člen řady násobený 0,75 + 35. člen řady násobený 0,25.

6400	7975	8434	9059	9808
9876	10015	10349	10560	10805
10825	10884	10923*	11143	11429
12000	12420	12956	13075	13213
13243	13630	13804	13809	13899
14390*	14406	14569	14890	14906
14928	15243	15300	15908	15963
16075	16343	16451	16983*	17473
17825	17875	17962	18563	19849
20016	20102	20377	22012	22865
25416				

Nalezení některých kvantilů v našich datech

- první decil \tilde{x}_{10} bude hodnota, která odpovídá prvku s pořadovým číslem $\frac{n+1}{10} = \frac{52}{10} = 5,2$, získáme $\tilde{x}_{10} = \frac{9808+9876}{2} = 9842$
- druhý decil \tilde{x}_{20} bude hodnota, která odpovídá prvku s pořadovým číslem $2\frac{n+1}{10} = 2\frac{52}{10} = 10,4$, získáme $\tilde{x}_{20} = \frac{10805+10825}{2} = 10815$
- poslední decil \tilde{x}_{90} bude hodnota, která odpovídá prvku s pořadovým číslem $9\frac{n+1}{10} = 9\frac{52}{10} = 46,8$, získáme $\tilde{x}_{90} = \frac{20016+20102}{2} = 20059$

Ověřte, že $\tilde{x}_{30} = 11714,5$, $\tilde{x}_{40} = 13228$, $\tilde{x}_{50} = 14390$, $\tilde{x}_{60} = 15090,5$, $\tilde{x}_{70} = 16209$, $\tilde{x}_{80} = 17850$

Obecný vztah pro stanovení přibližné hodnoty kvantilu pomocí interpolace

$$\frac{\tilde{x}_i - x_d}{x_h - x_d} = \frac{i - i_d}{i_h - i_d},$$

kde i je relativní kumulativní četnost

x_d je dolní hranice intervalu, x_h je horní hranice intervalu,

$$\tilde{x}_i = x_d + \frac{i - i_d}{i_h - i_d}(x_h - x_d).$$

Označme délku intervalu $x_h - x_d$ jako h a relativní třídní četnost v procentech $i_h - i_d$ jako $\frac{n_i}{n}100$ lze předchozí vztah psát jako

$$\tilde{x}_i = x_d + \frac{i - i_d}{\frac{n_i}{n}100}h.$$

Např. medián $\tilde{x}_{50} = 11700 + \frac{50-29,4}{23,5}2800 = 14155,6$,

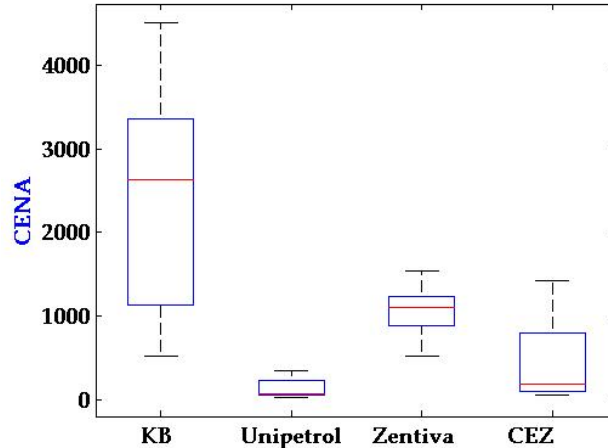
dolní kvartil $\tilde{x}_{25} = 8900 + \frac{25-5,9}{23,5}2800 = 11176,4$,

horní kvartil $\tilde{x}_{75} = 14500 + \frac{75-52,9}{23,5}2800 = 17132$,

první decil $\tilde{x}_{10} = 8900 + \frac{10-5,9}{23,5}2800 = 9387,2$.

2.3. Grafy

V popisné statistice se často pracuje s tzv. krabicovým diagramem (boxplot).

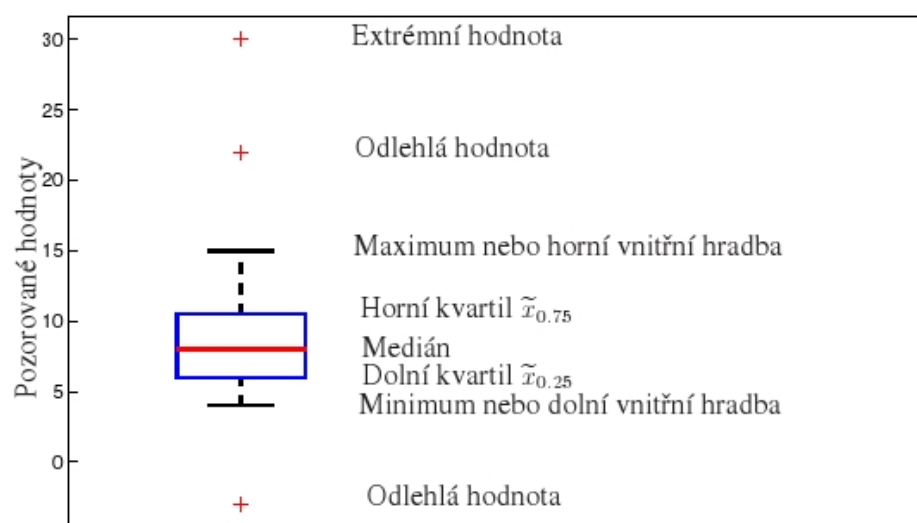


Obrázek: Boxplot

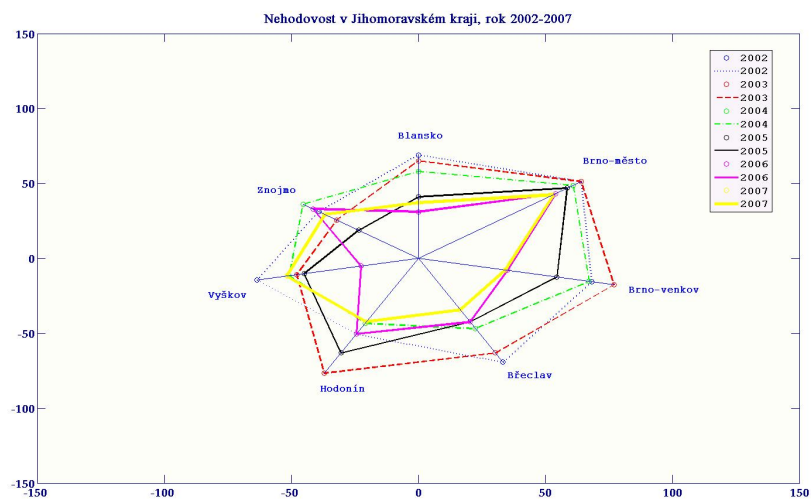
Krabicový diagram nám umožňuje posoudit symetrii a variabilitu námi zvolených dat a existenci extrémních hodnot. Význam boxplotu objasníme na dalším obrázku. Nejprve je ale třeba zavést následující pojmy:

- dolní vnitřní hradba: $\tilde{x}_{25} - 1,5R_Q$,
- horní vnitřní hradba: $\tilde{x}_{75} + 1,5R_Q$,
- dolní vnější hradba: $\tilde{x}_{25} - 3R_Q$,
- horní vnější hradba: $\tilde{x}_{75} + 3R_Q$.
- Kvartilové rozpětí je $R_Q = \tilde{x}_{75} - \tilde{x}_{25}$.

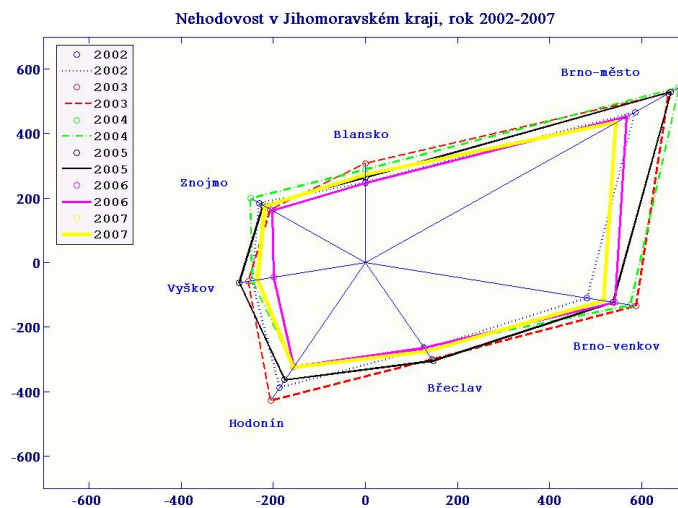
Symbol křížku v obrázku označuje extrémní hodnoty, které leží za vnějšími hradbami.



Obrázek: Konstrukce boxplotu



Obrázek: Hvězdicový diagram



Obrázek: Hvězdicový diagram

3

Popisná statistika

3.1. Momenty

Zatímco kvantily kvantitativního statistického znaku x jsou vždy určité konkrétní jednotlivé hodnoty znaku (nebo jen z několika blízkých hodnot znaků určené) jsou momenty vždy funkcemi všech hodnot daného znaku v daném souboru.

Mějme jednorozměrný statistický soubor o rozsahu n .

k -tý moment kvantitativního statistického znaku je obecně definován jako

$${}_am_{x,k} = \frac{1}{n} \sum_i^n (x_i - a)^k$$

kde

$x_i, i = 1, 2, \dots, n$ jsou jednotlivé hodnoty kvantitativního statistického znaku,

a je nějaká konstanta

k je přirozené číslo vyjadřující stupeň momentu

Symbol na levé straně rovnice čteme jako k -tý moment z x kolem a .

3.1.1. Obecné momenty

Pro $a = 0$ dostáváme k -tý moment x kolem nuly (kolem počátku) a moment nazýváme k -tý obecný moment x

$${}_0m_{x,k} = \frac{1}{n} \sum_i^n x_i^k.$$

Nejpoužívanější je 1. obecný moment, který se nazývá střední hodnota:

$${}_0m_{x,1} = \frac{1}{n} \sum_i^n x_i = \bar{x}.$$

Podobně pracujeme s vyššími obecnými momenty

$${}_0m_{x,2} = \frac{1}{n} \sum_i^n x_i^2, {}_0m_{x,3} = \frac{1}{n} \sum_i^n x_i^3, {}_0m_{x,4} = \frac{1}{n} \sum_i^n x_i^4,$$

atd.

3.1.2. Centrální momenty

Učiníme-li aritmetický průměr počátkem neboli centrem (středem) počítání, dostaneme k -tý centrální moment.

Do vzorce pro moment dosadíme $a = \bar{x}$:

$$\bar{x}m_{x,k} = \frac{1}{n} \sum_i^n (x_i - \bar{x})^k.$$

Nejpoužívanější je 2. centrální moment, který se nazývá rozptyl a je dán vztahem:

$$\begin{aligned}
\bar{x}m_{x,2} &= \frac{1}{n} \sum_i^n (x_i - \bar{x})^2. \\
\bar{x}m_{x,2} &= \frac{1}{n} \sum_i^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_i^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \\
&= \frac{\sum_i^n x_i^2}{n} - 2\bar{x} \frac{\sum_i^n x_i}{n} + \frac{n\bar{x}^2}{n} = \\
&= \frac{\sum_i^n x_i^2}{n} - 2\bar{x}^2 + \bar{x}^2 = \\
&= {}_0m_{x,2} - ({}_0m_{x,1})^2
\end{aligned}$$

Druhý centrální moment tedy lze vyjádřit jako druhý obecný moment zmenšený o čtverec prvního obecného momentu.

$$\begin{aligned}
\bar{x}m_{x,1} &= \frac{1}{n} \sum_i^n (x_i - \bar{x}) = \\
&= \frac{\sum_i^n x_i}{n} - \frac{n\bar{x}}{n} = \\
&= \bar{x} - \bar{x} = 0
\end{aligned}$$

První centrální moment je vždy roven nule.

Třetí centrální moment lze vyjádřit pomocí prvních tří obecných momentů

$$\begin{aligned}
\bar{x}m_{x,3} &= \frac{1}{n} \sum_i^n (x_i - \bar{x})^3 = \frac{1}{n} \sum_i^n (x_i^3 - 3x_i^2\bar{x} + 3x_i\bar{x}^2 - \bar{x}^3) = \\
&= \frac{\sum_i^n x_i^3}{n} - 3\bar{x} \frac{\sum_i^n x_i^2}{n} + 3\bar{x} \frac{\sum_i^n x_i}{n} - \frac{n\bar{x}^3}{n} = \\
&= {}_0m_{x,3} - 3{}_0m_{x,1} \cdot {}_0m_{x,2} - 2({}_0m_{x,1})^3
\end{aligned}$$

Čtvrtý centrální moment lze vyjádřit pomocí prvních čtyř obecných momentů

$$\begin{aligned}
\bar{x}m_{x,4} &= \frac{1}{n} \sum_i^n (x_i - \bar{x})^4 = \frac{1}{n} \sum_i^n (x_i^4 - 4x_i^3\bar{x} + 6x_i^2\bar{x}^2 - 4x_i\bar{x}^3 + \bar{x}^4) = \\
&= {}_0m_{x,4} - 4{}_0m_{x,1} \cdot {}_0m_{x,3} + 6({}_0m_{x,1})^2 \cdot {}_0m_{x,2} - 3({}_0m_{x,1})^4.
\end{aligned}$$

3.1.3. Charlierův test

Při ručních výpočtech docházelo při výpočtu momentů k chybám. Proto se pro kontrolu výpočtů používal test Karla Ludwiga Wilhelam Charliera (1861-1934).

Tato kontrola je založena na platnosti vztahu

$$\begin{aligned} \frac{1}{n} \sum_i^n (x_i + 1)^4 &= \frac{1}{n} \sum_i^n x_i^4 + 4 \frac{1}{n} \sum_i^n x_i^3 + 6 \frac{1}{n} \sum_i^n x_i^2 + 4 \frac{1}{n} \sum_i^n x_i + 1 = \\ &= {}_0m_{x,4} + 4{}_0m_{x,3} + 6{}_0m_{x,2} + 4{}_0m_{x,1} + 1. \end{aligned}$$

3.1.4. Úloha

Zvolte data, např. teplotní řadu nebo hodnoty znečištění z vybrané meteorologické stanice a určete všechny uvedené momenty.

3.2. Normované momenty

Jednotlivá pozorování (hodnoty) kvantitativního statistického znaku x u výběrového souboru o rozsahu n , tj. x_1, x_2, \dots, x_n jsou vždy vyjádřena v nějakých měrných jednotkách (v Kč, kusech, kg, metrech, apod.). Ve stejných měrných jednotkách jsou vyjádřeny i odchylky jednotlivých pozorování od aritmetického průměru, tj. $x_i - \bar{x}$, $i = 1, 2, \dots, n$ směrodatnou odchylkou, získá se tzv. směrodatná proměnná

$$u_i = \frac{x_i - \bar{x}}{s_x}, \quad i = 1, 2, \dots, n,$$

jejíž jednotlivé hodnoty u_i jsou čísla bez rozměru.

Je možné uvažovat momenty směrodatné proměnné, které se často nazývají normovanými momenty.

První obecný moment směrodatné proměnné

$${}_0m_{\bar{u},1} = \frac{1}{n} \sum_{i=1}^n u_i = \bar{u}$$

je aritmetickým průměrem směrodatné proměnné.

Po dosazení dostáváme

$${}_0m_{\bar{u},1} = \frac{1}{n} \sum_{i=1}^n \frac{x_i - \bar{x}}{s_x} = \frac{1}{ns_x} \sum_{i=1}^n (x_i - \bar{x}) = 0.$$

První obecný moment směrodatné proměnné ${}_0m_{\bar{u},1}$ neboli aritmetický průměr směrodatné proměnné \bar{u} je tedy roven nule. Z toho plyne, že obecné momenty směrodatné proměnné jsou zároveň centrálními momenty této proměnné:

$${}_0m_{\bar{u},k} = \frac{1}{n} \sum_{i=1}^n u_i^k = \frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})^k = m_{u,k}.$$

Druhý moment směrodatné proměnné a tedy zároveň i její rozptyl je:

$$\begin{aligned} m_{u,2} &= \frac{1}{n} \sum_{i=1}^n u_i^2 = \frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})^2 = s_u^2 = \frac{1}{n} \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{s_x^2} = \\ &= \frac{1}{s_x^2} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{s_x^2} s_x^2 = 1. \end{aligned}$$

Rozptýl směrodatné proměnné s_u^2 a tedy i směrodatná odchylka směrodatné proměnné jsou rovny jedné.

Třetí normovaný moment Třetí normovaný moment směrodatné proměnné lze vyjádřit pomocí centrálních momentů původní proměnné x takto:

$$\begin{aligned} m_{u,3} &= \frac{1}{n} \sum_{i=1}^n u_i^3 = \frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})^3 = \frac{1}{n} \sum_{i=1}^n \frac{(x_i - \bar{x})^3}{s_x^3} = \\ &= \frac{1}{s_x^3} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3 = \frac{m_{x,3}}{m_{x,2} \sqrt{m_{x,2}}} . \end{aligned}$$

Čtvrtý normovaný moment Čtvrtý normovaný moment směrodatné proměnné lze vyjádřit pomocí centrálních momentů původní proměnné x takto:

$$\begin{aligned} m_{u,4} &= \frac{1}{n} \sum_{i=1}^n u_i^4 = \frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})^4 = \frac{1}{n} \sum_{i=1}^n \frac{(x_i - \bar{x})^4}{s_x^4} = \\ &= \frac{1}{s_x^4} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4 = \frac{m_{x,4}}{(m_{x,2})^2} . \end{aligned}$$

3.2.1. Příklad – domácnosti a počet dětí

	četnost	x_i	x_i^2	x_i^3	x_i^4	$(x_i + 1)^4$
1.	3	9	27	81	256	
2.	1	1	1	1	16	
3.	1	1	1	1	16	
4.	3	9	27	81	256	
5.	1	1	1	1	16	
6.	1	1	1	1	16	
7.	2	4	8	16	81	
8.	1	1	1	1	16	
9.	0	0	0	0	1	
součet	13	27	67	183	674	

$${}_0m_{x,1} = \bar{x} = \frac{13}{9} \doteq 1,44 ,$$

$${}_0m_{x,2} = \frac{27}{9} \doteq 3,00 ,$$

$${}_0m_{x,3} = \frac{67}{9} \doteq 7,44 ,$$

$${}_0m_{x,4} = \frac{183}{9} \doteq 20,33 .$$

$$m_{x,2} \doteq 3,00 - 1,44^2 = 0,93 ,$$

$$m_{x,3} \doteq 7,44 - 3 \cdot 1,44 \cdot 3 + 2 \cdot 1,44^2 = -0,53 ,$$

$$m_{x,4} \doteq 20,33 - 4 \cdot 1,44 \cdot 7,44 + 6 \cdot 1,44^2 \cdot 3 - 3 \cdot 1,44^4 = 1,90 ,$$

$$m_{u,3} \doteq \frac{-2,53}{0,93 \sqrt{0,93}} \doteq -2,83 ,$$

$$m_{u,4} \doteq \frac{1,90}{0,93^2} \doteq 2,20.$$

Užijeme-li Charlierův test dostáváme

$$\begin{aligned} \frac{1}{n} \sum_i^n (x_i + 1)^4 &= \frac{1}{n} \sum_i^n x_i^4 + 4 \frac{1}{n} \sum_i^n x_i^3 + 6 \frac{1}{n} \sum_i^n x_i^2 + 4 \frac{1}{n} \sum_i^n x_i + 1 = \\ &= {}_0m_{x,4} + 4{}_0m_{x,3} + 6{}_0m_{x,2} + 4{}_0m_{x,1} + 1 \\ \frac{674}{9} &= 20,33 + 4 \cdot 7,44 + 6 \cdot 3,00 + 4 \cdot 1,44 + 1 \\ &74,89 \doteq 74,85. \end{aligned}$$

3.2.2. Výpočet momentů u rozdělení četností

U rozdělení četností kvantitativního statistického znaku x , kde uvažujeme varianty znaku x_i , $i = 1, 2, \dots, k$, které se vyskytují s četnostmi n_i , $i = 1, 2, \dots, k$, platí, že rozsah souboru n je roven součtu všech absolutních četností:

$$n = \sum_{i=1}^n n_i.$$

Součty $\sum_{i=1}^n x_i^k$, $k = 1, 2, 3, 4$ jsou u rozdělení četností ekvivalentní součtům $\sum_{i=1}^k x_i^k n_i$, $k = 1, 2, 3, 4$.

Proto první čtyři obecné momenty budou

$$\begin{aligned} {}_0m_{x,1} &= \bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i, \\ {}_0m_{x,2} &= \frac{1}{n} \sum_{i=1}^k x_i^2 n_i, \\ {}_0m_{x,3} &= \frac{1}{n} \sum_{i=1}^k x_i^3 n_i, \\ {}_0m_{x,4} &= \frac{1}{n} \sum_{i=1}^k x_i^4 n_i. \end{aligned}$$

3.2.3. Výpočet momentů u intervalového rozdělení četností

Metoda vhodně zvoleného počátku Pokud budeme počítat momenty u intervalového rozdělení četností, je třeba si uvědomit, že výpočty přinášejí pouze přibližné hodnoty.

Při výpočtu momentů u intervalového dělení četností můžeme postupovat stejně jako u rozdělení četností, uvažujeme-li středy třídních intervalů místo jednotlivých variant znaků a třídní četnosti místo četností jednotlivých variant.

Vzhledem k tomu, že středy intervalů mohou být několikacíferná čísla, zavádíme obvykle zjednodušení výpočtu, spočívající v zavedení nové proměnné v_i , $i = 1, 2, \dots, k$, kde k je počet intervalů,

$$v_i = \frac{x_i - a}{h}, i = 1, 2, \dots, k,$$

kde x_i jsou středy intervalů, h je délka intervalů a kde a je libovolná konstanta (předběžný průměr, vhodně zvolený počátek).

Metoda vhodně zvoleného počátku

Po zavedení pomocné proměnné v_i se momenty vypočtou pomocí vztahů:

$${}_0m_{v,1} = \bar{v} = \frac{1}{n} \sum_{i=1}^k v_i n_i ,$$

$${}_0m_{v,2} = \frac{1}{n} \sum_{i=1}^k v_i^2 n_i ,$$

$${}_0m_{v,3} = \frac{1}{n} \sum_{i=1}^k v_i^3 n_i ,$$

$${}_0m_{v,4} = \frac{1}{n} \sum_{i=1}^k v_i^4 n_i .$$

3.2.4. Příklad

		střed	četnost n_i	v_i	$v_i n_i$	$v_i^2 n_i$	$v_i^3 n_i$	$v_i^4 n_i$	$(v_i + 1)^4 n_i$
1.	6100–8900	7500	3	-3	-9	27	-81	243	48
2.	8900–11700	10300	3	-2	-24	48	-96	192	12
3.	11700–14500	13100	3	-1	-12	12	-12	12	0
4.	14500–17300	15900	3	0	0	0	0	0	12
5.	17300–20100	18700	3	1	7	7	7	7	112
6.	20100–22900	21500	3	2	8	16	32	64	324
7.	22900–25700	24300	3	3	3	9	27	81	256
součet	—	—	51	—	-27	119	-123	599	764

$$v_i = \frac{x_i - 15900}{2800}$$

$${}_0m_{v,1} = \bar{v} = -\frac{27}{51} \doteq -0,53 ,$$

$${}_0m_{v,2} = \frac{119}{51} \doteq 2,33 ,$$

$${}_0m_{v,3} = -\frac{123}{51} \doteq -2,41 ,$$

$${}_0m_{v,4} = \frac{599}{51} \doteq 11,75 .$$

Užijeme-li Charlierův test dostáváme

$$\begin{aligned} \frac{1}{n} \sum_i^n (v_i + 1)^4 &= {}_0m_{v,4} + 4{}_0m_{v,3} + 6{}_0m_{v,2} + 4{}_0m_{v,1} + 1 \\ \frac{64}{51} &= 11,75 - 4 \cdot 2,41 + 6 \cdot 2,33 - 4 \cdot 0,53 + 1 \\ 14,98 &\doteq 14,97. \end{aligned}$$

$$m_{v,2} = 2,33 - (-0,53)^2 \doteq 2,05 ,$$

$$m_{v,3} = -2,41 - 3(-0,53)^2 \cdot 2,33 + 2 \cdot (-0,53)^2 \doteq 1,00 ,$$

$$m_{v,4} = 11,75 - 4(-0,53) \cdot (-2,41) + 6 \cdot (-0,53)^2 \cdot 2,33 - 3(-0,53)^4$$

$$\begin{aligned}
&= \doteq 10,33, \\
{}_0m_{x,1} &= \bar{x} = 2800 \cdot (-0,53) + 15900 = 14416, \\
{}_0m_{x,2} &= s_x^2 = 2800^2 \cdot 2,05 \doteq 16072000, \\
m_{u,3} &= \frac{1,000}{2,05\sqrt{2,05}} \doteq 0,34, \\
m_{u,4} &= \frac{10,33}{2,05^2} \doteq 2,46.
\end{aligned}$$

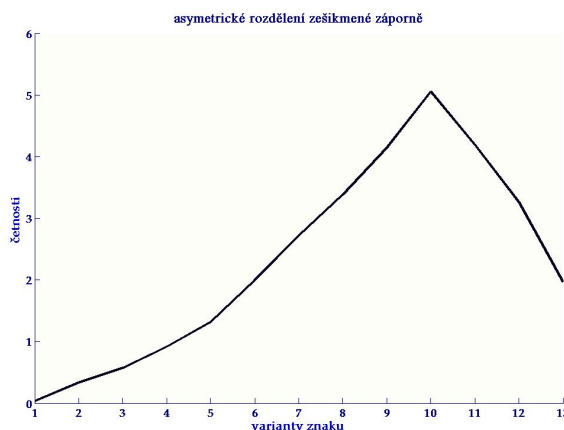
3.2.5. Vztahy mezi průměry

V symetrickém rozdělení spadá aritmetický průměr \bar{x} , medián \tilde{x} i modus \hat{x} do jednoho bodu.

Čím více se rozdělení četností blíží symetrickému, tím méně se tyto charakteristiky odlišují.

3.2.6. Rozdělení četností

Schéma asymetrického rozdělení četností zešikmeného záporně



3.2.7. Vztahy mezi průměry

V asymetrickém rozdělení zešikmeném záporně platí

$$\bar{x} < \tilde{x} < \hat{x}.$$

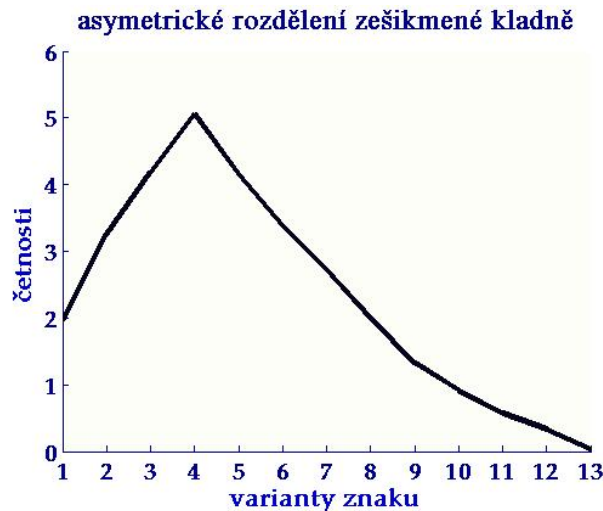
Není-li asymetrické rozdělení příliš (extrémně) nesouměrné, je vzdálenost mediánu od aritmetického průměru většinou přibližně jednou třetinou vzdálenosti mezi modem a aritmetickým průměrem.

V mírně nesouměrném rozdělení tedy platí

$$\bar{x} - \hat{x} \sim 3(\bar{x} - \tilde{x}).$$

Není-li asymetrické rozdělení příliš (extrémně) nesouměrné, je vzdálenost mediánu od aritmetického průměru většinou přibližně jednou třetinou vzdálenosti mezi modem a aritmetickým průměrem.

Schéma asymetrického rozdělení četností zešikmeného kladně



3.2.8. Míry variability

Dosud jsme se zejména zabývali mírami polohy. Jejich účelem je umožnit srovnání úrovně hodnot znaků ve dvou či více souborech.

Míra polohy ale sama nestačí ke srovnání sledovaného znaku u dvou či více souborů, neboť měří pouze obecnou úroveň hodnot zkoumaného znaku.

Zcela různé soubory či zcela různá rozdělení mohou mít stejné míry polohy.

Je tedy zřejmé, že kromě úrovně, na niž se pohybují hodnoty sledovaných znaků, je třeba zkoumat i to, jak se jednotlivé hodnoty znaku liší od míry polohy i vzájemně.

A právě tyto odlišnosti hodnot sledovaného znaku v souhrnu v daném souboru nazýváme měnlivostí neboli variabilitou (variancí).

3.2.9. Míry variability

Rozpětí

$$R_x = x_{\max} - x_{\min}.$$

Kvartilové rozpětí

$$\tilde{x}_{75} - \tilde{x}_{25}.$$

Decilové rozpětí

$$\tilde{x}_{90} - \tilde{x}_{10}.$$

Percentilové rozpětí

$$\tilde{x}_{99} - \tilde{x}_1.$$

Kvartilová odchylka

$$Q = \frac{(\tilde{x}_{75} - \tilde{x}) + (\tilde{x} - \tilde{x}_{25})}{2} = \frac{(\tilde{x}_{75} - \tilde{x}_{25})}{2}.$$

Decilová odchylka

$$\frac{(\tilde{x}_{90} - \tilde{x}_{80}) + (\tilde{x}_{80} - \tilde{x}_{70}) + \dots + (\tilde{x}_{20} - \tilde{x}_{10})}{8} = \frac{(\tilde{x}_{90} - \tilde{x}_{10})}{8}.$$

Obdobně percentilová odchylka

$$\frac{\tilde{x}_{99} - \tilde{x}_1}{98}.$$

Momentové míry variability

Odhad rozptylu s^2 .

Momentová míra relativní variability

$$c_x = \frac{s_X}{\bar{x}}.$$

Výpočet celkového rozptylu z dílčích rozptylů

$$\bar{x}^2 = \sum_{i=1}^k s_i^2 n_i \sum_{i=1}^k n_i.$$

Kvantilová šikmost

$$\tau = \frac{(x_{\max} - \tilde{x}) - (\tilde{x} - x_{\min})}{(x_{\max} - \tilde{x}) + (\tilde{x} - x_{\min})} = \frac{x_{\max} - x_{\min} - 2\tilde{x}}{x_{\max} - x_{\min}},$$

v případě symetrického rozdělení je rovna nule.

Percentilová šikmost

$$\tau_i = \frac{(\tilde{x}_{100-i} - \tilde{x}) - (\tilde{x} - \tilde{x}_i)}{(\tilde{x}_{100-i} - \tilde{x}) + (\tilde{x} - \tilde{x}_i)} = \frac{\tilde{x}_{100-i} + \tilde{x}_i - 2\tilde{x}}{\tilde{x}_{100-i} - \tilde{x}_i}.$$

3.2.10. Míry šikmosti

Výběrová šikmost

$$A_3 = \frac{m_{x,3}}{\sqrt{m_{x,2}^3}} = \frac{\sum_{i=1}^k (x_i - \bar{x})^3 n_i}{ns_x^2}$$

Pearsonova míra šikmosti

$$\tau' = \frac{\bar{x} - \hat{x}}{s_x}.$$

Jednoduché míra šikmosti

$$\tau'' = \frac{n' - n''}{n},$$

kde n' je počet hodnot menších než aritmetický průměr,

kde n'' je počet hodnot větších než aritmetický průměr.

Kvantilová šikmost

$$K = \frac{(x_{\max} - x_{\min})}{(\tilde{x}_{75} - \tilde{x}_{25})},$$

$$K' = \frac{(\tilde{x}_{99.5} - \tilde{x}_{0.5})}{(\tilde{x}_{75} - \tilde{x}_{25})},$$

$$K'' = \frac{(\tilde{x}_{99} - \tilde{x}_1)}{(\tilde{x}_{75} - \tilde{x}_{25})}.$$

3.2.11. Míry špičatosti

Výběrová špičatost

$$A_4 = \frac{m_{x,4}}{\sqrt{m_{x,2}^2}} = \frac{\sum_{i=1}^k (x_i - \bar{x})^4 n_i}{ns_x^4}$$

3.2.12. Empirická distribuční funkce

$F_n(x)$ je empirická (výběrová) distribuční funkce $X_{(1)}, X_{(2)}, X_{(3)}, \dots, X_{(n-1)}, X_{(n)}$.

$$F_n(x) = \begin{cases} 0 & x < X_{(1)} \\ \frac{i}{n} & X_{(i)} \leq x < X_{(i+1)} \\ 1 & x \geq X_{(n)} \end{cases}$$

Příklad (ověření rovnoměrného rozdělení)

Prvních 10 čísel z generátoru náhodných čísel bylo

0,93 0,35 0,66 0,93 0,14 0,23 0,08 0,23 0,21 0,59

Zjistěte, zda tyto veličiny lze pokládat za výběr z $R(0, 1)$.

V MATLABU LZE VYGENEROVAT TAKTO:

`X=rand(10,1)`

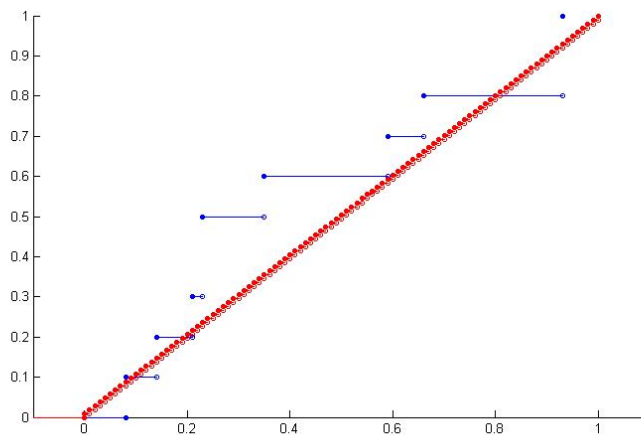
Nejprve čísla uspořádáme vzestupně: 0,08 0,14 0,21 0,23 0,23 0,35 0,59 0,66 0,93 0,93.

Dále vytvoříme empirickou (výběrovou) distribuční funkci

$$F_n(x) = \begin{cases} 0 & x < 0,08_{(1)} \\ \frac{1}{10} & 0,08_{(1)} \leq x < 0,14_{(2)} \\ \frac{2}{10} & 0,14_{(2)} \leq x < 0,21_{(3)} \\ \frac{3}{10} & 0,21_{(3)} \leq x < 0,23_{(4)} \\ \frac{5}{10} & 0,23_{(4,5)} \leq x < 0,35_{(6)} \\ \frac{6}{10} & 0,35_{(6)} \leq x < 0,59_{(7)} \\ \frac{7}{10} & 0,59_{(7)} \leq x < 0,66_{(8)} \\ \frac{8}{10} & 0,66_{(8)} \leq x < 0,93_{(9)} \\ 1 & x \geq 0,93_{(9,10)} \end{cases}$$

Hodnoty této distribuční funkce porovnáme s

$$F_0(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$



4

Dvourozměrný statistický soubor

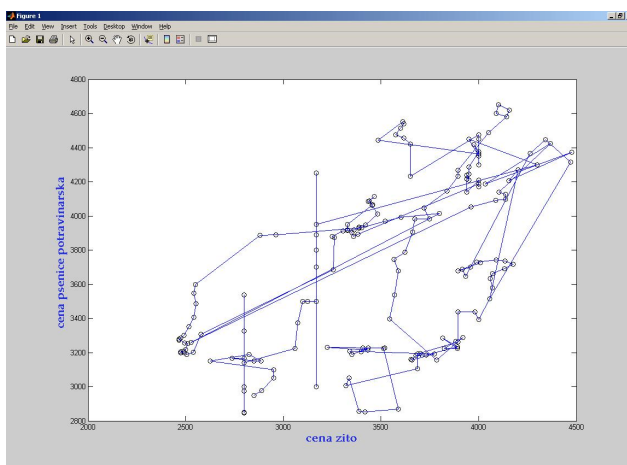
4.1. Charakteristiky náhodného vektoru, korelace

Jestliže vyšetřujeme na každé statistické jednotce dva znaky X, Y , máme podobně jako v případě jednorozměrného statistického souboru dvě možnosti:

- pracovat se všemi daty,
- data uspořádat do četnostní tabulky.

Graf ceny žita a potravinářské pšenice v období leden 1999 – leden 2006.

Data pochází z plodinové burzy v Brně.



ad a) Statistický soubor tvoří n uspořádaných dvojic $(x_1, y_1), \dots, (x_n, y_n)$. Základní charakteristiky jsou aritmetické průměry a rozptyly

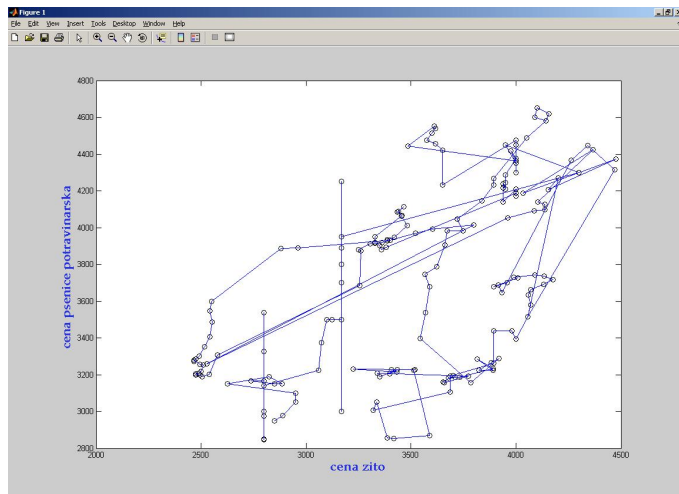
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad s_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2,$$

kovariance

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \left[\frac{1}{n} \sum_{i=1}^n x_i y_i \right] - \bar{x} \cdot \bar{y},$$

korelační koeficient

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2][n \sum y_i^2 - (\sum y_i)^2]}}, \text{ je-li } s_x \cdot s_y \neq 0.$$



$$\bar{x} = 3485,45 \text{ Kč}, \bar{y} = 3699,20 \text{ Kč},$$

$$s_x^2 = 283510,97, s_y^2 = 245469,57, s_{xy} = 156730,65, r_{xy} = 0,59411.$$

4.1.1. Četnostní tabulka

ad b) Nechť $a_1 < \dots < a_r$, resp. $b_1 < \dots < b_s$, označují buď varianty znaku X resp. Y nebo středy intervalů při intervalovém třídění obou znaků. Označme symbolem n_{ij} , $i = 1, \dots, r$, $j = 1, \dots, s$, počet těch pozorovaných dvojic $(x_1, y_1), \dots, (x_n, y_n)$ jejichž x -ová hodnota je rovna a_i a současně jejichž y -ová hodnota je rovna b_j resp. jejichž x -ová hodnota patří do i -tého intervalu hodnot znaku X a současně y -ová hodnota patří do j -tého intervalu hodnot znaku Y . Označme dále

$$n_{i.} = \sum_{j=1}^s n_{ij}, \quad n_{.j} = \sum_{i=1}^r n_{ij}, \quad n = \sum_{i=1}^r \sum_{j=1}^s n_{ij}.$$

Tabulka rozdělení četností vypadá následovně

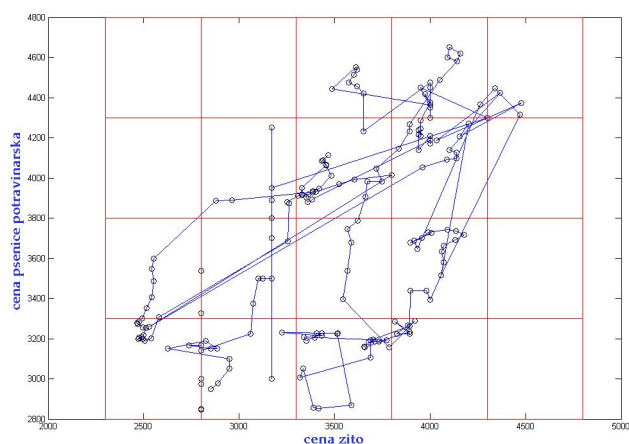
	b_1	\dots	b_j	\dots	b_s	\sum
a_1	n_{11}	\dots	n_{1j}	\dots	n_{1s}	$n_{1.}$
\vdots	\vdots		\vdots		\vdots	\vdots
a_i	n_{i1}	\dots	n_{ij}	\dots	n_{is}	$n_{i.}$
\vdots	\vdots		\vdots		\vdots	\vdots
a_r	n_{r1}	\dots	n_{rj}	\dots	n_{rs}	$n_{r.}$
\sum	$n_{.1}$	\dots	$n_{.j}$	\dots	$n_{.s}$	n

Výpočet číselných charakteristik

$$\bar{x} = \frac{1}{n} \sum_{i=1}^r a_i n_{i.}, \quad \bar{y} = \frac{1}{n} \sum_{j=1}^s b_j n_{.j},$$

$$s_x^2 = \frac{1}{n} \sum_{i=1}^r (a_i - \bar{x})^2 n_{i.}, \quad s_y^2 = \frac{1}{n} \sum_{j=1}^s (b_j - \bar{y})^2 n_{.j},$$

$$s_{xy} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^s (a_i - \bar{x})(b_j - \bar{y}) n_{ij} = \left[\frac{1}{n} \sum_{i=1}^r \sum_{j=1}^s a_i b_j n_{ij} \right] - \bar{x} \cdot \bar{y}.$$



Tabulka rozdělení četností cen žita a potravinářské pšenice vypadá následovně

	b_1	b_2	b_3	b_4	\sum
a_1	24	8	0	0	32
a_2	11	8	8	0	27
a_3	24	5	26	7	62
a_4	7	18	22	16	63
a_5	0	0	0	4	4
\sum	66	39	56	27	188

$$\bar{x} = 3496,81 \text{ Kč}$$

$$\bar{y} = 3667,02 \text{ Kč}$$

$$s_x^2 = 308340,88, s_y^2 = 292157,08, s_{xy} = 157820,28, r_{xy} = 0,52582$$

Tabulka rozdělení četností cen žita a potravinářské pšenice vypadá následovně

	$\langle 2800, 3300 \rangle$ $b_1 = 3050$	$\langle 3300, 3800 \rangle$ $b_2 = 3550$	$\langle 3800, 4300 \rangle$ $b_3 = 4050$	$\langle 4300, 4800 \rangle$ $b_4 = 4550$	\sum
$a_1 = 2550$	24	8	0	0	32
$a_2 = 3050$	11	8	8	0	27
$a_3 = 3550$	24	5	26	7	62
$a_4 = 4050$	7	18	22	16	63
$a_5 = 4550$	0	0	0	4	4
\sum	66	39	56	27	188

Četnostní tabulka (znak X reprezentuje výšku, znak Y reprezentuje váhu)

	$\langle 46, 59 \rangle$ $b_1 = 52,5$	$\langle 59, 72 \rangle$ $b_2 = 65,5$	$\langle 72, 85 \rangle$ $b_3 = 78,5$	$\langle 85, 91 \rangle$ $b_4 = 91,5$	\sum
$\langle 158, 169 \rangle$ $a_1 = 163,5$	6	1	0	1	8
$\langle 169, 180 \rangle$ $a_2 = 174,5$	1	8	3	0	12
$\langle 180, 191 \rangle$ $a_3 = 185,5$	0	0	5	1	6
$\langle 191, 202 \rangle$ $a_4 = 196,5$	0	0	1	3	4
\sum	7	9	9	5	30

$$\bar{x} = \frac{1}{30}(163,5 \times 8 + 174,5 \times 12 + 185,5 \times 6 + 196,5 \times 4) = 176,7 \text{ cm}$$

$$\bar{y} = \frac{1}{30}(52,5 \times 7 + 65,5 \times 9 + 78,5 \times 9 + 91,5 \times 5) = 70,7 \text{ kg}$$

Četnostní tabulka (znak X reprezentuje výšku, znak Y reprezentuje váhu)

	$\langle 46, 59 \rangle$ $b_1 = 52,5$	$\langle 59, 72 \rangle$ $b_2 = 65,5$	$\langle 72, 85 \rangle$ $b_3 = 78,5$	$\langle 85, 91 \rangle$ $b_4 = 91,5$	\sum
$\langle 158, 169 \rangle a_1 = 163,5$	6	1	0	1	8
$\langle 169, 180 \rangle a_2 = 174,5$	1	8	3	0	12
$\langle 180, 191 \rangle a_3 = 185,5$	0	0	5	1	6
$\langle 191, 202 \rangle a_4 = 196,5$	0	0	1	3	4
\sum	7	9	9	5	30

$$\bar{x} = 176,7 \text{ cm} \quad \bar{y} = 70,7 \text{ kg}$$

$$s_x^2 = \frac{1}{30} ((163,5 - 176,7)^2 \times 8 + (174,5 - 176,7)^2 \times 12 + (185,5 - 176,7)^2 \times 6 + (196,5 - 176,7)^2 \times 4) = 116,16 \text{ cm}^2$$

Četnostní tabulka (znak X reprezentuje výšku, znak Y reprezentuje váhu)

	$\langle 46, 59 \rangle$ $b_1 = 52,5$	$\langle 59, 72 \rangle$ $b_2 = 65,5$	$\langle 72, 85 \rangle$ $b_3 = 78,5$	$\langle 85, 91 \rangle$ $b_4 = 91,5$	\sum
$\langle 158, 169 \rangle a_1 = 163,5$	6	1	0	1	8
$\langle 169, 180 \rangle a_2 = 174,5$	1	8	3	0	12
$\langle 180, 191 \rangle a_3 = 185,5$	0	0	5	1	6
$\langle 191, 202 \rangle a_4 = 196,5$	0	0	1	3	4
\sum	7	9	9	5	30

$$\bar{x} = 176,7 \text{ cm} \quad \bar{y} = 70,7 \text{ kg} \quad s_x^2 = 116,16 \text{ cm}^2$$

$$s_y^2 = \frac{1}{30} ((52,5 - 70,7)^2 \times 7 + (65,5 - 70,7)^2 \times 9 + (78,5 - 70,7)^2 \times 9 + (91,5 - 70,7)^2 \times 5) = 175,76 \text{ kg}^2$$

Četnostní tabulka (znak X reprezentuje výšku, znak Y reprezentuje váhu)

	$\langle 46, 59 \rangle$ $b_1 = 52,5$	$\langle 59, 72 \rangle$ $b_2 = 65,5$	$\langle 72, 85 \rangle$ $b_3 = 78,5$	$\langle 85, 91 \rangle$ $b_4 = 91,5$	\sum
$\langle 158, 169 \rangle a_1 = 163,5$	6	1	0	1	8
$\langle 169, 180 \rangle a_2 = 174,5$	1	8	3	0	12
$\langle 180, 191 \rangle a_3 = 185,5$	0	0	5	1	6
$\langle 191, 202 \rangle a_4 = 196,5$	0	0	1	3	4
\sum	7	9	9	5	30

$$\bar{x} = 176,7 \text{ cm} \quad \bar{y} = 70,7 \text{ kg} \quad s_x^2 = 116,16 \text{ cm}^2, \quad s_y^2 = 175,76 \text{ kg}^2,$$

$$s_{xy} = 107,72,$$

$$r_{xy} = 0,7539.$$

5

Náhodná procházka

5.1. Definice bílého šumu

Náhodná složka (chybová složka, reziduální složka) je ideálně tvořena jen drobnými nesystematickými vlivy, chybami měření, zaokrouhlováním, = náhodná veličina, většinou předpokládáme bílý šum – white noise, navíc někdy normalitu.

Vlastnosti bílého šumu Náhodná veličina ε je bílý šum, jestliže

- má nulovou střední hodnotu, t.j. $E\varepsilon_t = 0$, $t = 1, \dots, n$,
- $\text{var } \varepsilon_t = \sigma^2$, $t = 1, \dots, n$,
- měření jsou nekorelovaná, t.j. $\text{cov}(\varepsilon_t, \varepsilon_s) = 0$, $t \neq s$.

5.2. Chybová složka – testování náhodnosti

Často se provádí na zvolené hladině významnosti testy nulové hypotézy

$$H_0 : y_t \sim \text{i.i.d.},$$

i.i.d znamená identically independent distribution.

Tato nulová hypotéza připouští konstantní nenulovou střední hodnotu.

Testy tohoto typu označujeme jako testy náhodnosti (tests of randomness) a jde většinou o testy neparametrické.

5.2.1. Test založený na znaménku diferencí

Test je založen na počtu kladných prvních diferencí dané řady, t.j. na počtu bodů, v nichž daná řada roste (tzv. body růstu).

Definujeme náhodnou veličinu V_t předpisem

$$V_t = \begin{cases} 1 & \text{pro } y_t < y_{t+1}, \\ 0 & \text{pro } y_t > y_{t+1}. \end{cases}$$

Střední hodnota počtu k bodů růstu je pak za platnosti nulové hypotézy rovna

$$E(k) = E\left(\sum_{t=1}^{n-1} V_t\right) = \sum_{t=1}^{n-1} \left(\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0\right) = \frac{n-1}{2}.$$

Dále za platnosti nulovosti nulové hypotézy je

$$\text{var}(k) = \sum_{t=1}^{n-1} \sum_{s=1}^{n-1} \text{cov}(V_t, V_s) = \sum_{t=1}^{n-1} \text{var}(V_t) + 2 \sum_{t=1}^{n-1} \text{cov}(V_t, V_s) = \frac{n-1}{4} - \frac{2(n-2)}{12} = \frac{n+1}{12}.$$

Testovou statistikou je

$$U = \frac{k - \frac{n-1}{2}}{\sqrt{\frac{n+1}{12}}} \geq u_{1-\alpha/2}.$$

Test založený na bodech zvratu Nechť r označuje počet horních a dolních bodů zvratu v testované řadě.

Analogicky lze odvodit, že střední hodnota počtu r bodů zvratu je za platnosti nulové hypotézy rovna

$$E(r) = \frac{2(n-2)}{3},$$

$$\text{var}(r) = \frac{16n-29}{90}.$$

Testovou statistikou je

$$U = \frac{r - 2(n-2)/3}{\sqrt{(16n-29)/90}} \geq u_{1-\alpha/2}.$$

5.2.2. Test založený na Spearmanově koeficientu

Nechť q_1, \dots, q_n označuje pořadí hodnot v testované řadě. Spearmanův koeficient pořadové korelace ρ je

$$\rho = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^n (i - q_i)^2.$$

Testovou statistikou je

$$\sqrt{n-1}|\rho| \geq u_{1-\alpha/2}.$$

5.2.3. Mediánový test

Nechť M označuje výběrový medián.

Označme m počet pozorování, které jsou menší než M . Označme u počet pozorování, které nejsou rovny M .

Testovou statistikou je

$$\frac{u - (m+1)}{\sqrt{\frac{m(m-1)}{2m-1}}} \geq u_{1-\alpha/2}.$$

5.2.4. Test založený na Kendallově koeficientu pořadové korelace τ

Definujeme

$$\tau = \frac{4v}{n(n-1)} - 1,$$

kde n označuje počet pozorování a v je počet dvojic y_s a y_t takových, že $y_s < y_t$ pro $s < t$.

Číslo v je při výpočtu τ vhodně normováno, neboť $E(v) = \frac{n(n-1)}{4}$, tedy $\tau \in \langle -1, 1 \rangle$.

Testovou statistikou je

$$\frac{\tau}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} \geq u_{1-\alpha/2}.$$

5.3. Chybová složka – testování autokorelace

Uvažujme model autokorelace

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

kde ρ je korelační koeficient a u_t je bílý šum.

Dále uvažujeme nekorelovanost u_t a ε_{t-1} .

Durbinův–Watsonův test

$$H_0 : \rho = 0, \text{ t.j. } \varepsilon_t \text{ je bílý šum.}$$

Testovou statistikou je

$$DW = \frac{\sum_{t=2}^n (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^n \hat{\varepsilon}_t^2}.$$

Lze také užít aproximaci

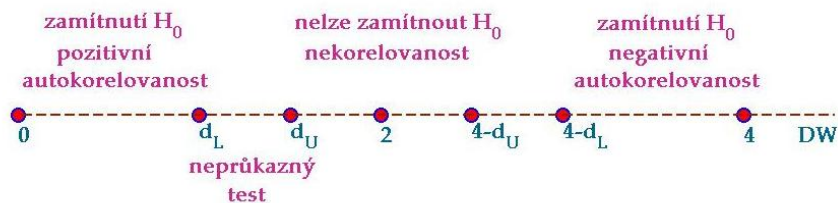
$$DW \sim 2(1 - \hat{\rho}),$$

když odhad korelačního koeficientu je dán vztahem

$$\hat{\rho} = \frac{\sum_{t=2}^n \hat{\varepsilon}_{t-1} \hat{\varepsilon}_t}{\sum_{t=1}^n \hat{\varepsilon}_t^2}.$$

Pro tři význačné hodnoty korelačního koeficientu dostáváme

$$DW = \begin{cases} 4 & \text{pro } \rho = -1, \\ 2 & \text{pro } \rho = 0, \\ 0 & \text{pro } \rho = 1. \end{cases}$$



Lze použít i další testy: Boxův–Pierceův test, Ljungův–Boxův test, Breuschův–Godfreyův test autokorelovanosti reziduí.

6

Index a koš zboží

Index je hospodářský ukazatel, indikátor pokroku nebo neúspěchu. Co indikuje index? Ukazuje nám průběh nějakého vývoje tím, že zaznamenává změny oproti dřívějšímu období. Není to však obyčejná časová řada, nýbrž v jisté míře časová řada koše zboží.

Zjistíme-li kolik stále tuna pšenice v roce 1900 a vedle toho uvedeme její cenu v letošním roce, dostáváme sice časový vývoj určité ceny, nikoliv však cenový index. Nebude tomu však ani v případě, když pro získání snadného přehledu stanovíme cenu za září 1995 jako základní cenu rovnou 100, takže pro následující léta dostáváme 102,4 % nebo 95,4 %, protože stále ještě nejde o index v běžném slova smyslu. Index totiž musí charakterizovat celkovou situaci, nejen situaci jednotlivého zboží. Index nám musí říci, jak se změnili životní náklady, kursy akcií, daňové zatížení apod.

Nejznámějším indexem je index spotřebitelských cen, většinou nazývaný indexem životních nákladů. Index je konstruován na základě spotřebního schematu, které přesně neodpovídá pro žádného spotřebitele.

Proti tomuto indexu se pak často namítá, že se v něm skutečný vzestup životních nákladů zrcadlí jen nedostatečně.

Index musí také reflektovat na změnu spotřebních zvyklostí. Spotřební koš se musí v několikaletých intervalech vždy znovu revidovat.

Dalším problémem je vážení.

Myšlenka měřit snižování hodnoty peněz časově průběžným měřením cen sahá k počátku 18. století. Anglický biskup Fleetwood se v díle *Chronican Preciosum* z r. 1707 zabývá řešením morálního hlediska ztráty kupní síly peněz. Na jedné anglické univerzitě je studentům udělováno stipendium s tou podmínkou, že jeho držitel je musí vrátit, jakmile by jeho majetek přesáhl 5 liber. Vznikla otázka, zda se má dodržet doslovný text zakládací listiny nebo zda je spravedlivé, aby se přihlíželo k poklesu kupní síly peněz, který nastal. Biskup Fleetwood na základě starých dokumentů porovnal ceny a zjistil, že ceny obilí a masa stouply na šestnásobek, nápoje a oděvy na pětinasobek a proto — uzavíral — je prý možno stipendistovi přiznat majetek 28 liber, aniž by tím byl změněn úmysl zakladatele stipendia.

O 30 let později se se Francouz Dutot zabýval otázkou, zda Ludvík XV. se svým příjmem 100 milionů franků byl na tom lépe než Ludvík XII. o dvě stě let dříve s necelými 8 milióny. Také on sestavil pro porovnání seznam zboží, který obsahoval mimo jiné i kozu, holuba, náklad sena, mzdu služby a nádeníka. Na tomto základě určil znehodnocení peněz na jednu dvaadvacetinu a usoudil, že Ludvík XV. na tom byl hůře.

Známým je propočet indexu od italského finančníka Gianrinaldo Carlina z r 1764. Jeho spotřební koš se skládal z vína, pšenice a oleje. Litř vína stál v r. 1500 10 lir, o 20 let později 15 lir, kilogram pšenice stál v r. 1500 8 lir, o 20 let později 14 lir, 1 litř oleje stál nejdříve 60 lir a později 80 lir.

Z toho vyplynul propočet indexu

$$\frac{\frac{15}{10} + \frac{14}{8} + \frac{80}{60}}{3} = 1,525.$$

Při základním roku 1500 by se měl jeho index zvýšit ze 100 na 152,5.

Přitom se však vůbec nepřihlíželo k faktoru vážení. Pro výpočet indexu se vůbec nevyužíval jakýkoliv údaj o zvyklostech spotřeby. Pouze pokud by cenový vývoj u všech druhů

zboží probíhal stejně by nebylo potřeba vážení. Dalším faktorem, který nebyl uvažován, je skutečnost že se mohou měnit spotřební zvyklosti.

Moderní index můžeme demonstrovat na Carliho metodě. Zvolíme základní rok 1500 a budeme uvažovat, že ve spotřebním koši je 5 litrů vína, 8 kg pšenice a 2 litry oleje. Náklady tohoto spotřebního koše v r. 1500 jsou $5 \times 10 + 8 \times 8 + 2 \times 60 = 234$ lir.

Pokud v r. 1501 podražilo víno o 1 liru a ostatní 2 položky se nezměnily, má náš spotřební koš hodnotu 239 lir. V daném případě se index roku 1501 rovná $239/234 = 102,1$. Lze říci, že se životní náklady v roce 1501 oproti r. 1500 zvýšili o 2,1%.

Pokud v r. 1502 podražila pšenice a stojí 9 lir místo 8 lir, stojí náš spotřební koš 274 lir a index má hodnotu 105,5.

Ve srovnání se základním rokem je situace jasná: vzestup o 5,5%.

Proti r. 1501 je vzestup třeba určit jako podíl $247/239$, resp. $105,5/102,1$.

Tento výpočet má ale nedostatek. Můžeme např. vědět, že ke zvýšení ceny pšenice došlo v důsledku neúrody. Současně s růstem ceny ale došlo i k poklesu odbytu a nahrazení pšenice jinými komplementárními plodinami — proto je třeba změnit spotřební schema.

K výpočtu po uvážení množství zboží lze použít Laspayresův index

$$I_L = \frac{\text{nové ceny} \times \text{stará množství}}{\text{staré ceny} \times \text{stará množství}}$$

nebo Paascheho index

$$I_P = \frac{\text{nové ceny} \times \text{nová množství}}{\text{staré ceny} \times \text{nová množství}}.$$

Předpokládejme, že se v roce 1501 spotřební schema skládalo z 4 litrů vína (místo 5), 10 kg pšenice (místo 8) a 2 litrů oleje (beze změny). Spočtěme nyní hodnoty uvedených indexů, pokud v r. 1510 stojí víno 12 lir, pšenice 7 lir a olej 65 lir.

$$I_L = \frac{60 + 56 + 130}{50 + 64 + 120} = 105,1.$$

$$I_P = \frac{48 + 70 + 130}{40 + 80 + 120} = 103,3.$$

Vidíme, že se výsledky značně liší. Pokud chce opozice ukázat, že ceny drasticky stouply zvolí metodu Laspeyresovu. Pokud budou ekonomové vládní strany chtít ukázat pouze pozvolný růst cen, použijí index Paascheho.

Kromě těchto indexů se lze setkat i s výpočtem podle vztahu

$$I = \frac{\text{nové ceny} \times \text{nová množství}}{\text{staré ceny} \times \text{stará množství}},$$

kdy získáme hodnotu

$$I = \frac{48 + 70 + 130}{50 + 64 + 120} = 105,9.$$

Použití tohoto indexu je ale nesmyslné, neboť nezohledňuje větší požadavky spotřebitelů.

Americký nárohhospodář Irving Fischer navrhl spojení Laspeyresova a Paascheho indexu do ideálního Fisherova indexu:

$$I_F = \sqrt{I_P I_L}.$$

V našem případě

$$I_F = 104,20.$$

Literatura

Cihelský, L.: Úvod do teorie popisné statistiky, SNTL, Praha, 1974.

Swoboda, H.: *Moderní statistika*. Nakladatelství Svoboda, Praha, 1977.

7

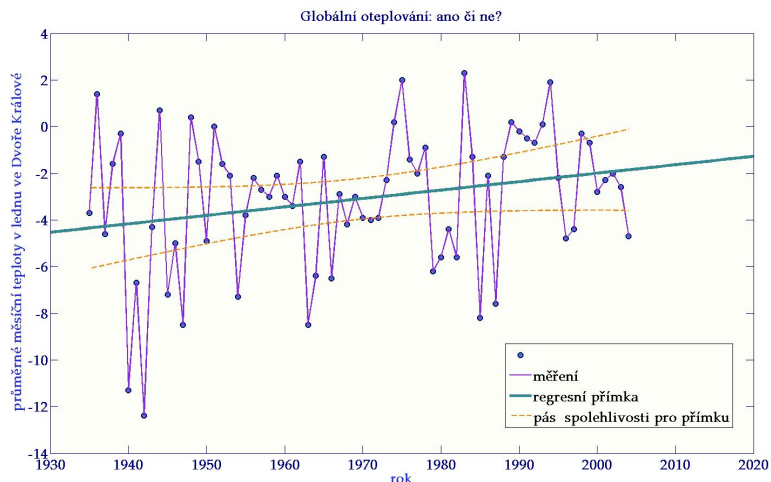
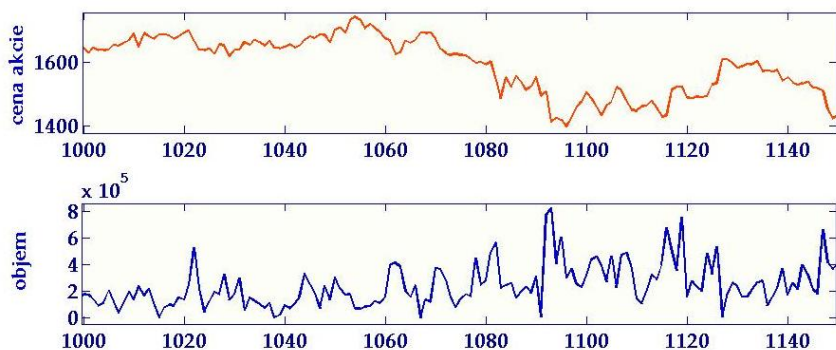
Časové řady

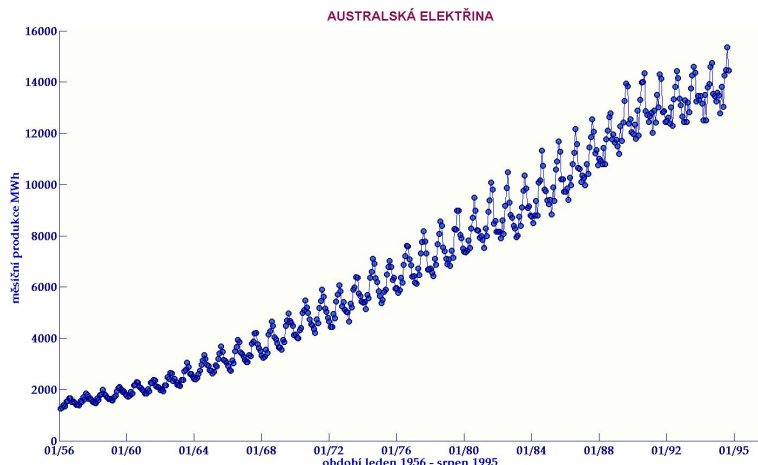
7.1. Úvod do časových řad a historie

Cílem této kapitoly je seznámit laskavého čtenáře s vybranými partie z teorie časových řad, pomoci mu získat nadhled a usnadnit mu bližší přístup k teoretičtějším publikacím.

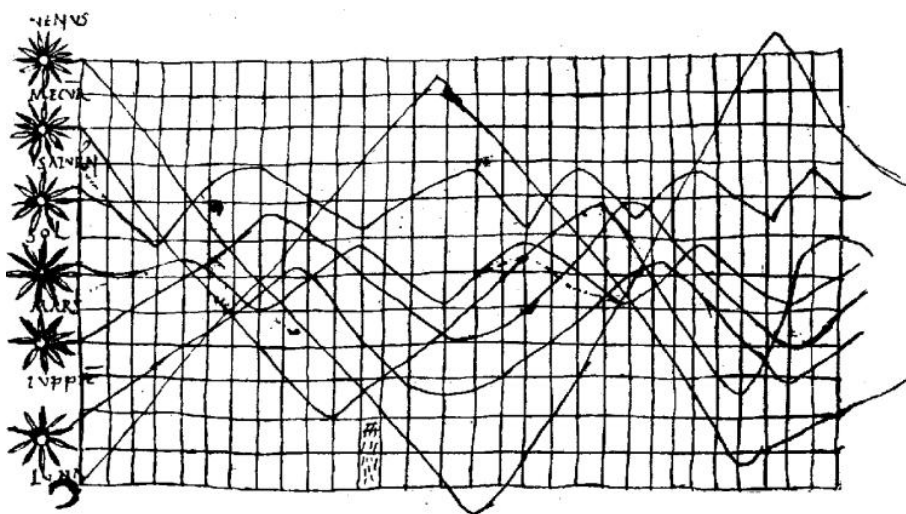
Analýza časových řad a predikce budoucího vývoje je důležitou statistickou disciplínou. Data vytvářející časovou řadu vznikají jako chronologicky uspořádaná pozorování. K práci s časovými řadami vede např. zpracování demografických údajů, hospodářských údajů, fyzikálních údajů, technických dat, údajů v medicíně, ekologii, atd.

Následující obrázky obsahují časové řady cen a objemů akcí v obchodních dnech, průměrných ročních teplot, spotřeby elektřiny.





K nejstarším záznamům můžeme řadit astronomická pozorování, viz diagram poloh planet neznámého autora (pořadí: Venuše, Merkur, Saturn, Slunce!, Mars, Jupiter, Měsíc) kolem roku 950.



Statistická grafika vzniká ve snaze zhustit číselné informace do snadněji vnímatelné formy různých grafů.

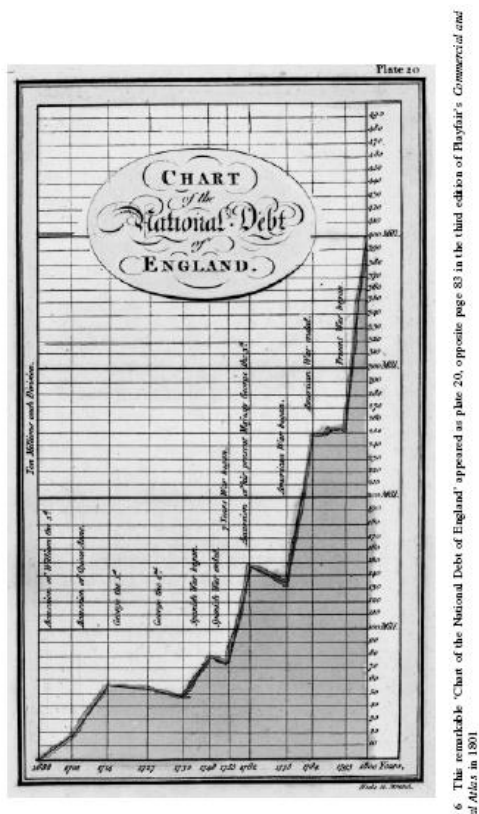
Skotský filosof Dugald Stewart ve stati *A general View of the Progress of Metaphysical, Ethical, and Political Philosophy since the Revival of Letters* (1811) napsané pro *Dodatky k Britské encyklopedii* konstatuje^{*)}, že historie, jako znalost určitých faktů a dějů, je především záležitostí naší paměti, která je subjektivní. Historické děje (a s nimi také ekonomické, populační aj.) sice mohou být a nejspíš jsou podřízeny nějakým zákonům, ty však nelze zjistit pozorováním jako zákony přírodní, ale pouze reflexí, uvažováním. Speciálně ekonomický stav státu je důsledkem subjektivního jednání lidí v jejich soukromých životech; to může probíhat např. na základě „zdravého rozumu“.

A jeho myšlenky jsou pro teorii časových řad podstatné — chceme najít jisté zákony, které ovlivňují zkoumané děje. Cílem analýzy časových řad je zejména konstrukce vhodného

^{*)} překlad viz Saxl, I., Ilucová, L.: *Historie grafického zobrazování statistických dat*, Robust, Praha, 2004, str. 363 – 384

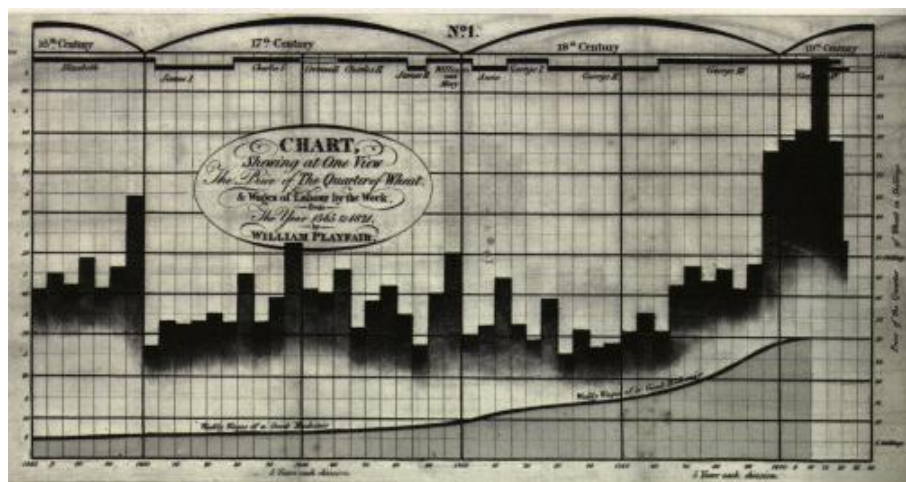
modelu, který popisuje sledovaný jev. Známe-li správný model, můžeme sestavit algoritmus, který umí predikovat budoucí vývoj.

Z historických ekonomických řad je známý graf růstu britského národního dluhu v letech 1699 až 1800, grafy vzájemného obchodu mezi Anglií a různými státy (např. s Německem, s Dánskem a Norskem, histogram zahraničního obchodu Skotska aj. od Williama Playfaira.



: 6. The remarkable 'Chart of the National Debt of England' appeared as plate 20, opposite page 83 in the third edition of Raynaud's 'Commercial and Political Atlas' in 1801.

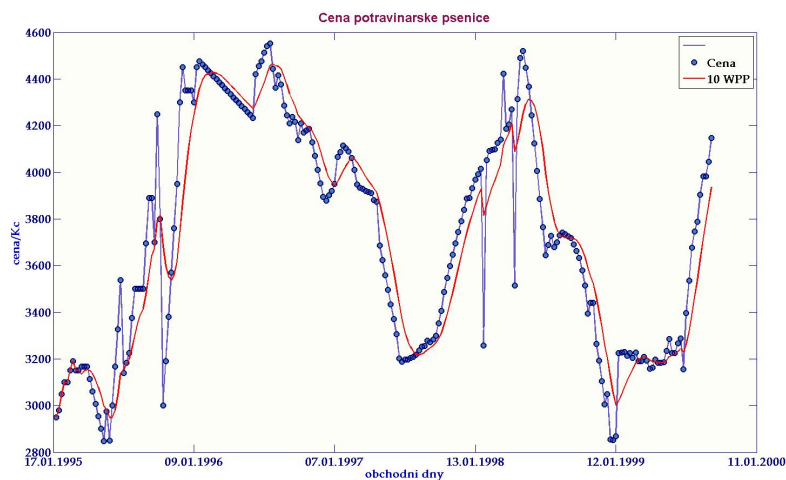
K nejznámějším patří mimořádně sugestivní graf porovnávající ceny pšenice a mzdy řemeslníků na pozadí vlád jednotlivých britských panovníků v letech 1665 až 1821.



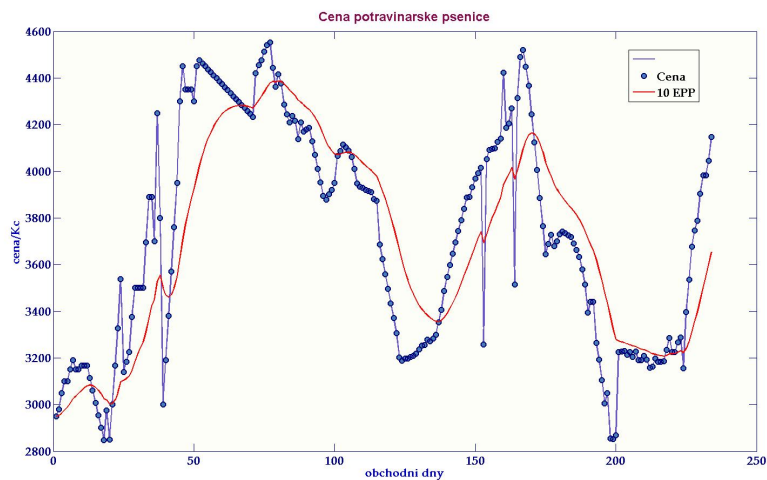
Pozoruhodné je, že právě tento graf na první pohled nesděluje autorův záměr a hrozivě rostoucí černý histogram (termín histogram však zavedl až K. Pearsons) mu spíše protiřečí. Playfairůvou snahou bylo totiž podle jeho vlastního vyjádření ukázat, že nikdy nebyla pšenice

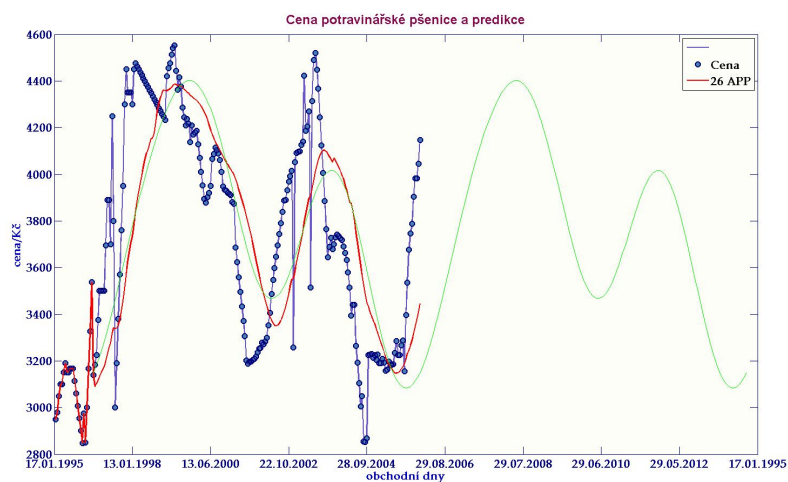
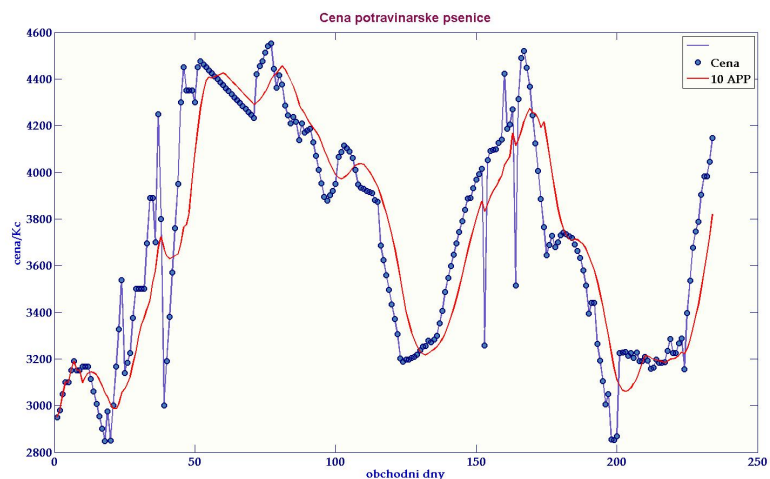
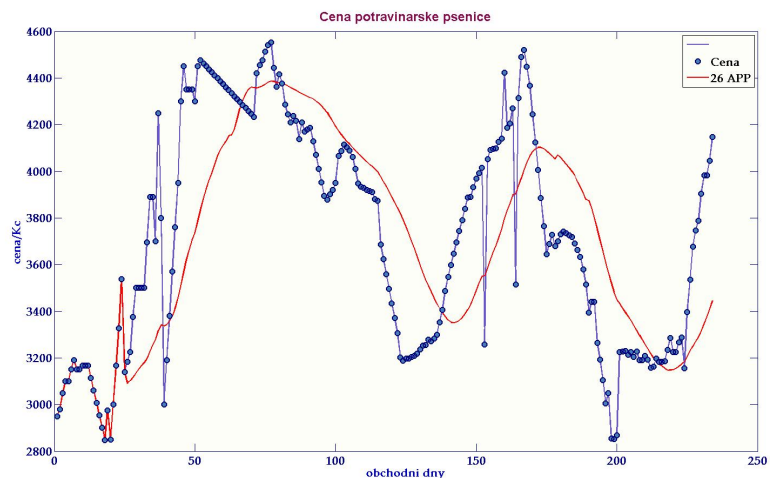
tak levná jako na počátku 19. století. To je však patrné teprve tehdy, když je vynesena graf poměru cen a mezd, který skutečně klesá od devíti ke dvěma. Playfairův graf tak ukazuje jednu z charakteristických vlastností grafického zobrazení, totiž na možnost vytvoření dojmu na první pohled opačného, než odpovídá skutečnému obsahu dat.

Grafické znázornění časových řad v dnešní podobě se objevilo až na počátku 19. století, jako záznamy zemědělské produkce (známá Beveridgeova řada popisuje cenový index pšenice v západní Evropě od r. 1500 do r. 1869).



Na konci 19. století a na začátku 20 století je při analýze časových řad věnována velká pozornost hledání periodických složek. Tento přístup má ale spoustu nevýhod.





Později se prokázalo, že pomocí modelů využívajících bílý šum (autoregresních modelů nebo modelů klouzavých součtů) lze modelovat skutečné řady z ekonomie, které se často vyznačují cyklickou složkou s proměnnou amplitudou a proměnnou vzdáleností mezi body zvratu.

Další možností pro analýzu časových řad je Boxův–Jenkinsův přístup, který zde studovat nebudeme.

7.2. Pojem a druhy časových řad

Časová řada je posloupnost věcně a prostorově srovnatelných porovnání, která jsou časově uspořádána ve směru minulost — přítomnost.

Poznámka. Pozorováním rozumíme měření, údaje, hodnoty, data atd.

Místo pojmu časová řada se můžeme setkat s termíny dynamická, chronologická, resp. vývojová řada.

Pozorování se vztahují k časovým okamžikům t_1, t_2, \dots, t_n , které vytváří body na časové ose.

Budeme užívat následujícího značení: y_{t_i} je pozorování uskutečněné v časovém okamžiku t_i , t_i je časový index.

7.2.1. Dělení časových řad

(1) neekvidistantní

délka intervalů $t_i - t_{i-1}$ není stejná

U neekvidistantní časové řady použijeme indexování

$y_{t_1}, y_{t_2}, \dots, y_{t_n};$

$y_{t_i}, i = 1, \dots, n;$

$\{y_{t_i}\}_{i=1}^n.$

(2) ekvidistantní

délka intervalů $t_i - t_{i-1}$ je stejná

U ekvidistantní časové řady použijeme indexování

$y_1, y_2, \dots, y_n;$

$y_t, t = 1, \dots, n;$

$\{y_t\}_{t=1}^n.$

Poznámka. Přirozené číslo n nazýváme délkou řady.

Někdy u časových řad užíváme označení denní, týdenní, měsíční, roční — jedná se o vymezení četnosti (frekvence) pozorování a s číslem n nesouvisí.

7.2.2. Druhy časových řad

(1) intervalová (časová řada intervalových ukazatelů)

vzniká akumulací (agregací) hodnot za dané časové období (úsek, interval). Hodnoty udávají kolik čeho vzniklo nebo zaniklo. Hodnoty závisí na délce období. Může se jednat např. o množství produkce, tržby, srážky, emise. Při nestejně dlouhých obdobích nejsou hodnoty vzájemně srovnatelné. Je třeba je očistit od kalendářních variací, tedy přepočítat na jednotkové období. Například měsíční data přepočítáme pomocí vztahu

$$y_t^0 = y_t \frac{\bar{k}}{k_t}, \quad (1)$$

kde \bar{k} je průměrný počet dní v měsíci a k_t je počet dní v měsíci.

Hodnoty se většinou přiřazují středu intervalu.

Můžeme rozlišovat intervalovou časovou řadu běžných hodnot a kumulovaných (úhrnných) hodnot.

(2) okamžiková

Data mohou být diskrétní nebo mohou vzniknout diskretizací spojité veličiny. Udává, kolik čeho je, resp. existuje, např. teplota, počet pracovníků. Je důležité si uvědomit, že součty nedávají smysl.

Můžeme se také setkat s následujícími pojmy (časovými řadami)

(3) extenzivních ukazatelů

(4) odvozených charakteristik

(5) dlouhodobá roční

(6) krátkodobá měsíční

(7) naturálních ukazatelů

Poznámka. Nabízí se otázka — jak často pozorovat? Mělo by to být přiměřeně často a ve stejných intervalech.

Na měření (způsob zjišťování) jsou kladeny jisté požadavky:

(1) srovnatelnost

– věčná

– prostorová (např. geografický prostor – kraj, ekonomický prostor – typ spotřebitele)

(2) časové

(3) cenové

– běžné (aktuální)

– stálé (fixované k určitému datu)

Práce s časovými řadami zahrnuje analýzu – popis, pochopení generujícího mechanismu (tedy způsob vzniku dat). Matematickým cílem je sestavit model, který je jednoduchý a dostatečně věrný. Tento model slouží k prognóze, předvídání budoucího vývoje. Při hledání modelu je také třeba vhodně volit vstupní proměnné a počáteční parametry, tak aby výstup byl optimální.

7.3. Základní charakteristiky časových řad

7.3.1. Popisné charakteristiky

Při práci s časovými řadami je někdy důležité zjistit jejich průměrné hodnoty. Průměrná hodnota intervalové časové řady se vypočítá pomocí *prostého aritmetického průměru*

$$\bar{y} = \sum_{t=1}^T \frac{y_t}{T}. \quad (2)$$

Průměrná hodnota okamžikové časové řady y_t , $t = 1, \dots, T$ se počítá pomocí *chronologického průměru*. Při stejné vzdálenosti mezi jednotlivými okamžiky sledování se používá *prostý chronologický průměr*

$$\bar{y} = \frac{\frac{1}{2}y_1 + \sum_{t=2}^{T-1} y_t + \frac{1}{2}y_T}{T-1}. \quad (3)$$

Při různé vzdálenosti jednotlivých okamžiků sledování se používá *vážený chronologický průměr*

$$\bar{y} = \frac{\frac{y_1+y_2}{2}d_2 + \frac{y_2+y_3}{2}d_3 + \dots + \frac{y_{T-1}+y_T}{2}d_T}{d_2+d_3+\dots+d_T}. \quad (4)$$

kde d_t , $t = 2, \dots, T$ je délka jednotlivých časových intervalů sledování daného okamžikového ukazatele.

Příklad 7.1

Vypočítejte průměrnou hodnotu roční časové řady hrubého domácího produktu České republiky na jednoho obyvatele v Kč v letech 1999 - 2008:

1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
202357	213110	230064	241593	252617	275770	291561	313868	342494	354410

Časová řada hrubého domácího produktu České republiky na jednoho obyvatele v letech 1999 - 2008 je intervalovou časovou řadou, proto se její průměrná hodnota vypočítá pomocí aritmetického průměru

$$\bar{y} = \sum_{t=1}^T \frac{y_t}{T} = 271780. \quad (5)$$

Průměrná hodnota časové řady v letech 1999 až 2008 je 271780 Kč.

Příklad 7.2 Vypočítejte průměrný počet registrovaných uchazečů o zaměstnání v České republice v druhé polovině roku 2007 v tis. osob.

datum	počet	délka období
31. 6. 2007	370791	
31. 7. 2007	376608	31
31. 8. 2007	372759	31
30. 9. 2007	364978	30
31. 10. 2007	348842	31
30. 11. 2007	341438	30
31.12. 2007	354878	31

Počet registrovaných uchazečů o zaměstnání v České republice v jednotlivých měsících roku 2007 je okamžitá časová řada, protože rozhodným okamžikem sledování je vždy poslední den daného měsíce. Protože vzdálenost mezi jednotlivými okamžiky sledování není stejná, použijeme pro výpočet průměrné hodnoty vážený chronologický průměr

$$\bar{y} = \frac{\frac{370791+376608}{2}31 + \frac{376608+372759}{2}31 + \dots + \frac{341438+354878}{2}31}{31 + 31 + 30 + 31 + 30 + 31} = \frac{66477236}{184} = 361290.$$

Průměrný počet registrovaných uchazečů o zaměstnání v České republice v druhé polovině roku 2007 byl 361,3 tis. osob.

7.3.2. Míry dynamiky

Jednoduché míry dynamiky časových řad umožňují charakterizovat základní rysy „chování“ časových řad a formulovat jistá kritéria pro jejich modelování. Předpokládejme časovou řadu y_t , $t = 1, \dots, T$. Nejjednodušší mírou dynamiky je absolutní přírůstek (první diference), který lze zapsat jako

$$\Delta y_t = y_t - y_{t-1}, t = 2, \dots, T. \quad (6)$$

Tato charakteristika vyjadřuje změnu hodnoty v čase t proti času $t - 1$.

Často se používá také průměrný absolutní přírůstek

$$\bar{\Delta} = \frac{(y_2 - y_1) + (y_3 - y_2) + \dots + (y_T - y_{T-1})}{T - 1} = \frac{\sum_{t=2}^T \Delta y_t}{T - 1} = \frac{y_T - y_1}{T - 1} \quad (7)$$

Diferencováním první difference lze získat druhou diferenci, diferencováním druhé difference dostaneme diferenci třetí. Diferencování má v analýze časových řad velký význam. Používá se při modelování trendu časových řad k výběru vhodné trendové funkce, nezastupitelná je jeho role při stochastickém modelování časových řad.

trend	informativní test
lineární	první difference $y_{t+1} - y_t$ jsou přibližně konstantní
kvadratický	druhé difference $y_{t+2} - 2y_{t+1} + y_t$ jsou přibližně konstantní
exponenciální	podíly sousedních hodnot y_{t+1}/y_t jsou přibližně konstantní
logistický	křivka prvních diferencí $y_{t+1} - y_t$ se podobá křivce hustoty normálního rozdělení
Gompertzova křivka	podíly $(\log y_{t+2} - \log y_{t+1})/(\log y_{t+1} - \log y_t)$ jsou přibližně konstantní

Velmi důležitou mírou dynamiky časových řad je *koeficient růstu*.

$$k_t = \frac{y_t}{y_{t-1}}, \quad t = 2, \dots, T$$

Jestliže se tento koeficient vynásobí stem, udává na kolik procent hodnoty v čase $t-1$ vzrostla hodnota v čase t .

Někdy se pro tento koeficient používá název *tempo růstu*.

Průměrný koeficient růstu (průměrné tempo růstu) se vypočítá jako geometrický průměr jednotlivých koeficientů růstu

$$\bar{k} = \sqrt[T]{k_2 k_3 \dots k_T} = \sqrt[T]{\frac{y_T}{y_1}}$$

Meziroční koeficient růstu je podíl hodnot časové řady ve stejných obdobích (sezónách) v po sobě jdoucích letech. V případě čtvrtletní časové řady má tvar

$$k_t = \frac{y_t}{y_{t-4}}, \quad t = 5, 6, \dots, T$$

lze jej vyjádřit také jako součin (čtvrtletních) koeficientů růstu

$$k_{4,t} = \frac{y_t}{y_{t-1}} \frac{y_{t-1}}{y_{t-2}} \frac{y_{t-2}}{y_{t-3}} \frac{y_{t-3}}{y_{t-4}}.$$

Další mírou dynamiky časových řad je relativní přírůstek

$$\delta_t = \frac{\Delta y_t}{y_{t-1}} = \frac{y_t - y_{t-1}}{y_{t-1}} = \frac{y_t}{y_{t-1}} - 1,$$

po vynásobení stem nám říká o kolik procent se změnila hodnota časové řady v čase t ve srovnání s časem $t-1$.

Průměrný relativní přírůstek se vypočítá jako

$$\bar{\delta} = \bar{k} - 1.$$

Další mírou dynamiky časových řad je tempo se stálým růstem

$$\frac{y_t}{y_1}.$$

Jinou charakteristikou je koeficient zrychlení

$$\phi_{y_t} = \frac{\Delta^{(2)} y_t}{\Delta y_{t-1}} = \frac{\Delta y_t - \Delta y_{t-1}}{\Delta y_{t-1}}, \quad t = 3, \dots, T. \quad (8)$$

u neekvidistantní řady

$$\phi_{y_{t_i}} = \frac{\Delta y_{t_i} - \Delta y_{t_{i-1}}}{t_i - t_{i-1}} \frac{1}{\Delta y_{t_{i-1}}}, t = 3, \dots, T. \quad (9)$$

Těžba uhlí 1994 (ve 100 tunách měsíčně)

t	y_t	y_t^0	1. difference	2. difference	koef.	koef.	koef. se	klouzavý
			1. difference	2. difference	zrychlení	růstu	stalým zákl.	úhrn délky 3
1	472	456.8	-	-	-	-	1	-
2	401	429.6	-27.2	-	-0.397	1.063	0.94	-
3	427	413.2	-16.4	-10.8	-1.171	1.040	0.961	1299.6
4	416	416	+2.8	19.2	-16.17	0.993	1.006	1558.4
5	386	373.5	-42.5	-45.3	-0.212	1.113	0.898	1202.7
6	340	340	-33.5	9	-0.209	0.91	0.91	1129.5
7	324	313.5	-26.5	7	-0.460	1.08	0.922	1027
8	284	274.8	-38.7	-12.2		1.141	0.876	928.3

Pro roční časovou řadu reálného hrubého domácího produktu České republiky v letech 1994 – 2000 v mld. Kč vypočítejte základní míry dynamiky.

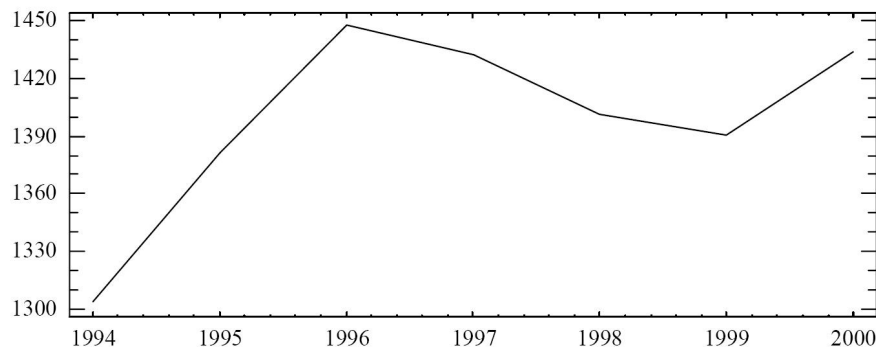
rok	HDP	Δy_t	k_t	δ_t
1994	1303,6	-	-	-
1995	1381,1	77,5	1,059	0,059
1996	1447,7	66,6	1,048	0,048
1997	1432,8	-14,9	0,990	-0,010
1998	1401,3	-31,5	0,978	-0,022
1999	1390,6	-10,7	0,992	-0,008
2000	1433,8	43,2	1,031	0,031

Průměrný absolutní přírůstek je

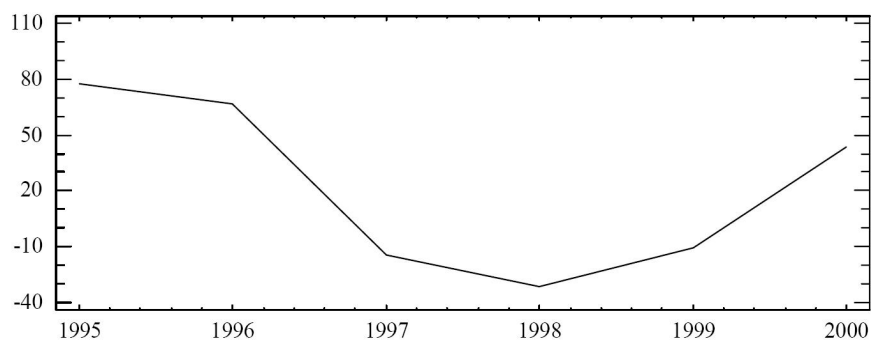
$$\overline{\Delta} = \frac{1433,8 - 1303,6}{6} = 21,7 \text{ mld. Kč,}$$

průměrný koeficient růstu

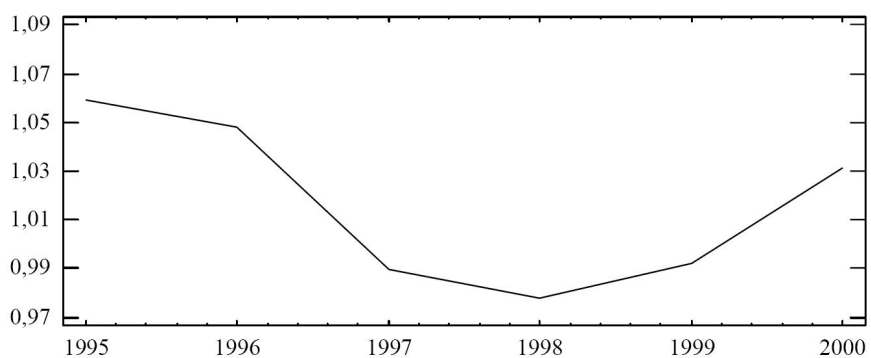
$$\bar{k} = \sqrt[6]{\frac{1433,8}{1303,6}} = 1,016.$$



Obr. : časová řada HPD ČR v letech 1994 - 2000 v mld. Kč



Obr. : absolutní přírůstky (1. difference) časové řady HPD ČR v letech 1994 – 2000 v mld. Kč



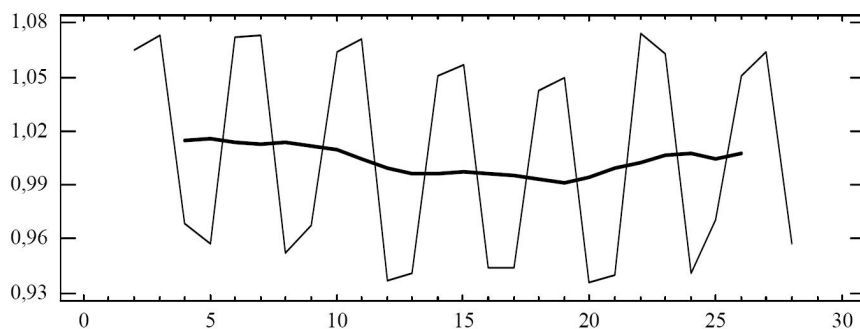
Obr. : koeficienty růstu časové řady HPD ČR v letech 1994 – 2000 v mld. Kč

Na základě čtvrtletní časové řady reálného hrubého domácího produktu ČR od I. čtvrtletí roku 1994 do IV. čtvrtletí roku 2000 v mld. Kč vypočítejte koeficienty růstu (čtvrtletní), meziroční koeficienty růstu a centrované klouzavé geometrické průměry délky 4 počítané z koeficientů růstu.

rok	kvartál	HDP	koef. růstu (čtvrťl.)	meziroční koef. růstu	centrované klouzavé geometrické průměry
			$k_t = \frac{y_t}{y_{t-1}}$	$k_{(4),t} = \frac{y_t}{y_{t-4}}$	$(\sqrt[4]{k_{(4),t}} \sqrt[4]{k_{(4),t-1}})^{1/2}$
1994	I.	302,2			
	II.	321,8	1,065		
	III.	345,2	1,073		
	IV.	334,4	0,969	1,015	
1995	I.	319,9	0,957	1,059	1,016
	II.	343,0	1,072	1,066	1,014
	III.	367,9	1,073	1,066	1,013
	IV.	350,3	0,952	1,048	1,014
1996	I.	339,1	0,968	1,060	1,012
	II.	360,7	1,064	1,052	1,010
	III.	386,2	1,071	1,050	1,004
	IV.	361,7	0,937	1,033	0,999
1997	I.	340,4	0,941	1,004	0,996
	II.	357,6	1,051	0,991	0,996
	III.	378,0	1,057	0,979	0,997
	IV.	356,8	0,944	0,986	0,996
1998	I.	336,8	0,944	0,989	0,995
	II.	351,0	1,042	0,982	0,993
	III.	368,6	1,050	0,975	0,991
	IV.	344,9	0,936	0,967	0,994
1999	I.	324,2	0,940	0,963	0,999
	II.	348,1	1,074	0,992	1,002
	III.	370,0	1,063	1,004	1,007
	IV.	348,3	0,941	1,010	1,008
2000	I.	338,2	0,971	1,043	1,005
	II.	355,5	1,051	1,021	1,008
	III.	378,2	1,064	1,022	
	IV.	361,9	0,957	1,039	

Koeficienty růstu (čtvrtletní) a centrované klouzavé geometrické průměry délky 4 — HPD ČR

Z hlediska symboliky koeficientů růstu je logické, že první hodnota časové řady meziročních koeficientů růstu je umístěna do 1. čtvrtletí druhého roku, tj. roku 1995. Vzhledem k tomu, že po odmocnění čtvrtou odmocninou lze tuto řadu chápat jako řadu klouzavých geometrických průměrů koeficientů růstu, je vhodné umístit její první hodnotu mezi třetí a čtvrté čtvrtletí roku 1994. Centrováním geometrickým průměrem sousedních dvojic hodnot se první hodnota centrovaného klouzavého geometrického průměru dostane do čtvrtého čtvrtletí roku 1994 a poslední hodnota do druhého čtvrtletí roku 2000.



Obr. : Koeficienty růstu (čtvrtletní) a centrované klouzavé geometrické průměry délky 4 — HPD ČR

7.4. Dekompozice časové řady

Dekompozice může být aditivní

$$y = T + S + C + E$$

nebo multiplikativní

$$y = T \cdot S \cdot C \cdot E$$

T ... trend, hlavní tendence dlouhodobého vývoje, popis jednoduchou funkcí.

S ... sezónnost, zachycuje pravidelné odchylky od trendu s periodou ne delší než jeden rok, odchylky souvisí s kalendářem a střídáním ročních období (průběh teploty, tržby, std.)

C ... cyklus, dlouhodobé kolísání kolem trendu s proměnlivou periodou a rozpětím (demografický cyklus, hospodářský cyklus).

E ... náhodná složka (chybová složka, reziduální složka), co zbyde, ideálně drobné nesystematické vlivy, chyby měření, zaokrouhlování, = náhodná veličina, většinou předpokládáme bílý šum – white noise, navíc někdy normalitu.

náhodná veličina ε je bílý šum, jestliže

- má nulovou střední hodnotu, t.j. $E\varepsilon_t = 0$, $t = 1, \dots, n$,
- $\text{var } \varepsilon_t = \sigma^2$, $t = 1, \dots, n$,
- měření jsou nekorelovaná, t.j. $\text{cov}(\varepsilon_t, \varepsilon_s) = 0$, $t \neq s$.

7.4.1. Modelování trendu a sezónní složky

Nejčastěji používané trendové funkce jsou:

- konstantní
- lineární
- kvadratická (parabolická)
- kubická
- exponenciální
- posunutá exponenciální
- logistická
- Gompertzova

U prvních 4 uvedených funkcí lze odhad získat pomocí MNČ. Získáváme BLUE. Odvození, viz přednáška, resp. kurz matematické statistiky – kapitola Regrese.

Funkci exponenciální lze transformovat pomocí logaritmické transformace a použít MNČ, avšak získaný odhad je vychýlený.

U posledních 3 funkcí nelze použít MNČ.

Popis trendové složky pomocí exponenciální funkce

$$y_t = \alpha \cdot \gamma^t, \quad \gamma > 0, t = 1, \dots, n.$$

Nelze přímo použít MNČ, ale lze linearizovat logaritmickou transformací:

$$\ln y_t = \ln \alpha + \ln \gamma \cdot t = a_0 + a_1 \cdot t.$$

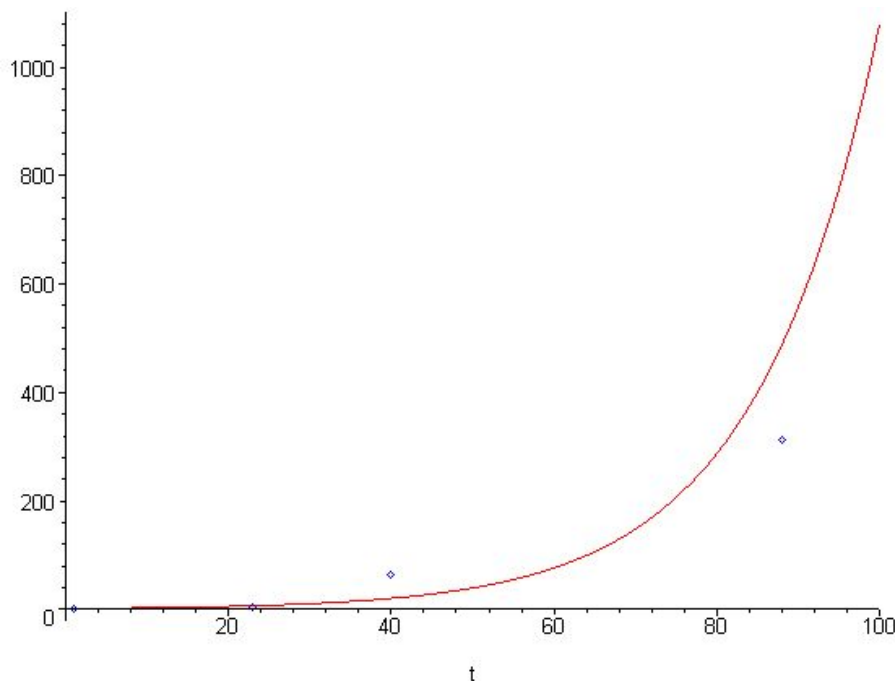
Z původního modelu jsme přešli k modelu s lineárním trendem a pomocí metody nejmenších čtverců najdeme odhady \hat{a}_0 a \hat{a}_1 a zpětnou transformací získáme odhady $\hat{\alpha} = e^{\hat{a}_0}$, $\hat{\gamma} = e^{\hat{a}_1}$.

Je také možné použít váženou metodu nejmenších čtverců (předmětem minimalizace jsou měření, kterým jsou přiděleny různé váhy)

$$\min_{a_0, a_1} \sum_t w_t (\ln y_t - a_0 - a_1 \cdot t)^2, \quad \text{kde } w_t = y_t^2.$$

chřipka A/H1N1

datum	25.5.09	17.6.09	27.7.09	23.10.09
x	1	23	40	88
počet ČR	1	5	63	314



Obr. : chřipka A/H1N1

Posunutý exponenciální trend

$$y_t = \delta + \alpha \cdot \gamma^t, \quad \gamma > 0, t = 1, \dots, n.$$

Tuto funkci nelze lehce převést na lineární regresní model.

Je možné fixovat (vhodně zvolit) hodnotu γ a položit $f(t) = \gamma^t$.

Pak lze trend psát jako lineární

$$T_t = \delta + \alpha \cdot f(t).$$

Použijeme MNČ a získáme odhady $\hat{\delta}$ a $\hat{\alpha}$.

Volbu γ lze upravit a pro jednotlivé volby najdeme odhady $\hat{\delta}$ a $\hat{\alpha}$ a co je podstatné — určíme reziduální součet čtverců. Vybereme γ s nejmenším součtem čtverců.

Logistický trend

$$y_t = \frac{\delta}{1 + \alpha \cdot \gamma^t}, \quad \gamma > 0, t = 1, \dots, n.$$

Tuto funkci nelze lehce převést na lineární regresní model.

Je ale možné provést inverzní transformaci

$$\frac{1}{T_t} = \frac{1 + \alpha \cdot \gamma^t}{\delta} = \beta_0 + \beta_1 \cdot \gamma^t,$$

kterou přejdeme na předchozí posunutý exponenciální trend.

Gompertzův trend

$$y_t = \delta \cdot \alpha^{\gamma^t}, \quad \gamma > 0, t = 1, \dots, n.$$

Tuto funkci nelze lehce převést na lineární regresní model.

Je ale možné provést logaritmickou transformaci

$$\ln T_t = \ln \delta + \ln \alpha \cdot \gamma^t = \beta_0 + \beta_1 \cdot \gamma^t,$$

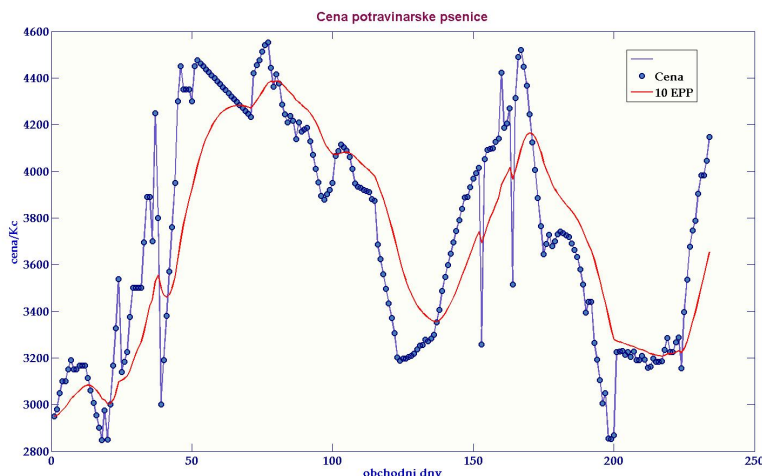
kterou přejdeme opět na posunutý exponenciální trend.

7.5. Klouzavé průměry a exponenciální vyhlazování

Klasická analýza časových řad předpokládá, že trendová funkce má v čase konstantní parametry. V delším časovém období je tento předpoklad nereálný, proto je vhodné využívat adaptivní techniky, jako jsou metoda klouzavých průměrů a exponenciální vyrovňování.

7.5.1. Klouzavé průměry

Metoda klouzavých průměrů se zakládá na myšlence, že časovou řadu y_t pro $t = 1, 2, \dots, T$ rozdělíme na kratší časové úseky, na kterých odhadujeme lokální polynomické trendy určitého stupně. Konstantní trend se popisuje polynomem nultého stupně, lineární trend polynomem prvního stupně apod.



Obr. : příklad vyhlazení

Postup je následující - nechť první část řady má $2m + 1$ hodnot, které označujeme $y_1, y_2, \dots, y_{2m+1}$, z nich odhadneme parametry lokálního trendu vhodným polynomem a vypočítáme jeho odhad \hat{T}_{m+1} , stejný polynom odhadneme na druhé skupině hodnot řady, $y_2, y_3, \dots, y_{2m+2}$ a vypočítáme odhad lokálního trendu \hat{T}_{m+2} , tímto klouzavým způsobem postupujeme až do konce časové řady.

Algoritmus:

- (1) zvolíme délku okna, ozn. m
- (2) daty v okně proložíme nějakou křivku — trendovou funkci
- (3) spočítáme vyrovnanou hodnotu pro střed okna
- (4) posuneme okno o jedno pozorování doprava,
- (5) pokud nejsme na konci řady, návrat na krok 2

Použijeme označení (t, i) , kde t je střed okna, i je pozice pozorování vůči středu okna

Předpokládejme, že řadu y_t pro $t = 1, 2, \dots, T$ budeme vyrovnávat na klouzavých úsecích polynomičtým trendem stupně r s časovou proměnnou i .

Odhady parametrů lokálního trendu (v rámci okna) $T_{t,i} = a_{0t} + a_{1t}i + a_{2t}i^2 + \dots + a_{rt}i^r$ získáme metodou nejmenších čtverců:

Pozn.: Střed okna má vždy index 0 a pro odhad – vyrovnanou hodnotu ve středu okna platí $\hat{y}_t = \hat{T}_{t,0} = \hat{a}_{0t}$.

Pozn.: Stupeň polynomu r nazýváme řád klouzavého průměru.

Najdeme řešení úlohy

$$\min_{a_0, \dots, a_r} \sum_i (y_{t,i} - a_{0t} - a_{1t}i - \dots - a_{rt}i^r)^2.$$

Dostaneme $r + 1$ rovnic

$$\begin{aligned} -2 \sum_i (y_{t,i} - a_{0t} - a_{1t}i - \dots - a_{rt}i^r) &= 0, \\ -2 \sum_i (y_{t,i} - a_{0t} - a_{1t}i - \dots - a_{rt}i^r)i &= 0, \\ &\vdots \\ -2 \sum_i (y_{t,i} - a_{0t} - a_{1t}i - \dots - a_{rt}i^r)i^r &= 0. \end{aligned}$$

(10)

$$\begin{aligned}
\hat{a}_0 &= \sum_{j=-p}^p w_j y_{t_j}, \\
w_j &= \sum_{j=-p}^p \frac{\sum_i i^4 - j^2 \sum i^2}{m \sum i^4 - (\sum i^2)^2}.
\end{aligned}
\tag{11}$$

Výpočet s odvozením vah pro konkrétní stupeň, viz přednáška.

Váhy klouzavých průměrů stupně 2 a 3 pro různé délky okna

3	$(0, 1, 0)$
5	$1/35(-3, 12, 17, \dots)$
7	$1/21(-2, 3, 6, 7, \dots)$
9	$1/231(-21, 14, 39, 54, 59, \dots)$
11	$1/429(-36, 9, 44, 69, 84, 89, \dots)$
13	$1/143(-11, 0, 9, 16, 21, 24, 25, \dots)$
15	$1/1105(-78, -13, 42, 87, 122, 147, 162, 167, \dots)$
17	$1/323(-21, -6, 7, 18, 27, 34, 39, 42, 43, \dots)$
19	$1/2261(-136, -51, 24, 89, 144, 189, 224, 249, 264, 269, \dots)$
21	$1/3059(-171, -76, 9, 84, 149, 204, 249, 284, 309, 324, 329, \dots)$

7.5.2. Exponenciální vyhlazování

Jde o tzv. adaptivní metodu, která rychle reaguje na změnu v datech.

Při vyrovnávání se používají všechna pozorování, ale těm starším se přiřazuje menší váha.

Předpokládáme, že trend analyzované časové řady je konstantní, tj.

$$T_{t-k} = \beta_0$$

Dosadíme-li do vážené podmínky metody nejmenších čtverců a minimalizujeme-li funkci

$$\sum_{k=0}^{t-1} (y_{t-k} - T_{t-k})^2 \alpha^k \rightarrow \min,$$

tedy

$$\sum_{k=0}^{t-1} (y_{t-k} - \beta_0)^2 \alpha^k \rightarrow \min,$$

po zderivování podle β_0 dostáváme rovnici

$$-2 \sum_{k=0}^{t-1} (y_{t-k} - \beta_0) \alpha^k = 0,$$

Vyřešením dostáváme

$$\sum_{k=0}^{t-1} \alpha^k y_{t-k} = \beta_0 \sum_{k=0}^{t-1} \alpha^k.$$

Dostáváme

$$\beta_0 = \frac{\sum_{k=0}^{t-1} \alpha^k y_{t-k}}{\sum_{k=0}^{t-1} \alpha^k}.$$

Řadu ve jmenovateli lze sečíst (při užití aproximace). Uvažujme, že

$$\sum_{k=0}^{t-1} \alpha^k = \sum_{k=0}^{\infty} \alpha^k,$$

součet této geometrické řady pak je $\frac{1}{1-\alpha}$.

Tedy

$$\beta_0 = (1 - \alpha) \sum_{k=0}^{t-1} \alpha^k y_{t-k}.$$

Tento odhad použijeme pro výpočet vyrovnané hodnoty

$$\hat{y}_{t-k} = \hat{\beta}_0, k = 0, 1, 2, \dots$$

$$\hat{y}_t = (1 - \alpha) \sum_{k=0}^{t-1} \alpha^k y_{t-k}$$

Vyraz $(1 - \alpha)\alpha^k$ je váha pozorování, která směrem do minulosti klesá.

Pro výpočet lze užít rekurentní vzorec

$$\begin{aligned} \hat{y}_t &= (1 - \alpha)y_t + (1 - \alpha) \sum_{k=0}^{t-1} \alpha^k y_{t-k} = \\ &= (1 - \alpha)y_t + \alpha(1 - \alpha) \sum_{k=1}^{t-1} \alpha^{k-1} y_{t-1+k-1} = \\ &= (1 - \alpha)y_t + \alpha(1 - \alpha) \sum_{j=0}^{t-2} \alpha^j y_{t-1-j} = \\ &= (1 - \alpha)y_t + \alpha \hat{y}_{t-1}. \end{aligned}$$

Numerický výpočet realizujeme dle posledního vztahu

$$\hat{y}_t = (1 - \alpha)y_t + \alpha \hat{y}_{t-1}.$$

Předpověď je

$$\hat{y}_{t+T} = \hat{y}_t, \text{ pro } T = 1, 2, \dots$$

Volba vyhlazovací konstanty je doporučována v intervalu $\langle 0,7, 1,0 \rangle$.

Pokud známe několik budoucích hodnot, můžeme s nimi srovnat kvalitu předpovědi pro různé α a zvolit nejvhodnější.

Předpokládejme, že trend analyzované časové řady je lineární, tj. ve vztahu

$$T_{n-k} = \beta_0 - \beta_1 k + \beta_2 k^2 + \dots + (-1)^k \beta_r k^r$$

položíme $r = 1$, takže

$$T_{n-k} = \beta_0 - \beta_1 k.$$

Dosadíme-li do vážené podmínky metody nejmenších čtverců a minimalizujeme-li funkci

$$\sum_{k=0}^{n-1} (y_{n-k} - T_{n-k})^2 \alpha^k \rightarrow \min,$$

dostaneme známým způsobem normální rovnice pro odhad parametrů b_0 a b_1 :

$$\begin{aligned} \sum_{k=0}^{n-1} \alpha^k y_{n-k} &= b_0 \sum_{k=0}^{n-1} \alpha^k - b_1 \sum_{k=0}^{n-1} k \alpha^k, \\ - \sum_{k=0}^{n-1} k \alpha^k y_{n-k} &= -b_0 \sum_{k=0}^{n-1} k \alpha^k + b_1 \sum_{k=0}^{n-1} k^2 \alpha^k. \end{aligned} \quad (*)$$

Jejich řešením dostaneme

$$\begin{aligned} b_0 &= \frac{\sum_{k=0}^{n-1} \alpha^k y_{n-k} \sum_{k=0}^{n-1} k^2 \alpha^k - \sum_{k=0}^{n-1} k \alpha^k \sum_{k=0}^{n-1} k \alpha^k y_{n-k}}{\sum_{k=0}^{n-1} \alpha^k \sum_{k=0}^{n-1} k^2 \alpha^k - (\sum_{k=0}^{n-1} k \alpha^k)^2}, \\ b_1 &= \frac{- \sum_{k=0}^{n-1} \alpha^k \sum_{k=0}^{n-1} k \alpha^k y_{n-k} + \sum_{k=0}^{n-1} \alpha^k y_{n-k} \sum_{k=0}^{n-1} k \alpha^k}{\sum_{k=0}^{n-1} \alpha^k \sum_{k=0}^{n-1} k^2 \alpha^k - (\sum_{k=0}^{n-1} k \alpha^k)^2}. \end{aligned}$$

Kromě přesného výpočtu z normálních rovnic (*) lze při odhadu parametrů použít pro vyšší hodnoty počtu pozorování n jednodušší aproximativní vzorce. Jestliže $n \rightarrow \infty$, potom vzhledem k tomu, že $0 < \alpha < 1$, platí vztahy

$$\begin{aligned} \sum_{k=0}^{n-1} \alpha^k &= \sum_{k=0}^{\infty} \alpha^k = \frac{1}{1 - \alpha}, \\ \sum_{k=0}^{n-1} k \alpha^k &= \sum_{k=0}^{\infty} k \alpha^k = \frac{1}{(1 - \alpha)^2}, \\ \sum_{k=0}^{n-1} k^2 \alpha^k &= \sum_{k=0}^{\infty} k^2 \alpha^k = \frac{\alpha(1 + \alpha)}{(1 - \alpha)^3}. \end{aligned}$$

Zavedeme-li pomocné veličiny

$$\begin{aligned} S_n &= (1 - \alpha) \sum_{k=0}^{n-1} \alpha^k y_{n-k}, \\ S_n^+ &= (1 - \alpha)^2 \sum_{k=0}^{n-1} \alpha^k y_{n-k}, \end{aligned}$$

dostaneme pro odhady parametrů β_0 a β_1 po provedení dalších algebraických úprav výpočtové vzorce:

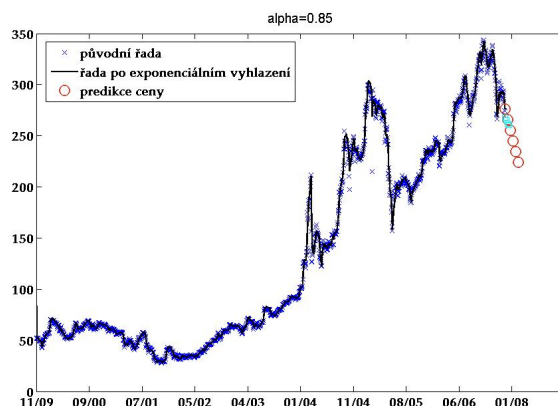
$$\begin{aligned} b_0 &= (1 + \alpha) S_n - S_n^+, \\ b_1 &= (1 - \alpha) S_n - \frac{1 - \alpha}{\alpha} S_n^+. \end{aligned}$$

V následujícím grafu je znázorněna cena akcie a k ní odpovídající řada vyhlazená pomocí exponenciálního vyrovnání.

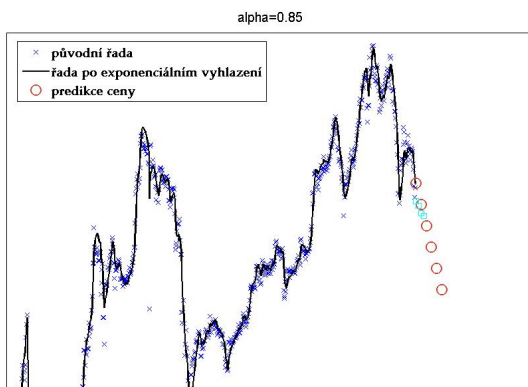
V uvedeném algoritmu je třeba pro výpočet predikovaných hodnot zvolit vhodný koeficient α . Nejčastěji se volí α splňující nerovnost $0,7 \leq \alpha < 1$.

Hodnotu α lze vybrat z trénovacího souboru obsahujícího data např. cen akcie s délkou dat kratší např. o posledních 50 obchodních dní. Na takové časové řadě se aplikuje exponenciální vyrovňávání a získané odhady budoucích cen se porovnají se skutečnými, nám známými, hodnotami cen akcií v posledních 50 obchodních dnech.

Hodnota α se postupně upravuje tak, aby byly minimalizovány korekce mezi předpovědí a skutečnou cenou v těchto 50 obchodních dnech.



Obr. : Exponenciální vyrovňávání a předpověď ceny akcie Unipetrol



Obr. : Detail předpovědi ceny akcie Unipetrol

7.6. Analýza periodické složky

7.6.1. Fourierova řada

Fourierova řada je velice užitečný matematický nástroj, který dovoluje vyjádřit nesinusovou periodickou funkci pomocí řady sinusových složek. Tato možnost se často využívá i v elektrotechnice při zkoumání nejrozličnějších periodických průběhů elektrických napětí a proudů, nebo i jiných veličin.

Mějme periodický signál s dobou periody T_0 , jenž je vyjádřen funkcí času $f(t)$. Nechť funkce $f(t)$ splňuje tzv. **Dirichletovy podmínky**:

1. Uvnitř intervalu T_0 je jednoznačná, má v tomto intervalu konečný počet bodů nespojitosti a konečný počet maxim a minim.
2. Je absolutně integrovatelná, tedy vyhovuje podmínce

$$\int_{-T_0/2}^{T_0/2} |f(t)| dt < \infty. \quad (12)$$

Jsou-li uvedené podmínky splněny, je možné funkci $f(t)$ vyjádřit pomocí nekonečné Fourierovy řady

$$f(t) = a_0 + 2 \sum_{n=1}^{\infty} [a_n \cos(n2\pi f_0 t) + b_n \sin(n2\pi f_0 t)], \quad (13)$$

$$t \in (-\infty, \infty).$$

Uvažovanou funkci je tedy možné rozložit na stejnosměrnou složku a_0 a dále na řadu harmonických (sinusových a kosinusových) složek, o frekvencích odpovídajících celočíselným násobkům $n = 1, 2, \dots$ kmitočtu $f_0 = 1/T_0$ základního průběhu (základní tj. první harmonické). Výrazy $\cos(n2\pi f_0 t)$ a $\sin(n2\pi f_0 t)$ se nazývají **bázové funkce**.

Amplitudy harmonických složek jsou

$$a_n = 1/T_0 \int_{-T_0/2}^{T_0/2} f(t) \cos(n2\pi f_0 t) dt, \quad n = 1, 2, \dots, \quad (14)$$

$$b_n = 1/T_0 \int_{-T_0/2}^{T_0/2} f(t) \sin(n2\pi f_0 t) dt, \quad n = 1, 2, \dots, \quad (15)$$

a stejnoměrná složka

$$a_0 = 1/T_0 \int_{-T_0/2}^{T_0/2} f(t) dt. \quad (16)$$

V předchozích vztazích se provádí integrace na intervalech $\langle -T_0/2, T_0/2 \rangle$, tyto meze však mohou ohraničovat libovolný časový interval o délce odpovídající době periody T_0 (tedy například interval $\langle 0, T_0 \rangle$ apod.). Páry harmonických složek (dále označovaných stručně jako harmonické) je možné sloučit do jediné složky o určité amplitudě A_n a fázi Θ_n .

Vyjádříme-li totiž amplitudy harmonických ve tvaru

$$a_n = A_n \cos(\theta_n); \quad b_n = A_n \sin(\theta_n), \quad (17)$$

lze řadu (1.12) převést do tvaru

$$f(t) = a_0 + 2 \sum_{n=1}^{\infty} [A_n \cos(n2\pi f_0 t + \theta_n)], \quad (18)$$

$$-T_0/2 \leq t \leq T_0/2,$$

přičemž

$$A_n = \sqrt{a_n^2 + b_n^2}; \quad \theta_n = -\arctg(b_n/a_n).$$

7.6.2. *Periodogram*

V této části se budeme zabývat periodogramem, který se velmi často používá např. v hydrologii. V další kapitole použijeme periodogram pro nalezení statisticky významných frekvencí ve vzorcích různých signálů.

Mějme konečnou posloupnost náhodných veličin X_1, \dots, X_N . Definujme funkci $I(\lambda)$ vzorcem

$$I(\lambda) = \frac{1}{2\pi N} \left| \sum_{t=1}^N X_t e^{-it\lambda} \right|^2, \quad -\pi \leq \lambda \leq \pi, \quad (19)$$

Funkce $I(\lambda)$ se nazývá *periodogram posloupnosti* X_1, \dots, X_N .

Při každém pevném N je $I(\lambda)$ náhodná veličina. Proto $I(\lambda)$, $-\pi \leq \lambda \leq \pi$, tvoří náhodný proces. Každá realizace tohoto procesu je spojitá funkce.

Bude-li délka posloupnosti N malá ve srovnání s délkou skutečné periody

$$T_j = \frac{2\pi}{\lambda_j}, \quad (20)$$

kde $\lambda_1, \dots, \lambda_p$ jsou vzájemně různá čísla z intervalu $(-\pi, \pi)$, bude se tato perioda jevit spíše jako trend.

Naproti tomu příliš krátké periody rovněž nebude možno správně rozeznat. Je zřejmé, že nejkratší zjištělná periodicitá má délku $T = 2$. Podle (3.2) jí odpovídá frekvence $\lambda = \pi$, které se říká **Nyquistova frekvence**.

Pro snadnější vyšetření statistické vlastnosti periodogramu, odvodíme nejprve jiné tvary vzorce. Omezíme se na případ, že posloupnost X_t je reálná.

LEMMA 1. Položme

$$A(\lambda) = (2/N)^{1/2} \sum_{t=1}^N X_t \cos t\lambda, \quad B(\lambda) = (2/N)^{1/2} \sum_{t=1}^N X_t \sin t\lambda. \quad (21)$$

Pak platí

$$I(\lambda) = \frac{1}{4\pi} A^2(\lambda) + \frac{1}{4\pi} B^2(\lambda). \quad (22)$$

Zřejmě platí

$$\begin{aligned} I(\lambda) &= \frac{1}{2\pi N} \left| \sum_{t=1}^N X_t e^{-it\lambda} \right|^2 = \frac{1}{2\pi N} \left| \sum_{t=1}^N X_t \cos(t\lambda) - i \sum_{t=1}^N X_t \sin(t\lambda) \right|^2 = \\ &= \frac{1}{2\pi N} \left[\left(\sum_{t=1}^N X_t \cos t\lambda \right)^2 + \left(\sum_{t=1}^N X_t \sin t\lambda \right)^2 \right] = \frac{1}{4\pi} [A^2(\lambda) + B^2(\lambda)]. \end{aligned}$$

LEMMA 2. Nechť

$$C_k = \frac{1}{N} \sum_{t=1}^{N-k} X_t X_{t+k}, \quad k = 0, 1, \dots, N-1. \quad (23)$$

Pak

$$I(\lambda) = \frac{1}{2\pi} (C_0 + 2 \sum_{k=1}^{N-1} C_k \cos k\lambda). \quad (24)$$

Máme

$$I(\lambda) = \frac{1}{2\pi N} \left| \sum_{t=1}^N X_t e^{-it\lambda} \right|^2 = \frac{1}{2\pi N} \sum_{s=1}^N \sum_{t=1}^N X_s X_t e^{-i(s-t)\lambda}.$$

7.6.3. Test R. A. Fishera

R. A. Fisher odvodil test, kterým se dá zjistit statistická významnost nejvyšších hodnot periodogramu.

S použitím tohoto testu u periodogramu budeme schopni na dané hladině významnosti schopni určit statisticky významné frekvence v naší časové řadě.

Uvažujme posloupnost náhodných veličin X_1, \dots, X_N . Budeme testovat hypotézu, že X_1, \dots, X_N je posloupnost nezávislých náhodných veličin s rozdělením $N(0, \sigma^2)$. Předpokládejme, že N je číslo liché. Z praktického hlediska se můžeme omezit na $N \geq 3$, takže

$$N = 2m + 1, \quad (25)$$

kde m je přirozené číslo. Toto omezení na lichá N je výhodné z toho důvodu, že se tím podstatně zjednoduší další matematické odvozování, viz [1]. Je-li dána posloupnost náhodných veličin se sudým počtem členů, první z nich se v praxi obvykle vynechává jakožto časově nejvzdálenější od současnosti.

Uvažujme hodnoty periodogramu v bodech

$$\lambda = 2\pi r/N, \quad r = 1, 2, \dots, m. \quad (26)$$

Srovnajme hodnoty $I(\lambda_1), \dots, I(\lambda_m)$ sestupně podle velikosti. Označíme V_1 největší z nich atd. až V_m nejmenší. Položme

$$W = \frac{V_1}{V_1 + V_2 + \dots + V_m}. \quad (27)$$

Budou-li všechny veličiny V_1, \dots, V_m téměř stejné, bude hodnota W blízká číslu $\frac{1}{m}$. Bude-li naopak veličina V_1 nabývat velmi vysokých hodnot ve srovnání s ostatními veličinami V_2, \dots, V_m , bude hodnota W blízká jedné. Hodnoty W (které jsou blízké jedné) budou tvořit kritický obor naší hypotézy. Pětiprocentní kritická hodnota je takové číslo α , pro které platí $P(W > \alpha) = 0.05$. Lze tedy postupovat tak, že se z dané realizace posloupnosti X_1, \dots, X_N vypočtou hodnoty periodogramu $I(\lambda_1), \dots, I(\lambda_m)$. Pak se vypočítá W podle (3.10). Překročí-li W kritickou hodnotu, zamítneme hypotézu, že X_1, \dots, X_N jsou nezávislé veličiny s rozdělením $N(0, \sigma^2)$.

7.6.4. Periodogram jako rozklad rozptylu

Nechť X_1, X_2, \dots, X_N je posloupnost reálných veličin, přičemž $N = 2m + 1$, kde m je přirozené číslo. Označme

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i, \quad s^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2, \quad (28)$$

$$I(\lambda) = \frac{1}{2\pi N} \left| \sum_{t=1}^N X_t - \bar{X} e^{-it\lambda} \right|^2. \quad (29)$$

V dalším textu nebudeme uvažovat původní hodnoty X_1, \dots, X_N , ale jejich odchylky od aritmetického průměru \bar{X} .

Asymptotické chování pravděpodobnosti $P(W > \alpha)$ Místo pouhého porovnání s kritickou hodnotou dáme přednost vypočtení hodnoty $P(W > \alpha)$. R. A. Fisher zjistil, že pro $m \leq 50$ se dostane dobrá aproximace pravděpodobnosti $P(W > \alpha)$, vztahem

$$P(W > \alpha) \cong m(1 - \alpha)^{m-1}, \quad (30)$$

$$\begin{aligned} h &= m(1 - \alpha)^{m-1}, & \text{kde } h \dots \text{ kritická hodnota,} \\ \frac{h}{m} &= (1 - \alpha)^{m-1}, & \alpha \dots \text{ hladina významnosti,} \\ \left(\frac{h}{m}\right)^{\frac{1}{m-1}} &= 1 - \alpha, & m \dots \text{ přirozené číslo, počet pozorování,} \\ \alpha &= 1 - \left(\frac{h}{m}\right)^{\frac{1}{m-1}}, & W \dots \text{ kritický obor, viz [2] .} \end{aligned}$$

Tabulka

Aproximace kritické hodnoty na hladině významnosti			
m	$\alpha=10\%$	$\alpha=5\%$	$\alpha=1\%$
5	0.6239397	0.6837722	0.7885257
6	0.5590699	0.6161481	0.7217919
7	0.5074122	0.5611542	0.6644038
8	0.4652756	0.5156875	0.6151665
9	0.4302036	0.4774944	0.5727130
10	0.4005157	0.4449527	0.5358411
11	0.3750278	0.4168803	0.5035669
12	0.3528819	0.3924008	0.4751026
13	0.3334418	0.3708529	0.4498201
14	0.3162250	0.3517283	0.4272170
15	0.3008588	0.3346307	0.4068885
16	0.2870507	0.3192463	0.3885063
17	0.2745676	0.3053236	0.3718014
18	0.2632216	0.2926583	0.3565521
19	0.2528590	0.2810831	0.3425738
20	0.2433533	0.2704594	0.3297118
21	0.2345990	0.2606714	0.3178356
22	0.2265076	0.2516215	0.3068344
23	0.2190041	0.2432271	0.2966132

Aproximace			
kritické hodnoty na hladině významnosti			
m	$\alpha=10\%$	$\alpha=5\%$	$\alpha=1\%$
24	0.2120248	0.2354176	0.2870907
25	0.2055148	0.2281322	0.2781962
26	0.1994268	0.2213185	0.2698685
27	0.1937198	0.2149308	0.2620542
28	0.1883580	0.2089294	0.2547062
29	0.1833100	0.2032792	0.2477834
30	0.1785483	0.1979496	0.2412490
31	0.1740484	0.1929131	0.2350707
32	0.1697886	0.1881458	0.2292196
33	0.1657496	0.1836258	0.2236699
34	0.1619141	0.1793341	0.2183983
35	0.1582668	0.1752532	0.2133841
36	0.1547936	0.1713676	0.2086085
37	0.1514820	0.1676632	0.2040546
38	0.1483206	0.1641272	0.1997069
39	0.1452992	0.1607483	0.1955514
40	0.1424084	0.1575157	0.1915755
41	0.1396396	0.1544201	0.1877674
42	0.1369851	0.1514526	0.1841167
43	0.1344376	0.1486052	0.1806134
44	0.1319907	0.1458706	0.1772488
45	0.1296384	0.1432421	0.1740145
46	0.1273751	0.1407134	0.1709029
47	0.1251957	0.1382788	0.1679072
48	0.1230955	0.1359330	0.1650207
49	0.1210701	0.1336711	0.1622375
50	0.1191155	0.1314886	0.1595521

Příklad – periodogram

V tabulce jsou uvedeny průměrné měsíční průtoky vody na řece Moravě ve stanici Strážnice. Poněvadž je jich sudý počet, nebrali jsme v úvahu první pozorování. Od ostatních hodnot jsme odečetli jejich aritmetický průměr, abychom se přiblížili předpokladu o nulové střední hodnotě posloupnosti. Z těchto odchylek od průměru jsme vypočetli periodogram, který je uveden v tabulce 3. Pro hodnotu periodogramu v bodě $\lambda = \lambda_6 = 0.5310$ získáváme $3617.1 \cdot 10^{-6}$. Tato frekvence podle vzorce (3.2) odpovídá periodě délky $T = \frac{2\pi}{\lambda} \cong 11.83$ měsíce.

V tabulce jsou uvedeny průměrné měsíční průtoky vody na řece Moravě ve stanici Strážnice. Poněvadž je jich sudý počet, nebrali jsme v úvahu první pozorování. Od ostatních hodnot jsme odečetli jejich aritmetický průměr, abychom se přiblížili předpokladu o nulové střední hodnotě posloupnosti. Z těchto odchylek od průměru jsme vypočetli periodogram, který je uveden v tabulce 3. Pro hodnotu periodogramu v bodě $\lambda = \lambda_6 = 0.5310$ získáváme $3617.1 \cdot 10^{-6}$. Tato frekvence podle vzorce (3.2) odpovídá periodě délky $T = \frac{2\pi}{\lambda} \cong 11.83$ měsíce.

Tabulka:

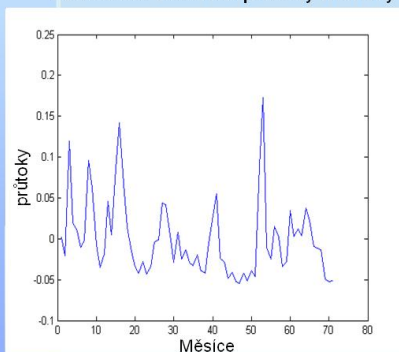
Roky	1965	1966	1967	1968	1969	1970	1971
Leden		0,0487	0,0748	0,069	0,0277	0,0307	0,0815
Únor		0,189	0,156	0,114	0,0641	0,0233	0,074
Březen		0,0889	0,212	0,112	0,0986	0,159	0,107
Duben		0,0794	0,137	0,075	0,125	0,242	0,092
Květen		0,0597	0,0838	0,0408	0,046	0,0598	0,0602
Červen		0,0676	0,0545	0,0781	0,0408	0,045	0,0581
Červenec		0,166	0,0353	0,0453	0,0215	0,0845	0,0564
Srpen		0,132	0,0278	0,0561	0,0282	0,0719	0,0205
Září		0,0626	0,0421	0,0411	0,0172	0,0362	0,0178
Říjen		0,0347	0,0261	0,0366	0,0148	0,0419	0,0187
Listopad	0,0268	0,0509	0,0353	0,0501	0,0278	0,104	
Prosinec	0,0718	0,116	0,065	0,0309	0,0186	0,0726	

j	λ_j	$T_j = \frac{2\pi}{\lambda_j}$	$I(\lambda_j)10^{-6}$	j	λ_j	$T_j = \frac{2\pi}{\lambda_j}$	$I(\lambda_j)10^{-6}$
1	0,0885	71,00	864,3	19	1,6814	3,74	418,1
2	0,177	35,50	352,3	20	1,7699	3,55	97,6
3	0,2655	23,67	779,4	21	1,8584	3,38	180,6
4	0,354	17,75	20,3	22	1,9469	3,23	4,5
5	0,4425	14,2	190,9	23	2,0354	3,09	371,3
6	0,5310	11,83	3617,1	24	2,1239	2,96	21,2
7	0,6195	10,14	133,3	25	2,2124	2,84	60,5
8	0,708	8,88	436	26	2,3009	2,73	69,5
9	0,7965	7,89	105,5	27	2,3894	2,63	102,6
10	0,885	7,1	413,5	28	2,4779	2,54	78,1
11	0,9735	6,45	1244,9	29	2,5664	2,45	186,5
12	1,0619	5,92	266,3	30	2,6549	2,37	368,2
13	1,1504	5,46	388,1	31	2,7434	2,29	3,4
14	1,2389	5,07	139,5	32	2,8319	2,22	2,5
15	1,3274	4,73	80,8	33	2,9204	2,15	36,9
16	1,4159	4,44	419,3	34	3,0088	2,09	71,1
17	1,5044	4,18	324	35	3,0973	2,03	115,4
18	1,5929	3,94	586,7				

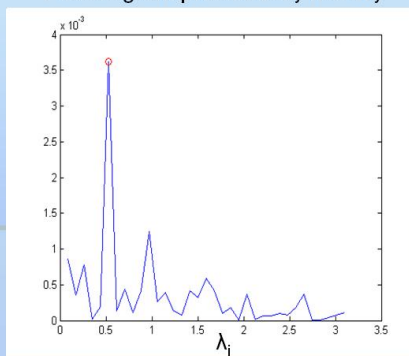
Významnost hodnoty $I(\lambda_6)$ ověříme pomocí Fisherova testu. Dostaneme $W = 0.2882$. Použijeme uvedenou aproximaci. V tabulce najdeme kritickou hodnotu pro $m = 35$ na hladině významnosti $\alpha = 5\%$, což je 0.1753. Zjistíme, že jde o významnou hodnotu. Na grafu jsme významnou hodnotu periodogramu, zvýraznili kroužkem.

Použití periodogramu v hydrologii

Průměrné měsíční průtoky Moravy



Periodogram průtoků řeky Moravy



$$\lambda_j = 0.5310$$

$$T = 11.83 \text{ měsíců}$$

Obr. : Průměrné roční průtoky vody

8

Lineární regrese

8.1. Úvod

Cílem regresní analýzy je popsat závislost hodnot znaku Y na hodnotách znaku x . K dispozici bývají měření závislé (vysvětlované) proměnné $\mathbf{Y} = (Y_1, \dots, Y_n)'$ realizované v n bodech $\mathbf{x} = (x_1, \dots, x_n)'$. Vektor \mathbf{x} představuje proměnnou vysvětlující.

Nejprve je třeba najít vhodný model, tj. zvolit vhodnou aproximující funkci $Y = f(x, \boldsymbol{\beta})$. Poté je třeba nalézt nevychýlený odhad $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_k)$ neznámých parametrů $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)$ této již známé funkce. Pokud je aproximující funkce lineární v parametrech β_j , hovoříme o lineární regresi.

V úlohách lineární regrese volíme aproximující funkci ve tvaru

$$y = \sum_{j=0}^k \beta_j \phi_j(x),$$

kde funkce $\phi_j(x)$ jsou známé funkce neobsahující neznámé parametry a $\phi_0(x) = 1$. Většinou volíme $\phi_j(x) = x^j$, $j = 0, 1, \dots, k$.

Aproximující funkce může být i složitější a nemusí být v parametrech β_j lineární, pak se jedná o problém nelineární regrese. Například regresní funkce obsahuje mocniny neznámých parametrů β_j anebo tyto parametry jsou argumenty vnitřních funkcí, které tvoří aproximující funkci. V tomto kurzu budeme studovat pouze problém lineární regrese. Nelineární regresi můžete dostudovat v [4], [5].

V další části seznámíme čtenáře s historií algoritmů pro aproximaci dat přímkou, jejichž vznik bývá označován za první statistickou revoluci.

8.2. První statistická revoluce

Dnes se pro řešení problému lineární regrese obvykle používá metoda nejmenších čtverců, ojediněle se volí ortogonální metoda nejmenších čtverců (korekce měření se provádí po normále k aproximující funkci).

Pro seznámení čtenáře s regresí popíšeme principy algoritmů pro řešení (dnes nepoužívaných a pozapomenutých) a uvedeme jednu historickou úlohu, která značně přispěla k tzv. první statistické revoluci.

8.2.1. Vývoj algoritmů

Chorvatský matematik Roger Joseph Bošković v r. 1753 navrhuje pro určení neznámých parametrů regresní přímky následující metodu. Přímka má procházet těžištěm dat, tedy bodem $[\bar{x}, \bar{Y}]$. Bošković dále usuzuje, že v jednom z měřených bodů nutně musela být nejmenší chyba měření. Tento „nejlepší bod“ spolu s těžištěm určí „nejlepší přímku“. Kritériem pro volbu „nejlepšího bodu“ se stává funkce

$$\Phi(\beta_0, \beta_1) = \sum_{i=1}^n |Y_i - \beta_0 - \beta_1 x_i|.$$

Odhady parametrů β_0 a β_1 jsou určeny pro i -tou přímku určenou body $[x_i, Y_i]$ a $[\bar{x}, \bar{Y}]$. Nejlepší přímka je ta, pro kterou nastává $\min \Phi(\beta_0, \beta_1)$.

Boškovič navrhuje i algoritmus pro nalezení této přímky, viz literatura [Hald]. Čtenář ale ještě zvládne na počítači vyřešit v jednoduchém cyklu.

Později úlohu modifikuje Laplace v r. 1768, když použije stejné kritérium Φ . Také požaduje, aby aproximující přímka procházela těžištěm. Netrvá ale na tom, že přímka má projít některým z měřených bodů $[x_i, Y_i]$. Neznámé parametry β_0 a β_1 jsou tak hledány v množině R^2 . Algoritmus je uvedený v [Hald]. Výpočet je poměrně komplikovaný.

Další metodu navrhuje Lambert v r. 1790. Měření seřadí vzestupně podle x a rozdělí na 2 množiny, jejichž počet je stejný (pro sudé n) nebo se liší o jedničku (pro liché n). Pro každou množinu určí těžiště a jejich spojením získá regresní přímku.

Na začátku 19. století postupně publikují další (všichni stejnou) metodu Legendre v r. 1801, Audrey v r. 1802 a Gauss v r. 1804. Prvenství je nakonec připsáno Gaussovi, který prokáže použití metody pro výpočet polohy v důsledku několika týdnů zamračené oblohy ztracené komety v r. 1897. Princip metody je založen na hledání minima funkce

$$\Phi(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2.$$

Metoda dostává název metoda nejmenších čtverců (MNČ, anglická zkratka je LSM).

8.2.2. Měření obvodu Země (po poledníku)

Příklad 8.1 (měření délky jednoho stupně zeměpisné délky)

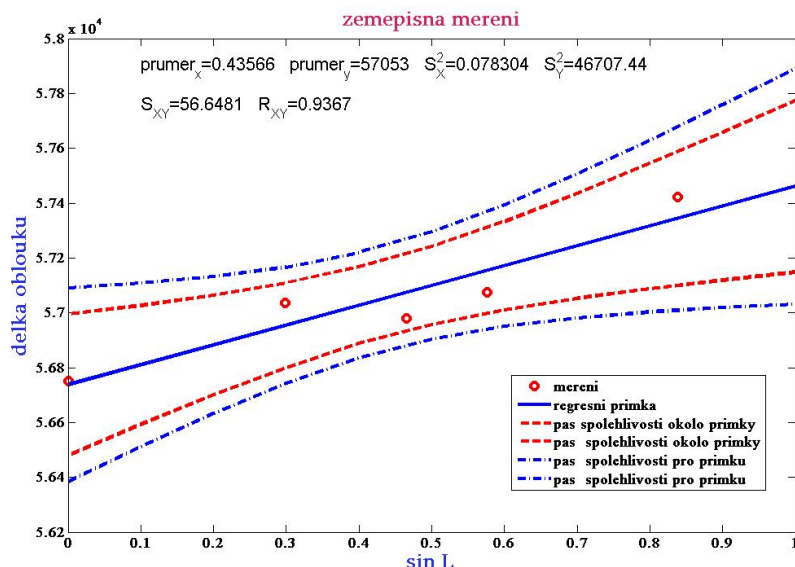
Okolo roku 1750 byla zorganizována měření délky oblouku poledníku s cílem potvrdit či vyvrátit hypotézy o tvaru Země jako rotačního elipsoidu. Měření v Římě organizuje papežský stát, měření a výpočtů se osobně zúčastní papež Benedikt XIV. Měření organizuje Roger Joseph Boškovič.

Výsledky uvádí tabulka (délky oblouku jsou uváděny v toisích):

i	zem. poloha	zem. šířka L	$x = \sin^2 L$	délka oblouku y
1	Quito	0°0'	0,0000	56751
2	Mys Dobré Naděje	33°18'	0,2987	57037
3	Řím	42°59'	0,4648	56979
4	Paříž	49°23'	0,5762	57074
5	Laponsko	66°19'	0,8386	57422

V následujícím obrázku vidíme měřené údaje, když na ose x je znázorněna poloha (druhá mocnina sinu zeměpisné šířky) a osa Y udává délku jednoho stupně zeměpisné délky na uvedené poloze.





Měření délky jednoho stupně bylo realizováno mezi body ležícími na stejném poledníku se zeměpisnou šířkou lišící se o jeden stupeň. Zeměpisná šířka byla určena z výšky slunce. Poloha na stejném poledníku byla zajištěna pomocí kompasu a chronografu. Uvedeným měřením předcházela měření délky jednoho stupně (s velkou, neakceptovatelnou nepřesností měření) zjišťovaná pomocí sledování počtu otočení kol na kočáru po silnici směřující přibližně severně.

Hodnoty uvedené v tabulce byly získány pomocí měření úhlů a vzdáleností ve vetknutých polygonálních tazích mezi dvěma danými body. Výpočet souřadnic mezilehlých bodů byl již realizován pomocí metod vyrovnávacího počtu (geodézie).

V polovině 18. století došlo již ke shodě vědců nad hypotézou, že rovník má tvar kružnice. Nezodpovězenou otázkou zůstávalo, zda poloměr rovníku je stejný, větší nebo menší než vzdálenost od středu Země k pólům.

Nalezená odhadnutá rostoucí regresní přímka vedla k závěru, že Země má tvar elipsoidu protaženějšího směrem k pólům.

Obrazek obsahuje regresní přímky určené pomocí popsaných algoritmů. Neuvádíme odhady získané z těchto metod. Cílem této úlohy bylo jen seznámit čtenáře se zadáním regresní úlohy a smyslem hledání jejího řešení.

Pro regresní úlohy je typické, že se snažíme najít zákonitosti. Snažíme se zjistit, jak závisí vysvětlovaná proměnná Y na vysvětlující proměnné x .

8.3. Odvození odhadů v metodě nejmenších čtverců

8.3.1. Přímka procházející počátkem

Nejprve najdeme řešení pomocí MNČ pro aproximující funkci dvou $\phi_0(x) = 0$ a $\phi_1(x) = x$. Jde tedy o přímku $Y = \beta_1 \phi_1(x) = \beta_1 x$ procházející počátkem.

Chceme najít

$$\min \Phi(\beta_1) = \min \sum_{i=1}^n (Y_i - \beta_1 x_i)^2.$$

Najdeme body podezřelé z extrému a pak minimum pomocí standardního postupu známého z diferenciálního počtu funkcí více proměnných.

$$\frac{\partial \Phi}{\partial \beta_1} = 2 \sum_{i=1}^n (Y_i - \beta_1 x_i)(x_i) = 0.$$

Po úpravě dostáváme

$$\sum_{i=1}^n Y_i x_i - \beta_1 \sum_{i=1}^n x_i^2 = 0.$$

Vyjádříme

$$\beta_1 = \frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2}.$$

Definice 8.1 Hodnotu kritéria Φ pro získaný odhad nazýváme reziduálním součtem čtverců a označujeme S_e .

Platí

$$S_e = \sum_{i=1}^n (Y_i - \hat{\beta}_1 x_i)^2 = \left(\sum_{i=1}^n Y_i^2 - \hat{\beta}_1 \sum_{i=1}^n x_i Y_i \right).$$

Při použití této metody předpokládáme, že $Y \sim N(\beta_1 x, \sigma^2)$. Odhad neznámého σ^2 získáme (viz [Anděl]) takto $s^2 = \frac{S_e}{n-k}$, kde n je počet měření a k je počet možných parametrů regresní přímky. Pro regresní přímku procházející počátkem je tedy $s^2 = \frac{S_e}{n-1}$.

Příklad 8.2 Mějme měděnou trubku o délce $L_0 = 1000$ mm při teplotě $t_0 = 20^\circ$ C. Bylo naměřeno, o kolik milimetrů se tato trubka prodlouží, stoupne-li její teplota o Δ° C. Je známo, že pro délkovou roztažnost platí vzorec $\Delta L = \alpha L_0 \Delta t$, kde α je tzv. koeficient tepelné roztažnosti.

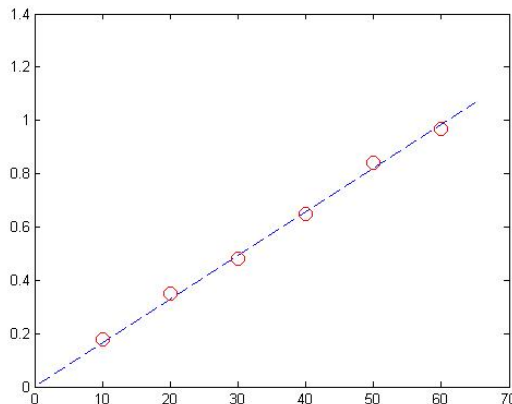


Změna teploty Δt :	10	20	30	40	50	60
Prodloužení trubky ΔL :	0,18	0,35	0,48	0,65	0,84	0,97

Je třeba odhadnout koeficient α . Proto provedeme úpravu dat z tabulky s cílem přejít na model $Y_i = \beta x_i + e_i$, $i = 1 \dots, n$, kde x_i je hodnota výrazu $L_0 \Delta t$ v i -tém měření, Y_i je prodloužení trubky, β píšeme místo α a e_i jsou chyby měření.

x_i :	10000	20000	30000	40000	50000	60000
Y_i :	0,18	0,35	0,48	0,65	0,84	0,97

Vyjde $n = 6$, $\hat{\beta} = 1,64 \times 10^{-5}$, $s^2 = 3,0264 \times 10^{-4}$.



8.3.2. Obecná přímka

Věta 8.3 V modelu lineární regrese $Y = \beta_0 + \beta_1 x$ jsou odhadem neznámých parametrů

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n Y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i Y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}, \quad (31)$$

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i Y_i - \sum_{i=1}^n x_i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}, \quad (32)$$

$$S_e(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n Y_i^2 - \hat{\beta}_0 \sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n x_i Y_i. \quad (33)$$

Věta 8.4 Pokud výpočet neznámých parametrů regresní přímky realizujeme pro vysvětlující proměnnou, od které odečteme střední hodnotu \bar{x} , dostáváme odhady regresních parametrů

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x},$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{Y} + \hat{\beta}_1 (x - \bar{x}).$$

Definice 8.2 Odchylky $e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = Y_i - \bar{Y} - \hat{\beta}_1 (x_i - \bar{x})$, $i = 1, \dots, n$, se nazývají *rezidua*.

Věta 8.5 Reziduální součet čtverců (součet čtverců reziduí) je dán vztahem

$$S_e = S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n [Y_i - \bar{Y} - \hat{\beta}_1 (x_i - \bar{x})]^2.$$

Definice 8.3 Kvalitu s jakou jsou data popsána regresní přímkou, udává index determinace I .

$$I = \sqrt{\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$

Při aplikacích se mnohdy také zajímáme o hodnotu $\beta_0 + \beta_1 x$, kde x je nějaké dané číslo, $x \in \langle \min x_i, \max x_i \rangle$.

Definice 8.4 Uvažujme dvojici statistik T_d a T_h takových, že

$$P(T_d \leq \beta_0 + \beta_1 x \leq T_h) = 1 - \alpha.$$

Dosazujeme-li do T_d, T_h různé hodnoty $x \in \langle \min x_i, \max x_i \rangle$, dostaneme při spojitě se měnícím x tzv. *pás spolehlivosti kolem regresní přímky*. Tento pás má nejmenší šířku pro $x = \bar{x}$, vzdaluje-li se x od \bar{x} , šířka pásu roste.

Věta 8.6 Oboustranný intervalový odhad hodnoty parametrické funkce $\beta_0 + \beta_1 x$ pro dané x tvoří uspořádaná dvojice statistik (T_d, T_h) :

$$T_d = \hat{\beta}_0 + \hat{\beta}_1 x - t_{n-2, 1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}, \quad (34)$$

$$T_h = \hat{\beta}_0 + \hat{\beta}_1 x + t_{n-2, 1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}. \quad (35)$$

Věta 8.7 Uvažujme model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon$, $i = 1, \dots, n$, kde $\varepsilon \sim N(0, \sigma^2)$. Lze pak psát $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$, kde $\boldsymbol{\beta} = (\beta_0, \beta_1)$ a kde

$$\mathbf{X} = \begin{pmatrix} 1, & x_1 \\ 1, & x_2 \\ \dots & \\ 1, & x_n \end{pmatrix}.$$

Pak

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Věta 8.8 Při výpočtech lze užít následující skutečnosti

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix},$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

Věta 8.9 Varianční matice odhadu je

$$\begin{aligned} \text{var}(\hat{\boldsymbol{\beta}}) &= \text{var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\text{var}(\mathbf{Y})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Věta 8.10 Při výpočtech lze užít následující skutečnosti

$$\text{var}(\hat{\beta}_0) = \sigma_{\hat{\beta}_0}^2 = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \sigma^2 \left[\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right],$$

$$\text{var}(\hat{\beta}_1) = \sigma_{\hat{\beta}_1}^2 = \frac{n\sigma^2}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Příklad 8.11 (mravenec průzkumník)

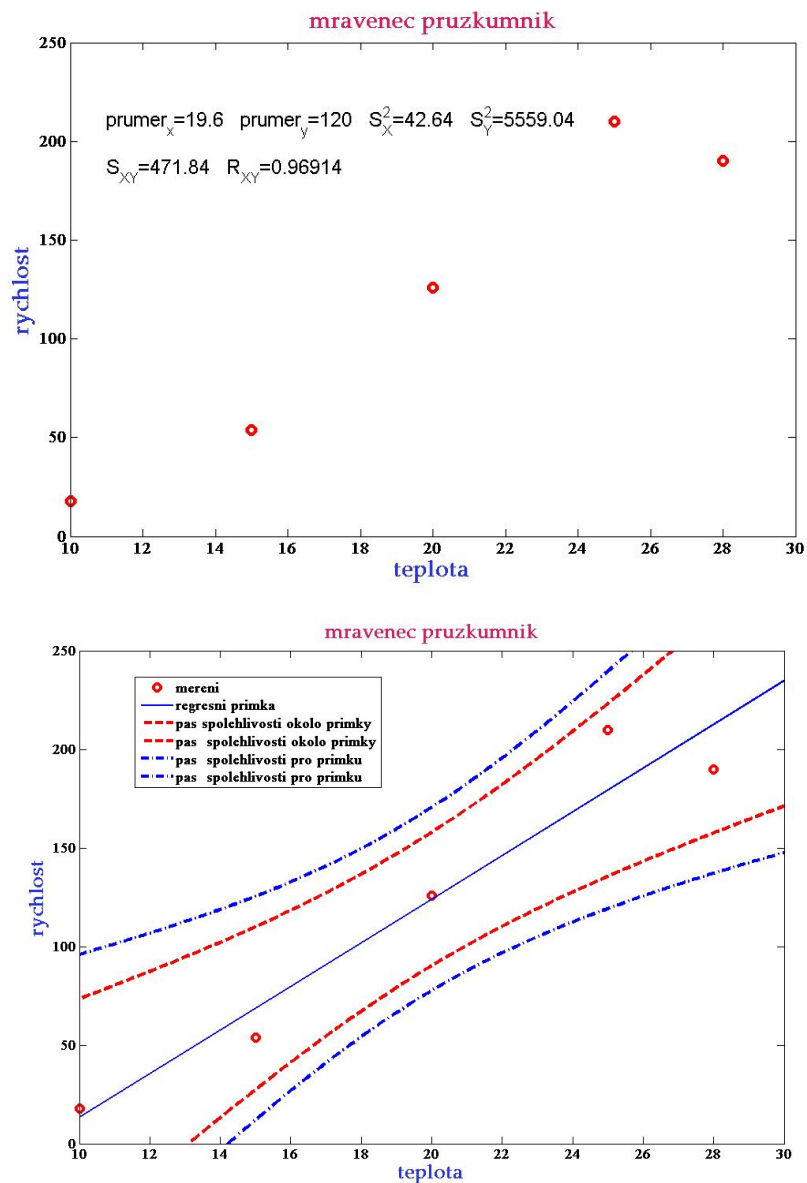


Mravenec průzkumník se probouzí při teplotě okolo 5° C, při teplotě 10° C už může dosáhnout rychlosti 18 m/hod., při teplotě 15° C vyvine rychlost 54 m/hod., při teplotě 20° C běží rychlostí 126 m/hod., při teplotě 25° C uhání rychlostí 210 m/hod., při teplotě 28° C jeho rychlost klesá na 190 m/hod.

Bernard Werber: Mravenci, KK, 2005

Najděte regresní přímku $\hat{\beta}_0 + \hat{\beta}_1 x$ pro závislost rychlosti y mravence průzkumníka na teplotě okolí x . Určete index determinace. Zkonstruujte pás spolehlivosti pro regresní přímku a okolo regresní přímky.





Výsledky uvádí tabulka

i	x_i	Y_i	x_i^2	$x_i \cdot Y_i$	Y_i^2	\hat{Y}_i	e_i	e_i^2
1	10	18	100	180	324	13,37	4,630	21,441
2	15	54	225	810	2916	68,70	-14,698	216,030
3	20	126	400	2520	15876	124,03	1,974	3,896
4	25	210	625	5250	44100	179,35	30,645	939,140
5	28	190	784	5320	36100	212,55	-22,552	508,570
Σ	98	598	2134	14080	99316	598	0	1689,1

$$\begin{aligned} \hat{\beta}_0 &= \frac{\sum_{i=1}^n Y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i Y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \\ &= \frac{598 \cdot 2134 - 98 \cdot 14080}{5 \cdot 2134 - 98^2} = \frac{-103708}{1066} = -97,287. \end{aligned}$$

$$\begin{aligned}\hat{\beta}_1 &= \frac{n \sum_{i=1}^n x_i Y_i - \sum_{i=1}^n x_i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \\ &= \frac{5 \cdot 14080 - 98 \cdot 598}{5 \cdot 2134 - 98^2} = \frac{11796}{1066} = 11,066, \\ S(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^n Y_i^2 - \hat{\beta}_0 \sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n x_i Y_i = \\ &99316 + 97,287 \cdot 598 - 11,066 \cdot 14080 = 1689,10.\end{aligned}$$

$$\begin{aligned}\hat{\beta}_0 + \hat{\beta}_1 x &= -97,287 + 11,066x, \\ S_e = 1689,1, \quad S^2 &= \frac{S_e}{n-2} = \frac{1689,10}{5-2} = 563,03, \quad S = 23,728.\end{aligned}$$

Výpočet lze realizovat pomocí těžiště $\bar{x} = 19,6$, $\bar{Y} = 119,6$:

i	x_i	Y_i	$x_i - \bar{x}$	$(x_i - \bar{x}) \cdot Y_i$	$(x_i - \bar{x})^2$	\hat{Y}_i	e_i	e_i^2
1	10	18	-9,6	-172,8	92,16	13,37	4,630	21,441
2	15	54	-4,6	-248,4	21,16	68,70	-14,698	216,030
3	20	126	0,4	50,	0,16	124,03	1,974	3,896
4	25	210	5,4	1134,0	29,16	179,35	30,645	939,140
5	28	190	8,4	1596,0	70,56	212,55	-22,552	508,570
Σ	98	598	0	2359,2	213,20	598,00	0	1689,1

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{2359,2}{213,2} = 11,066 \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{x} = 119,6 - 11,066 \cdot 19,6 = -97,287.\end{aligned}$$

$$\begin{aligned}\hat{\beta}_0 + \hat{\beta}_1 x &= -97,287 + 11,066x \\ S_e = 1689,1, \quad S^2 &= \frac{S_e}{n-2} = \frac{1689,10}{5-2} = 563,03, \quad S = 23,728.\end{aligned}$$

Určíme index determinace

$$I = \sqrt{\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}} = \sqrt{\frac{26105,0}{27795,2}} = 0,969.$$

i	x_i	Y_i	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$	$\hat{Y}_i - \bar{Y}$	$(\hat{Y}_i - \bar{Y})^2$
1	10	18	-101,6	10322,56	-106,23	11284,81
2	15	54	-65,6	4303,36	-50,90	2590,81
3	20	126	6,4	40,96	4,43	19,62
4	25	210	90,4	8172,16	59,75	3570,06
5	28	190	70,4	4956,16	92,95	8639,70
Σ	98	598	0	27795,2	0	26105,00

Výpočtem zjistíme, že

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix},$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

V našem případě

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 5 & 98 \\ 98 & 2134 \end{pmatrix},$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{5 \cdot 2134 - 98^2} \begin{pmatrix} 2134 & -98 \\ -98 & 5 \end{pmatrix} = \begin{pmatrix} 2,0019 & -0,0919 \\ -0,0919 & 0,0047 \end{pmatrix}.$$

$$\begin{aligned} \text{var}(\hat{\beta}) &= \text{var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\text{var}(\mathbf{Y})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Proto

$$\begin{aligned} \text{var}(\hat{\beta}_0) &= \sigma_{\hat{\beta}_0}^2 = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \sigma^2 \left[\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] = \\ &= 563,03 \cdot 2,0019 = 1127,1298 \end{aligned}$$

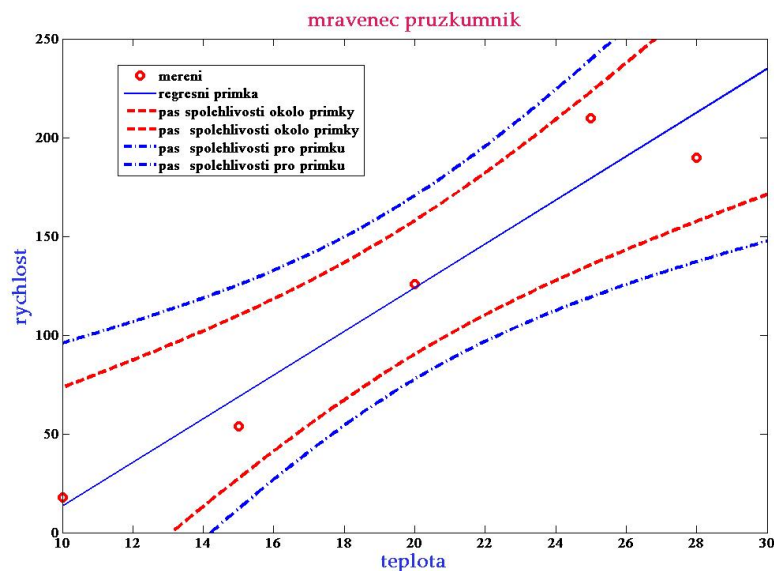
$$\begin{aligned} \text{var}(\hat{\beta}_1) &= \sigma_{\hat{\beta}_1}^2 = \frac{n\sigma^2}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \\ &= 563,03 \cdot 0,0047 = 2,6409 \end{aligned}$$

Určeme interval spolehlivosti pro $x = 20$.

$$\begin{aligned} T_d &= \hat{\beta}_0 + \hat{\beta}_1 x - t_{n-2, 1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = \\ &= -97,287 + 11,066 \cdot 20 - 3,182 \cdot \sqrt{\frac{1}{5} + \frac{(20 - 19,6)^2}{213,20}} = 90,20 \end{aligned}$$

$$\begin{aligned} T_h &= \hat{\beta}_0 + \hat{\beta}_1 x + t_{n-2, 1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = \\ &= -97,287 + 11,066 \cdot 20 + 3,182 \cdot \sqrt{\frac{1}{5} + \frac{(20 - 19,6)^2}{213,20}} = 157,86 \end{aligned}$$

Dosazujeme-li do T_d, T_h různé hodnoty $x \in \langle \min x_i, \max x_i \rangle$, dostaneme při spojitě se měnícím x tzv. *pás spolehlivosti kolem regresní přímky*.



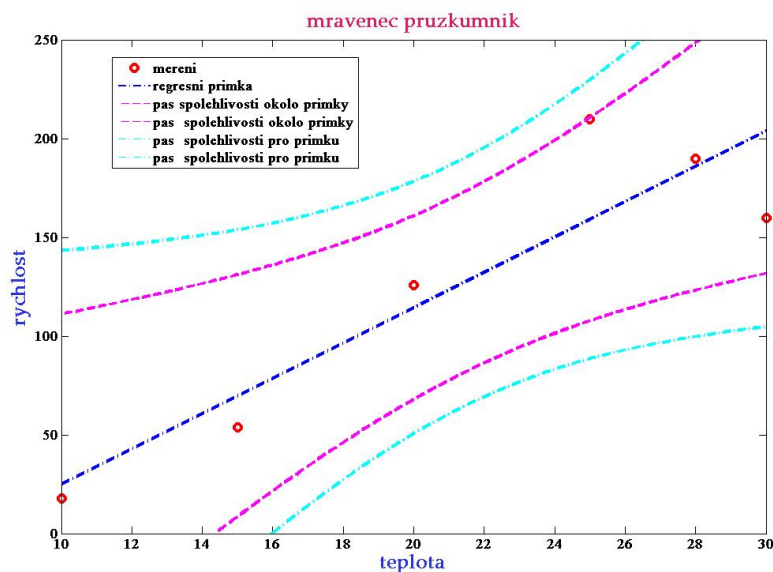
Příklad 8.12 (mravenec průzkumník a vyšší teplota)



Přepočtete předchozí úlohu, když navíc uvažujete šesté měření při teplotě 30° C s naměřenou rychlostí mravence průzkumníka 140 m/hod.



Najděte regresní přímku $\hat{\beta}_0 + \hat{\beta}_1 x$ pro závislost rychlosti y mravence průzkumníka na teplotě okolí x . Určete index determinace. Zkonstruuje pás spolehlivosti pro regresní přímku a okolo regresní přímky. Při výpočtu užíjte tabulek z předchozího příkladu.



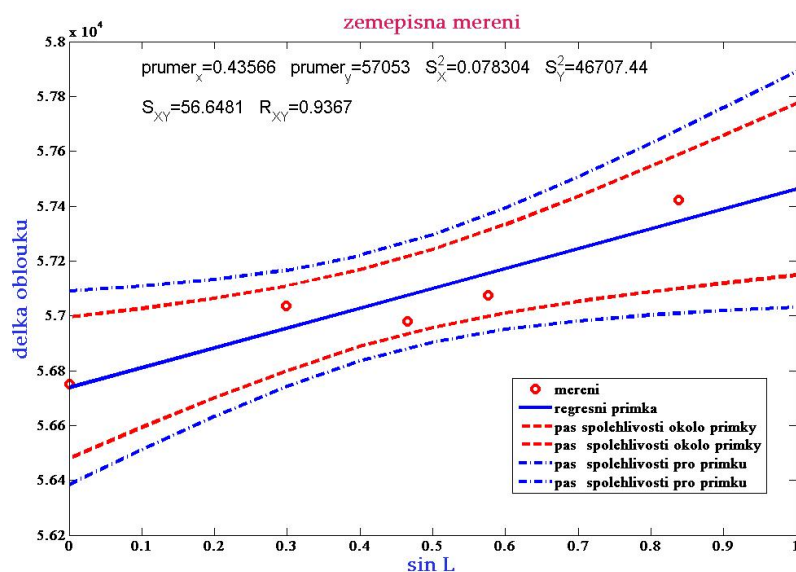
Příklad 8.13 (měření délky jednoho stupně zeměpisné délky)

Okolo roku 1750 byla zorganizována měření délky oblouku poledníku s cílem potvrdit či vyvrátit hypotézy o tvaru Země jako rotačního elipsoidu. (Na měření v Římě se podílil Roger Joseph Boškovič na objednávku papeže Benedikta XIV.)



Výsledky uvádí tabulka (délky oblouku jsou uváděny v toisích):

i	zem. poloha	zem. šířka L	$x = \sin^2 L$	délka oblouku y
1	Quito	$0^\circ 0'$	0,0000	56751
2	Mys Dobré Naděje	$33^\circ 18'$	0,2987	57037
3	Řím	$42^\circ 59'$	0,4648	56979
4	Paříž	$49^\circ 23'$	0,5762	57074
5	Laponsko	$66^\circ 19'$	0,8386	57422



Výsledky uvádí tabulka

i	x_i	Y_i	x_i^2	$x_i \cdot Y_i$	Y_i^2	\hat{Y}_i	e_i	e_i^2
1	0	56751	0	0	$3,2207e + 9$	56737	13,574	184,3
2	0,29870	57037	0,08922	17037	$3,2532e + 9$	56954	83,482	6969,3
3	0,46480	56979	0,21604	26484	$3,2466e + 9$	57074	-94,681	8964,5
4	0,57620	57074	0,33201	32886	$3,2574e + 9$	57154	-80,272	6443,6
5	0,83860	57422	0,70325	48154	$3,2973e + 9$	57344	77,897	6067,9
Σ	2,1783	285263	1,3405	124561	$1,6275e + 10$	0	$1,6e - 10$	28630

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n Y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i Y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} =$$

$$= \frac{285263 \cdot 1,3405 - 2,1783 \cdot 124561}{5 \cdot 1,3405 - 2,1783^2} = 56737,426$$

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i Y_i - \sum_{i=1}^n x_i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} =$$

$$= \frac{5 \cdot 124561 - 2,1783 \cdot 285263}{5 \cdot 1,3405 - 2,1783^2} = 723,440$$

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n Y_i^2 - \hat{\beta}_0 \sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n x_i Y_i =$$

$$28630,$$

$$\hat{\beta}_0 + \hat{\beta}_1 x = 56737,426 + 723,440x,$$

$$S_e = 2,8630e + 004, \quad S^2 = \frac{S_e}{n-2} = \frac{2.8630e + 004}{5-2} = 9,5432e + 003, \quad S = 97,689.$$

Příklad: měření poledníku

Test $H_0 : \beta_1 = 0$ proti $H_a : \beta_1 \neq 0$ je založen na statistice $T_1 = b_1 \sqrt{\sum_{i=1}^n x_i^2 - n\bar{x}^2} / s = 723,4401 \sqrt{1,3405 - 5 \cdot 0,4357^2} / \sqrt{9543,2} = 11,8353$.

Statistika $|T_1| > t_{n-3, 1-\alpha/2} = t_{5-3, 1-0,05/2} = 4,3027$. Nulovou hypotézu zamítáme.

Příklad 8.14 (mravenec průzkumník a vyšší teplota)



Přepočtete předchozí úlohu, když navíc uvažujete šesté měření při teplotě 30° C s naměřenou rychlostí mravence průzkumníka 140 m/hod.



Najděte regresní přímku $\hat{\beta}_0 + \hat{\beta}_1 x$ pro závislost rychlosti y mravence průzkumníka na teplotě okolí x . Určete index determinace. Zkonstruuje pás spolehlivosti pro regresní přímku a okolo regresní přímky. Při výpočtu užíjte tabulek z předchozího příkladu.

Výsledky uvádí tabulka

i	x_i	Y_i	x_i^2	$x_i \cdot Y_i$	Y_i^2	\hat{Y}_i	e_i	e_i^2
1	10	18	100	180	324	25,11	-7,105	50,488
2	15	54	225	810	2916	69,76	-15,765	248,530
3	20	126	400	2520	15876	114,42	11,576	134,000
4	25	210	625	5250	44100	159,08	50,916	2592,488
5	28	190	784	5320	36100	185,88	4,121	16,982
6	30	140	900	4800	19600	203,74	-43,743	1913,438
Σ	128	758	3034	18880	124916	598	0	4955,9

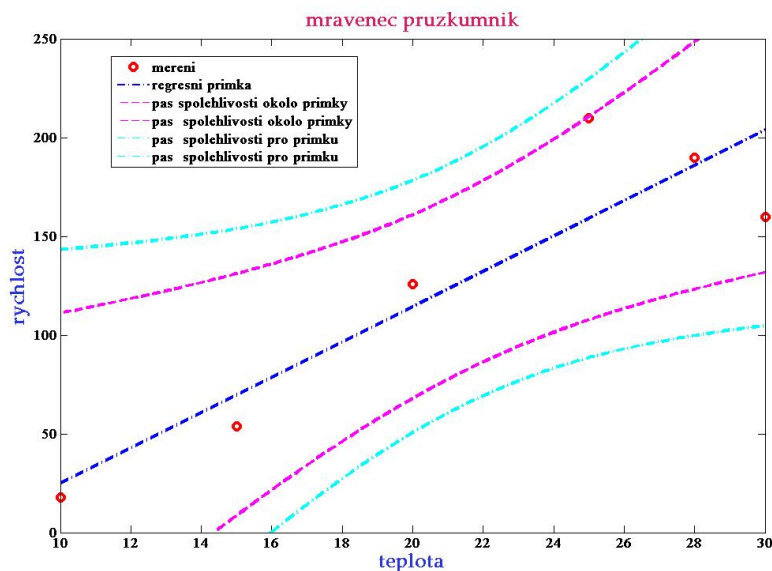
$$\begin{aligned} \hat{\beta}_0 &= \frac{\sum_{i=1}^n Y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i Y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \\ &= \frac{758 \cdot 3034 - 128 \cdot 18880}{6 \cdot 3034 - 128^2} = \frac{-116868}{1838} = -64,2132 \end{aligned}$$

$$\begin{aligned} \hat{\beta}_1 &= \frac{n \sum_{i=1}^n x_i Y_i - \sum_{i=1}^n x_i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \\ &= \frac{6 \cdot 124916 - 128 \cdot 758}{6 \cdot 3034 - 128^2} = \frac{16256}{1838} = 8,9319, \end{aligned}$$

$$\begin{aligned} S(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^n Y_i^2 - \hat{\beta}_0 \sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n x_i Y_i = \\ &= 124916 + 64,2132 \cdot 758 - 8,9319 \cdot 18880 = 4955,9. \end{aligned}$$

$$\hat{\beta}_0 + \hat{\beta}_1 x = -64,2132 + 8,9319x$$

$$S_e = 4955,9, \quad S^2 = \frac{S_e}{n-2} = \frac{4955,9}{6-2} = 1238,98, \quad S = 35,199.$$



8.4. Kvadratická regrese

Věta 8.15 V modelu kvadratické regrese $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ lze odhad $\hat{\beta}$ získat opět z maticového vztahu

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \quad (36)$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{pmatrix}, \quad (37)$$

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \\ \sum x_i^2 Y_i \end{pmatrix}.$$

Příklad 8.16

Uvažujme příklad mravenec průzkumník a vyšší teplota. Najdeme kvadratickou aproximaci.

V našem případě dostáváme

$$\mathbf{X} = \begin{pmatrix} 1 & 10 & 100 \\ 1 & 15 & 225 \\ 1 & 20 & 400 \\ 1 & 25 & 625 \\ 1 & 28 & 784 \\ 1 & 30 & 900 \end{pmatrix}, \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} 6 & 128 & 3034 \\ 128 & 3034 & 76952 \\ 3034 & 76952 & 2035906 \end{pmatrix}, \quad (38)$$

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 738 \\ 18280 \\ 470560 \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} -248,4195 \\ 30,3290 \\ -0,5450 \end{pmatrix}.$$

Pro reziduální součet čtverců platí

$$S_e = S(b_0, b_1, b_2) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n Y_i^2 - b_0 \sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n x_i Y_i - b_2 \sum_{i=1}^n x_i^2 Y_i.$$

Pro odhad disperze pak platí $s^2 = \frac{S_e}{n-3}$.

V našem případě dostáváme

$$\hat{\mathbf{Y}} = \begin{pmatrix} 0,3684 \\ 83,8857 \\ 140,1519 \\ 169,1671 \\ 173,4957 \\ 170,9312 \end{pmatrix}, \quad S_e = 4300,8, \quad s^2 = \frac{S_e}{6-3} = 86,0947.$$

Proveďte také testy:

a) $H_0 : \beta_1 = 0$ proti $H_a : \beta_1 \neq 0$

b) $H_0 : \beta_2 = 0$ proti $H_a : \beta_2 \neq 0$,

c) $H_0 : (\beta_1, \beta_2)' = \mathbf{0}$ proti $H_a : (\beta_1, \beta_2)' \neq \mathbf{0}$.

Test $H_0 : \beta_2 = 0$ proti $H_a : \beta_2 \neq 0$ (test linearity regrese proti alternativě kvadratické regrese) je založen na statistice $T_2 = \frac{b_2}{\sqrt{s^2 v_{22}}} = \frac{30,3290}{\sqrt{86,0947 \cdot 1,1255 \cdot 10^{-4}}} = -5,5366$

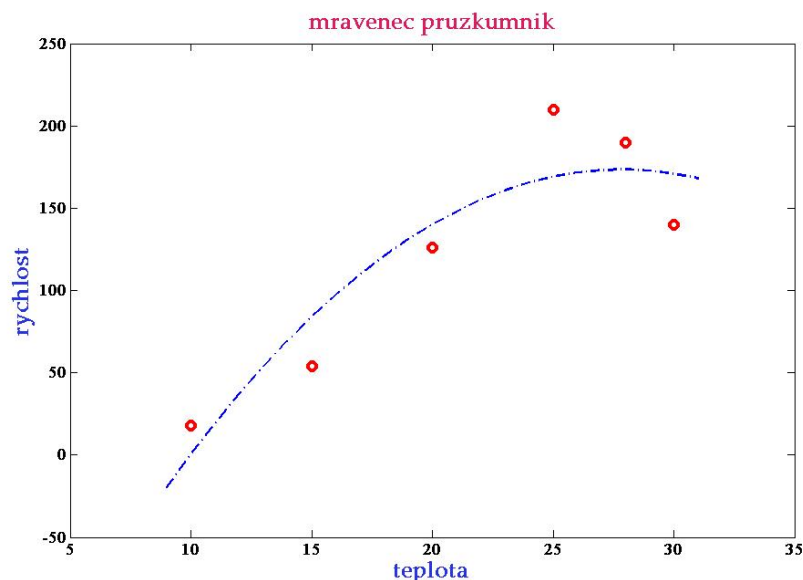
Statistika $|T_2| > t_{n-3, 1-\alpha/2} = t_{6-3, 1-0,05/2} = 3,1824$. Nulovou hypotézu zamítáme.

Test $H_0 : (\beta_1, \beta_2)' = \mathbf{0}$ proti $H_a : (\beta_1, \beta_2)' \neq \mathbf{0}$ je založen na statistice

$$Z = \frac{1}{2s^2} (b_1, b_2) \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}^{-1} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

Statistika $|Z| = 138,4594 > F_{2, n-3, 1-\alpha} = F_{2, 3, 1-0,05} = 9,5521$. Nulovou hypotézu zamítáme.

Závěr: zamítáme hypotézu $H_0 : (\beta_1, \beta_2)' = \mathbf{0}$.



8.5. Regrese se dvěma nezávislými proměnnými

Najděte odhady parametry β_0 , β_1 a β_2 tak, aby platilo $Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i + \varepsilon_i$.

Věta 8.17 V uvedeném modelu regrese se dvěma nezávisle proměnnými lze odhad $\hat{\beta}$ získat opět z maticového vztahu

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \quad (39)$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \dots & \dots & \dots \\ 1 & x_n & z_n \end{pmatrix}, \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum x_i & \sum z_i \\ \sum x_i & \sum x_i^2 & \sum x_i z_i \\ \sum z_i & \sum x_i z_i & \sum z_i^2 \end{pmatrix}, \quad (40)$$

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \\ \sum z_i Y_i \end{pmatrix}.$$

Příklad 8.18 (savci)



Pokusíme se u savců popsat závislost podílu tělesné vahy potomků po narození a váhy matky (ozn. Y , v %) na délce těla matky (ozn. x , v metrech) a době březosti (ozn. z , v měsících).

i	savec	Y_i	x_i	z_i
1.	hraboš	53,00%	0,14	0,66
2.	malý pes	30,00%	0,50	2,07
3.	antilopa	10,00%	1,90	10,50
4.	šimpanz	4,25%	1,30	8,30
5.	plejtvák	1,00%	25,00	12,00

V našem případě dostáváme

$$\mathbf{X} = \begin{pmatrix} 1 & 0,14 & 0,66 \\ 1 & 0,50 & 2,07 \\ 1 & 1,90 & 10,50 \\ 1 & 1,30 & 8,30 \\ 1 & 25,00 & 12,00 \end{pmatrix}, \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} 5,00 & 28,84 & 33,53 \\ 28,84 & 630,57 & 331,87 \\ 33,53 & 331,87 & 327,87 \end{pmatrix}, \quad (41)$$

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 98,25 \\ 71,95 \\ 249,36 \end{pmatrix} \quad \hat{\beta} = \begin{pmatrix} 49,9601 \\ 0,2003 \\ -4,2448 \end{pmatrix}. \quad \text{prvky matice } (\mathbf{X}'\mathbf{X})^{-1} \text{ ozn. } v_{ij}$$

Pro reziduální součet čtverců platí

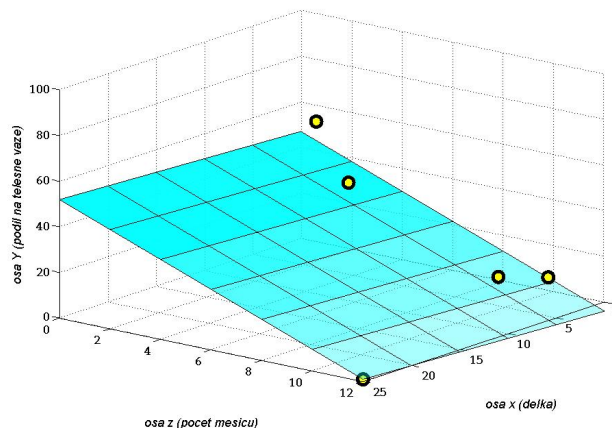
$$S_e = S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n Y_i^2 - \beta_1 \sum_{i=1}^n x_i Y_i - \beta_2 \sum_{i=1}^n z_i Y_i.$$

Pro odhad disperze pak platí

$$s^2 = \frac{S_e}{n-3}.$$

V našem případě dostáváme

$$\hat{\mathbf{Y}} = \begin{pmatrix} 44,1865 \\ 38,2735 \\ 2,7703 \\ 11,9887 \\ 1,0309 \end{pmatrix}, \quad S_e = 258,2841, \quad s^2 = \frac{S_e}{5-3} = 129,1421.$$



$$Y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z = 49,9601 + 0,2003x - 4,2448z.$$

U ženy s výškou 165 cm, dostáváme odhad $\hat{Y} = 49,9601 + 0,2003 \cdot 1,65 - 4,2448 \cdot 9 = 12,08\%$, což je velký podíl tělesné váhy matky a dítěte — odhad nasvědčuje na dvojčata.

Pozn.: Při hledání odhadů neznámých parametrů regresní přímky je třeba vždy posoudit, zda lze opravdu závislost zvoleným modelem aproximovat. Realita může být mnohem složitější, závislá proměnná může být závislá na dalších faktorech. Snaha vše popsat vzorcem může vést k tragikomickým závěrům.

Vhodnost modelu je možné ověřit pomocí vhodných testů.

a) Test $H_0 : \beta_1 = 0$ proti $H_a : \beta_1 \neq 0$ je založen na statistice $T_1 = \frac{\hat{\beta}_1}{\sqrt{s^2 v_{11}}} = \frac{0,2003}{\sqrt{129,1421 \cdot 0,0036}} = 0,2940$.

Statistika $|T_1| < t_{n-3, 1-\alpha/2} = t_{5-3, 1-0,05/2} = 4,3027$. Nulovou hypotézu nezamítáme.

b) Test $H_0 : \beta_2 = 0$ proti $H_a : \beta_2 \neq 0$ je založen na statistice $T_2 = \frac{\hat{\beta}_2}{\sqrt{s^2 v_{22}}} = \frac{-4,2448}{\sqrt{129,1421 \cdot 0,0162}} = -2,3942$.

Statistika $|T_2| < t_{n-3, 1-\alpha/2} = t_{5-3, 1-0,05/2} = 4,3027$. Nulovou hypotézu nezamítáme.

c) Test $H_0 : (\beta_1, \beta_2)' = \mathbf{0}$ proti $H_a : (\beta_1, \beta_2)' \neq \mathbf{0}$ je založen na statistice

$$Z = \frac{1}{2s^2} (\hat{\beta}_1, \hat{\beta}_2) \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}.$$

Statistika $|Z| = 6,3464 < F_{2, n-3, 1-\alpha} = F_{2, 2, 1-0,05} = 19$. Nulovou hypotézu nezamítáme.

Závěr: nelze zamítnout hypotézu $H_0 : (\beta_1, \beta_2)' = \mathbf{0}$.

8.6. Regrese – periodická funkce

Příklad 8.19 V tabulce jsou uvedeny průměrné měsíční teploty Y v Dunaji v Bratislavě.

X_i	1	2	3	4	5	6	7	8	9	10	11	12
Y_i	1,3	1,9	6,0	9,7	14,6	17,6	19,9	18,4	14,9	10,2	6,0	3,5



Najděte aproximaci trigonometrickým polynomem prvního stupně ve tvaru

$$Y_i = \beta_0 + \beta_1 \sin\left(\frac{\pi}{6}x_i\right) + \beta_2 \cos\left(\frac{\pi}{6}x_i\right) + \epsilon_i \quad i = 1, \dots, 12.$$

Úlohu převedeme na regresi se dvěma nezávislými proměnnými. Zavedeme substituci $t_i = \sin(\frac{\pi}{6}x_i)$ a $z_i = \cos(\frac{\pi}{6}x_i)$.

V uvedeném modelu regrese se dvěma nezávisle proměnnými lze odhad $\hat{\beta}$ získat opět z maticového vztahu

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \quad (42)$$

$$\mathbf{X} = \begin{pmatrix} 1 & t_1 & z_1 \\ 1 & t_2 & z_2 \\ \dots & \dots & \dots \\ 1 & t_n & z_n \end{pmatrix} = \begin{pmatrix} 1 & \sin(\frac{\pi}{6}) & \cos(\frac{\pi}{6}) \\ 1 & \sin(\frac{\pi}{3}) & \cos(\frac{\pi}{3}) \\ \dots & \dots & \dots \\ 1 & \sin(2\pi) & \cos(2\pi) \end{pmatrix}.$$

V našem případě

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 12 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 6 \end{pmatrix}, \quad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} 124 \\ -28,6224 \\ -45,6559 \end{pmatrix}.$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0,0833 & 0 & 0 \\ 0 & 0,1667 & 0 \\ 0 & 0 & 0,1667 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 124 \\ -28,6224 \\ -45,6559 \end{pmatrix}.$$

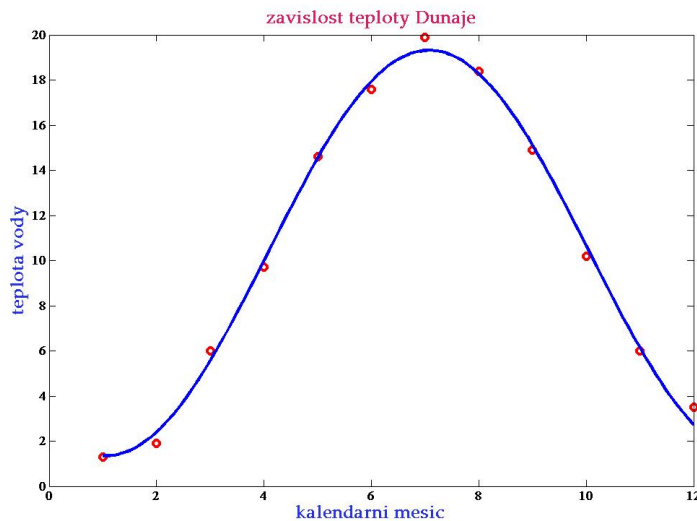
Odtud po dosazení dostáváme

$$\hat{\beta} = \begin{pmatrix} 10,3333 \\ -4,7704 \\ -7,6093 \end{pmatrix}$$

$$s^2 = \frac{S_e}{n-3} = \frac{\sum_{i=1}^n Y_i^2 - b_0 \sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n t_i Y_i - b_2 \sum_{i=1}^n z_i Y_i}{n-3} = 0,2107$$

Regresní funkce je

$$Y = 10,3333 - 4,7704 \sin\left(\frac{\pi}{6}x\right) - 7,6093 \cos\left(\frac{\pi}{6}x\right).$$



a) Test $H_0 : \beta_1 = 0$ proti $H_a : \beta_1 \neq 0$ je založen na statistice $T_1 = \frac{\beta_1}{\sqrt{s^2 v_{11}}} = -25,458$.

Statistika $|T_1| > t_{n-3,1-\alpha/2} = t_{12-3,1-0.05/2} = 2,2622$. Nulovou hypotézu zamítáme.

b) Test $H_0 : \beta_2 = 0$ proti $H_a : \beta_2 \neq 0$ je založen na statistice $T_2 = \frac{\beta_2}{\sqrt{s^2 v_{22}}} = -40,609$

Statistika $|T_2| > t_{n-3,1-\alpha/2} = t_{12-3,1-0.05/2} = 2,2622$. Nulovou hypotézu zamítáme.

c) Test $H_0 : (\beta_1, \beta_2)' = \mathbf{0}$ proti $H_a : (\beta_1, \beta_2)' \neq \mathbf{0}$ je založen na statistice

$$Z = \frac{1}{2s^2}(\hat{\beta}_1, \hat{\beta}_2) \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}.$$

Statistika $|Z| = 1148,6 > F_{2,n-3,1-\alpha} = F_{2,9,1-0.05} = 4,2565$. Nulovou hypotézu zamítáme.

Závěr: nelze zamítnout hypotézu $H_0 : (\beta_1, \beta_2)' = \mathbf{0}$.

9

Nelineární regrese

9.1. Linearizace

V případě nelineárního modelu $\mathbf{Y} = \phi(\mathbf{x}, \beta)$ nelze výpočet realizovat přímo metodou z předchozí kapitoly. Nicméně lze užít Taylorův rozvoj nelineární funkce v bodě $\beta^{(0)}$, kde $\beta^{(0)}$ je přibližné řešení aproximační úlohy.

Předpokládáme, že měření — náhodný vektor \mathbf{Y} vyhovuje nelineárnímu regresnímu modelu

$$\mathbf{Y} = \phi(\beta) + \varepsilon, \quad \varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{V}), \quad (43)$$

kde $\phi : \mathbb{R}^k \rightarrow \mathbb{R}^n$ je známá nelineární funkce, $\beta \in \mathbb{R}^k$ je neznámý parametr, \mathbb{R}^k je parametrický prostor a \mathbf{V} je známá pozitivně definitní matice. Dále uvažujme bod $\beta^0 \in \mathbb{R}^k$ a jeho okolí $\mathcal{O}(\beta^0)$ v parametrickém prostoru \mathbb{R}^k takové, že skutečná hodnota parametru β se nachází uvnitř $\mathcal{O}(\beta^0)$.

Dalšími předpoklady pro náš model (43) jsou:

- (1) model je regulární v bodě β^0 , t.j. $r(\mathbf{F}) = k$ kde $\mathbf{F} = \frac{\partial \phi(\beta)}{\partial \beta'}|_{\beta^0}$;
- (2) pro libovolné $\beta \in \mathcal{O}(\beta^0)$ and $\forall i, j, l \in 1, \dots, k : \frac{\partial^3 \phi(\beta)}{\partial \beta_i \beta_j \beta_l} = 0$.

Uvedené předpoklady na model (43) vedou k tomu, že parametrický prostor \mathbb{R}^k může být zúžen na množinu $\mathcal{O}(\beta^0)$, a model může být aproximován kvadratickým modelem, t.j.

$$\mathbf{Y} - \phi_0 \sim N_n(\mathbf{F}\delta\beta + \frac{1}{2}\kappa(\delta\beta), \sigma^2 \mathbf{V}), \quad \beta \in \mathcal{O}(\beta^0) \quad (44)$$

kde

$$\phi_0 = \phi(\beta^0), \quad \kappa(\delta\beta) = \begin{pmatrix} \kappa_1(\delta\beta) \\ \vdots \\ \kappa_n(\delta\beta) \end{pmatrix},$$

$$\kappa_i(\delta\beta) = (\beta - \beta^0)' h_i (\beta - \beta^0)', \quad h_i = \frac{\partial^2 \phi_i(\beta^0)}{\partial \beta' \partial \beta'}, \quad i = 1, \dots, n.$$

Linearizací kvadratického modelu (44) zanedbáváme členy Taylorova rozvoje druhého a vyššího řádu. Výsledný linearizovaný model má tvar

$$\mathbf{Y} \sim N_n(\phi_0 + \mathbf{F}\delta\beta, \sigma^2 \mathbf{V}), \quad (45)$$

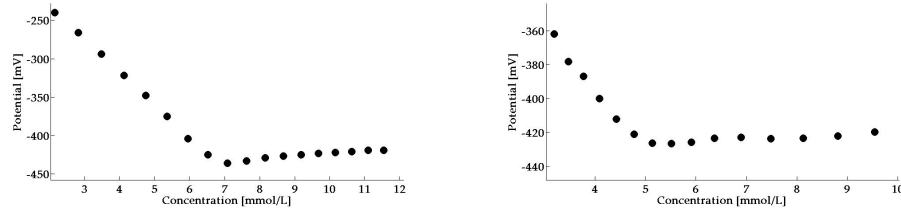
kde $\phi_0 = \phi(\beta^0)$, $\mathbf{F} = \frac{\partial \phi(\beta)}{\partial \beta'}|_{\beta^0}$ and $\delta\beta = \beta - \beta^0$. Za speciálních okolností $\beta^0 = \mathbf{0}$ a $\phi(\beta^0) = \mathbf{0}$ můžeme lineární model zapsat

$$\mathbf{Y} \sim N_n(\mathbf{F}\beta, \sigma^2 \mathbf{V}). \quad (46)$$

Je důležité si uvědomit, že volba počátečního parametru β^0 hraje významnou roli v celém procesu linearizace. Nepřijatelný odhad parametru β může být způsoben málo kvalitním počátečním řešením. Z těchto důvodů se určuje tzv. linearizační oblast pro difference $\delta\beta$, které jsou akceptovatelné při linearizaci modelu.

9.2. Příklad

Mějme k dispozici měření napětí (osa y) při různých koncentracích roztoku x znázorněné na obrázcích.



9.2.1. Model A:

Uvažujme, že závislost mezi potenciálem a koncentrací je dána funkcí

$$f_A(\beta, x) = \beta_1 x + \beta_2 + \beta_3 |x - \beta_4|, \quad (47)$$

kde proměnná x reprezentuje koncentraci a $\beta_1, \beta_2, \beta_3, \beta_4$ jsou neznámé parametry. Parametr β_4 má speciální význam, určuje bod zlomu. Uvažujeme a to je velmi důležité si uvědomit, že hodnoty koncentrací jsou dány deterministicky (tedy jsou měřeny bezchybně).

Všechna měření napětí vytváří n -dimenzionální náhodný vektor \mathbf{Y} , pro který uvažujeme nelineární regresní model

$$\mathbf{Y} = \phi(\beta) + \varepsilon, \quad \varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{V}), \quad (48)$$

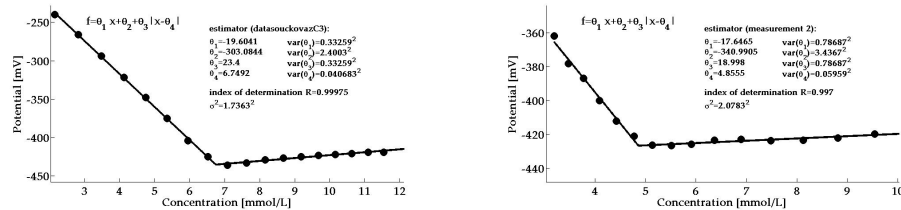
kde $\phi_i(\beta) = f_A(\beta, x_i) \quad \forall i = 1, \dots, n$, x_i jsou hodnoty odpovídající měření Y_i . Nejistota měřených dat je popsána směrodatnou odchylkou 0,5 [mV], takže kovarianční matice má tvar $\sigma^2 \mathbf{V} = 0,25 \cdot \mathbf{I}_n$.

V dalším kroku provedeme linearizaci modelu (48). Použijeme postup popsany v předchozí podkapitole. Uvažujme počáteční řešení $\beta^0 = \mathbf{0}$ a $\phi(\beta^0) = \mathbf{0}$, pak náš linearizovaný model má tvar

$$\mathbf{Y} \sim N_n(\mathbf{F}\beta, \sigma^2 \mathbf{V}),$$

kde

$$\begin{aligned} \mathbf{F}_i &= \frac{\partial f_A(x_i, \beta^0)}{\partial \beta'} = \left(\frac{\partial f_A(x_i, \beta^0)}{\partial \beta_1}, \frac{\partial f_A(x_i, \beta^0)}{\partial \beta_2}, \frac{\partial f_A(x_i, \beta^0)}{\partial \beta_3}, \frac{\partial f_A(x_i, \beta^0)}{\partial \beta_4} \right) = \\ &= (x_i, 1, |x_i - \beta_4^0|, |x_i - \beta_4^0|(-\beta_3^0)), \quad i = 1, \dots, n. \end{aligned} \quad (49)$$



Výpočet neznámých parametrů je založen na vztazích z části 7.6.4, zejména na

$$\hat{\beta} = (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\mathbf{Y}, \quad (50)$$

$$\begin{aligned} \text{var}(\hat{\beta}) &= \text{var}[(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\mathbf{Y}] = (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'[\text{var}(\mathbf{Y})]\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1} = \\ &= (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'(\sigma^2 \mathbf{I})\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1} = \sigma^2(\mathbf{F}'\mathbf{F})^{-1}. \end{aligned} \quad (51)$$

Měřená data pro dvě různá měření a jejich aproximace zvolenou funkcí jsou ilustrovány na obrázku.

Další možností při aproximaci dat nelineární funkcí je použít Levenberg–Marquardtův algoritmus.

10

ANOVA

10.1. Jednoduché třídění

Mějme realizace náhodných veličin

$$Y_{11}, Y_{12}, \dots, Y_{1n_1}$$

$$Y_{21}, Y_{22}, \dots, Y_{2n_2}$$

...

$$Y_{k1}, Y_{k2}, \dots, Y_{kn_k}$$

které jsou nezávislé a pro které platí $Y_{ij} \sim N(\mu_i, \sigma^2)$. Uvažujeme tedy, že všech k výběrů má shodný rozptyl, střední hodnota může být odlišná.

Označme $\bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ průměr v i -tém výběru, $\bar{Y}_{.j} = \frac{1}{k} \sum_{i=1}^k Y_{ij}$ průměr j -tého sloupce, $\bar{Y}_{..} = \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$ průměr ze všech pozorování.

Definujme reziduální součet čtverců $SS_e = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$, celkový součet čtverců $SS_T = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$.

Pro vektor

$$\mathbf{d} = \hat{Y} - \hat{Y}_0 = \begin{pmatrix} (\bar{Y}_{1.} - \bar{Y}_{..})_{n_1} \\ \vdots \\ (\bar{Y}_{k.} - \bar{Y}_{..})_{n_k} \end{pmatrix}$$

platí

$$\|\mathbf{d}\|^2 = SS_A = SS_T - SS_e = \sum_{i=1}^k n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2.$$

Zde jsme použili symbol SS_A pro součet čtverců vysvětlený faktorem A .

Testujme hypotézu $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ proti alternativě $H_a : i, j \in \{1, 2, \dots, k\} : \mu_i \neq \mu_j$.

Testovou statistikou je

$$F = \frac{\frac{SS_A}{(k-1)}}{\frac{SS_e}{(n-1)}} = \frac{MS_A}{MS_e}.$$

Hodnota MS_X se nazývá průměrný čtverec.

Tato testová statistika má Fisher-Snedecorovo rozdělení o $(k-1, n-k)$ stupních volnosti. Obvykle se celý proces výpočtu zapisuje v tabulce analýzy rozptylu jednoduchého třídění.

ANOVA table – v Matlabu příkaz `anova1`

zdroj variability	součet čtverců	stupně volnosti	průměrné čtverce	testová statistika
řádky (ošetření)	SS_A	$k-1$	$MS_A = SS_A/(k-1)$	F
reziduální	SS_e	$n-k$	-	-
celkový	SS_T	$n-1$	-	-

Pokud zamítneme nulovou hypotézu o shodě středních hodnot, zajímáme se o řádky, které způsobily její zamítnutí. K tomu se užívá několik metod (Schéffe, Tukey).

Ukážeme si nyní použití Bonferroniho metody mnohonásobného porovnání. Tato metoda považuje za různé ty výběry, které vyhovují nerovnici $|\bar{Y}_{i.} - \bar{Y}_{j.}| \geq t_{n-k, 1-\frac{\alpha}{m}} \sqrt{\frac{SS_e}{n-k} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$.

Symbol m zde představuje počet všech možných porovnávaných dvojic, tedy $m = k(k-1)/2$.

10.1.1. Bartlettův test

Před použitím popsané testové statistiky je třeba otestovat shodnost rozptylů jednotlivých výběrů. Většinou se používá Bartlettův test. Alternativou je použití Leveneova testu, Cochranova testu, Hartleyova testu nebo Brown-Forsythova testu.

Označme

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2,$$

$$S^2 = \frac{1}{n - k} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 = \sum_{i=1}^k \frac{n_i - 1}{n - k} S_i^2.$$

Bartlettův test má testovou statistiku

$$B = \frac{1}{C} \left((n - k) \log S^2 - \sum_{i=1}^k (n_i - 1) \log S_i^2 \right) =$$

$$= \frac{n - k}{C} \left(\log S^2 - \sum_{i=1}^k \frac{n_i - 1}{n - k} \log S_i^2 \right), \text{ kde}$$

$$C = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{n - k} \right).$$

Testová statistika $B \sim \chi_{k-1}^2$. Za předpokladem použití tohoto testu bývá považován rozsah výběrů $n_i \geq 6$.

Pokud zamítneme nulovou hypotézu o shodě rozptylů nelze pomocí ANOVY testovat shodu středních hodnot. V takovém případě používáme místo ANOVY neparametrický Kruskal-Wallisův test.

10.1.2. Příklad

Cílem úlohy je ověřit, zda se liší produkce semen smrku v rámci dané populace. K zodpovězení této otázky byl náhodně vybrán určitý počet stromů a u každého z nich se stanovil počet semen X_{ij} v náhodně vybraných šiškách.

strom č.	
1.	33 35 36 34 35 36
2.	36 38 35 35 35 37
3.	39 35 37 36 37 38
4.	42 43 46 42 44 42
5.	44 45 42 44 41 43
6.	39 40 43 39 41 42

i	n_i	\bar{X}_i	s_i^2	$\ln s_i^2$
1.	6	209/6=34.833	1.3667	0.3124
2.	6	216/6=36.000	1.6000	0.4700
3.	6	222/6=37.000	2.0000	0.6931
4.	6	259/6=43.167	2.5667	0.9426
5.	6	259/6=43.167	2.1667	0.7732
6.	6	244/6=40.667	2.6667	0.9808
	48	1409/6=39.139	2.0611	0.7232

Otestujeme nejprve pomocí Bartlettova testu shodu rozptylů. Dostáváme

$$C = 1 + \frac{1}{3(6-1)} \left(\sum_{i=1}^6 \frac{1}{6-1} - \frac{1}{36-6} \right) = 1 + \frac{1}{15} 6 \left(\frac{1}{5} - \frac{1}{30} \right) = 1,067.$$

Hodnota C bývá jen o málo větší než 1. Testová statistika má hodnotu

$$B = \frac{1}{1,067} \left((36-6) \log 2,0611^2 - \sum_{i=1}^6 (6-1) \log S_i^2 \right) = 1,056.$$

Hodnota testovací statistiky je menší než kritická hodnota $\chi_{6-1}^2(1-0,05) = 11,07$. Nezamítáme nulovou hypotézu o shodě rozptylů v jednotlivých výběrech. Můžeme tedy použít metodu analýzy rozptylu.

Závěrečná tabulka:

ANOVA table – v Matlabu příkaz `anova1`

Source	SS	df	MS	F
Rows	406.47	5	81.2944	39.442*
Error	61.83	30	2.0611	
Total	468.00	35		

Príslušný kvantil na zvolené hladině významnosti $\alpha = 5\%$ je $F_{k-1, n-k}(\alpha) = F_{5,30}(0.05) = 2,5336$.

Výsledek analýzy variance je skutečně vysoce průkazný. Lze tedy zamítnout nulovou hypotézu, že produkce semen se v populaci smrku významně neliší.

10.2. Dvojné třídění bez interakcí

Zdrojem variability nemusí být jen řádky (ošetření) ale i sloupce (faktory). V této části budeme testovat vliv řádkových a sloupcových zdrojů variability, což budeme označovat jako dvojné třídění bez interakcí.

Uvažujme, že střední hodnoty $\mu_{ij} = \mu + \alpha_i + \beta_j$.

Testujme hypotézu $H_0 : \alpha_i = 0, \beta_j = 0$ proti alternativě $H_a : \alpha_i \neq 0, \beta_j \neq 0$.

Testovou statistikou je

$$F = \frac{\frac{SS_X}{(k-1)}}{\frac{SS_e}{(n-1)}} = \frac{MS_X}{MS_e}.$$

Hodnota MS_X se nazývá průměrný čtverec. Zápisem MS_X se myslí MS_A resp. MS_B .

$$SS_{AB} = \sum_{i=1}^k \sum_{j=1}^n Y_{ijn_{ij}} (\bar{Y}_{ij} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2.$$

$$\text{Celkový součet čtverců je } SS_T = \sum_{s=1}^{n_{ij}} \sum_{i=1}^k \sum_{j=1}^n (Y_{ijs} - \bar{Y}_{...})^2.$$

Výpočty jsou poměrně náročné na počet operací. Analýza se většinou provádí v softwaru, který má knihovnu pro ANOVU. Výsledek výpočtu bývá prezentován v závěrečné tabulce ANOVY, což je demonstrováno na příkladech.

10.2.1. *Příklad*

Liší se obsah celkového dusíku v keřovitém stepním porostu v závislosti na tom, zda (1) je půda ve svrchním nebo spodním horizontu, (2) zda je v blízkosti keře nebo v otevřeném porostu?

Lokalita	svrchní horizont		spodní horizont	
Půda v blízkosti keře:	56, 48, 60, 49, 52, 67, 62	39, 37, 46, 51, 49, 41, 44		
Půda v otevřeném porostu:	51, 48, 41, 38, 60, 45, 52	28, 27, 38, 16, 31, 22, 18		
Celkem				

Lokalita	svrchní horizont		spodní horizont	
Půda v blízkosti keře:	$X_{11.} = 394,$ $x_{11.} = 56.3$	$X_{12.} = 307,$ $x_{12.} = 43.9$	$X_{1..} = 701,$ $x_{1..} = 50.07$	
Půda v otevřeném porostu:	$X_{21.} = 335,$ $x_{21.} = 47.9$	$X_{22.} = 180,$ $x_{22.} = 25.7$	$X_{2..} = 515,$ $x_{2..} = 36.79$	
	$X_{.1.} = 729,$ $x_{.1.} = 52.07$	$X_{.2.} = 487,$ $x_{.2.} = 34.79$	$X_{...} = 1216,$ $x_{...} = 43.4286$	

Po dlouhém výpočtu určíme $\sum \sum \sum X_{ijk}^2 = 57444$. Pak již můžeme snadno určit S_T a S_A, S_B, S_E .

Závěrečná tabulka:

ANOVA table – v Matlabu příkaz anova2

Source	SS	df	MS	F
Rows	1235.57	1	1235.57	23.62*
Columns	2091.57	1	2091.57	39.98*
Error	1307.72	25	52.31	
Total	4634.86	27		

$$F_{I-1, n-I-J+1}(\alpha) = F_{1,25}(0.05) = 4.2417$$

Závěrečné komentáře:

Jde o příklad s dvěma nezávislými kategoriálními proměnnými (horizont a stanoviště); závislá proměnná je velikost jedinců. Analyzovali jsme dvojnou (dvoucestnou) ANOVOU.

Vliv stanoviště ($F_A = 23.62$, signifikantní), vliv horizontu ($F_B = 39.98$, signifikantní).

Poznámka k založení pokusu: jednoduchou dvoucestnou ANOVOU lze analyzovat jen tehdy, pokud vzorky z obou horizontů jsou nezávislé (čili vzorek z horního horizontu není odebírán na stejném místě jako vzorek ze spodního horizontu).

10.3. **Analýza rozptylu – dvojně třídění s interakcemi**

Zdrojem variability nemusí být jen řádky (ošetření) a sloupce (faktory) ale i jejich interakce.

Uvažujme, že střední hodnoty $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$.

Testujme hypotézu $H_0 : \alpha_i = 0, \beta_j = 0, \gamma_{ij} = 0$ proti alternativě $H_a : \alpha_i \neq 0, \beta_j \neq 0, \gamma_{ij} \neq 0$.

Testovou statistikou je opět stejná

$$F = \frac{MS_X}{MS_e}.$$

Zápisem MS_X se myslí MS_A, MS_B nebo MS_{AB} .

10.3.1. Příklad

Liší se velikost dospělých octomilek v závislosti na výživě a genotypu? Samičky od obou genotypů byly pěstovány na třech druzích výživy. Jedinci následující generace byli změřeni. Hodnoty jsou v relativních jednotkách.

genotyp	výživa 1	výživa 2	výživa 3
1.	18,17,18,16,21,20,18,19	26,25,26,29,22,22,20,23	28,21,29,31,26,25,23,28
2.	13,19,18,15,16,18,14,17	18,16,18,19,17,19,20,18	22,26,17,19,27,21,24,23
Celkem			

	výživa 1	výživa 2	výživa 3	
genotyp 1.	$X_{11.} = 147, x_{11.} = 18.4$	$X_{12.} = 193, x_{12.} = 24.1$	$X_{13.} = 211, x_{13.} = 26.4$	$X_{1..} = 551, x_{1..} = 23.0$
genotyp 2.	$X_{21.} = 130, x_{21.} = 16.3$	$X_{22.} = 145, x_{22.} = 18.1$	$X_{23.} = 179, x_{23.} = 22.4$	$X_{2..} = 454, x_{2..} = 18.9$
	$X_{.1.} = 277, x_{.1.} = 17.3$	$X_{.2.} = 338, x_{.2.} = 21.1$	$X_{.3.} = 390, x_{.3.} = 24.4$	$X_{...} = 1005, x_{...} = 20.9375$

Závěrečná tabulka:

ANOVA table – v Matlabu příkaz anova2					
Source	SS	df	MS	F	Prob>F
Columns	399.875	2	199.938	30.55	0
Rows	196.021	1	196.021	29.95	0
Interaction	30.042	2	15.021	2.3	0.1132
Error	274.875	42	6.545		
Total	900.813	47			

$$F_{I-1, n-I \cdot J}(\alpha) = F_{1, 42}(0, 05) = 4.0727$$

Závěrečné komentáře:

To je příklad s dvěma nezávislými (faktoriálně kombinovanými) kategoriálními proměnnými; závislá proměnná je velikost jedinců. Analyzuje se dvoucestnou ANOVOU.

Vliv genotypu ($F_A = 29.95$), vliv výživy ($F_B = 30.55$) a interakce ($F_{AB} = 2.30$ - ta jediná není signifikantní na ($\alpha = 0.05$)).

Poznámka k založení pokusu: data lze jednoduše takto analyzovat jen tehdy, pokud jsou si všichni jedinci stejně příbuzní. To znamená, že buď všichni jedinci jednoho genotypu musí být potomky jedné samičky (ale pak je riziko, že se netestuje genotyp, ale efekt matky, ať už je to cokoli) nebo od každé samičky každého genotypu se musí vybrat jen jeden potomek (a samiček musí od každého genotypu být hodně). Ve všech ostatních případech by vznikl další faktor "samička", vřazený do faktoru „genotyp“.

11

Řízení jakosti

11.1. Analýza způsobilosti

Způsobilostí výrobního procesu (process capability) se rozumí jeho schopnost trvale dosahovat předem stanovená kritéria kvality. Je samozřejmě vhodné způsobilost vyjadřovat kvantitativně, tedy nějakým číselným ukazatelem. Při konstrukci těchto ukazatelů je vyžadována z hlediska výpočtového jednoduchost, srozumitelnost a dobrá vypovídací schopnost. Dalším požadavkem je univerzálnost. Z tohoto hlediska se takový univerzální ukazatel jen těžce hledá a i když existuje poměrně mnoho ukazatelů pro posuzování způsobilosti procesu, každý je použitelný jen při splnění určitých konkrétních předpokladů. V praxi například nebývá splněn častý požadavek užívaný v matematické statistice: normalita dat. Ve výrobních procesech se často objevuje spíše rovnoměrné rozdělení.

Cílem hodnocení způsobilosti technologického procesu je obvykle sledovat:

- (1) schopnost procesu udržet cílovou hodnotu (target value T) ukazatele kvality,
- (2) míru variability kolem cílové hodnoty.

Pokud se honocení způsobilosti provádí, omezuje se prakticky výhradně na výpočet některého z tzv. indexů způsobilosti. Při komplexním hodnocení způsobilosti by ale měly být provedeny tyto kroky:

- (1) test předpokladů, za kterých lze použít vybrané ukazatele způsobilosti,
- (2) vlastní výpočet vybraných ukazatelů,
- (3) testování významnosti vypočítaných ukazatelů a jejich využití k analýze výrobního procesu.

Pro měřitelné i neměřitelné atributy se způsobilost procesu vyjadřuje procentem výrobků, které vyhovují požadovanému ukazateli kvality. Označíme-li relativní četnost špatných výrobků

$$V = \frac{\text{počet nevyhovujících výrobků mezi sledovanými}}{\text{celkový počet sledovaných výrobků}}, \quad (52)$$

pak způsobilost bude procento vyhovujících výrobků

$$C = 100(1 - V). \quad (53)$$

U tohoto ukazatele není obecně stanovena obecně platná minimální hodnota C . Obvykle přijímaná úroveň je 98-99%.

Předpokládejme, že byl stanoven nejdůležitější znak jakosti sledovaného výrobního procesu a že je tento proces statisticky zvládnut. O znaku jakosti dále předpokládejme, že je to náhodná veličina s normálním rozdělením se střední hodnotou μ a směrodatnou odchylkou σ .

Střední hodnota sledovaného znaku udává hodnotu nastavení (seřízení) výrobního procesu. Je-li toto nastavení správné, měla by být tato hodnota ve středu tolerančního pole. Tedy o střední hodnotě μ by mělo platit:

$$\mu = \frac{USL + LSL}{2},$$

kde USL a LSL jsou mezní hodnoty.

11.2. Indexy způsobilosti

Způsobilost výrobního procesu měříme indexy způsobilosti.

Specifikace výrobního procesu je určena trojicí (\bar{x}, T, s) . Skutečně dosahovaná přesnost je vyjádřena rozptylem, a je známo, že má-li soubor normální rozdělení $N(\mu, \sigma^2)$, pak v intervalu $(\mu - 3\sigma, \mu + 3\sigma)$ leží 99,73% hodnot. Délka tohoto intervalu je 6σ . Porovnáním délky intervalu (\bar{x}, T) a intervalu 6σ získáme představu o poměru předepsané a skutečně dosahované přesnosti. Na tomto principu jsou indexy způsobilosti konstruovány:

$$\text{způsobilost} = \frac{\text{délka intervalu, kde mají být všechny hodnoty}}{\text{délka intervalu, kde jsou všechny hodnoty}}, \quad (54)$$

Nejznámější je index způsobilosti výrobního procesu (Process Capability Index) C_P , který vypočítáme podle následujícího vzorce:

$$C_P = \text{PCI} = \frac{\text{USL} - \text{LSL}}{6 \cdot \sigma} \quad (55)$$

V normě ČSN ISO 8258 se index způsobilosti značí zkratkou anglického názvu PCI.

Koeficient způsobilosti může nabývat různých hodnot:

- 1) $C_p = 1$: To je např. tehdy, když znak jakosti má normální rozdělení, μ je ve středu tolerančního pole, pak 99.73% výrobků, které jsou výsledkem výrobního procesu, je jakostních, neboť pro normální rozdělení platí, že v intervalu $\mu - 3\sigma$, $\mu + 3\sigma$ (interval o šířce 6σ) leží 99.73% hodnot. Pro $C_p = 1$ je šířka tolerančního pole právě rovna 6σ . V normě ČSN ISO 8258 je uvedeno, že proces je stěží způsobilý.
- 2) $C_p > 1$: V tomto případě, je ještě větší procento jakostních výrobků. Platí, že šířka tolerančního pole je větší, než 6σ . Proces je způsobilý. V praxi se požaduje minimální přípustná hodnota $C_p = 1.33$.
- 3) $C_p < 1$: Za této situace je počet výrobků, jejichž znak jakosti leží vně tolerančního pole, větší než 0.27% a výrobní proces je považován za nezpůsobilý. Kromě indexu C_p , který porovnává pouze variabilitu a nekontroluje polohu rozdělení se užívá index C_{pk} , který polohu rozdělení zohledňuje. Tento index je definován jak pro jednostranné, tak pro dvoustranné mezní hodnoty dané technickým předpisem:

Při předepsané dolní mezní hodnotě

$$C_{pk} = C_{pL} = \frac{\mu - \text{LSL}}{3\sigma},$$

Při předepsané horní mezní hodnotě

$$C_{pk} = C_{pU} = \frac{\text{USL} - \mu}{3\sigma},$$

Při předepsané dolní i horní mezní hodnotě

$$C_{pk} = \min \{C_{pL}, C_{pU}\},$$

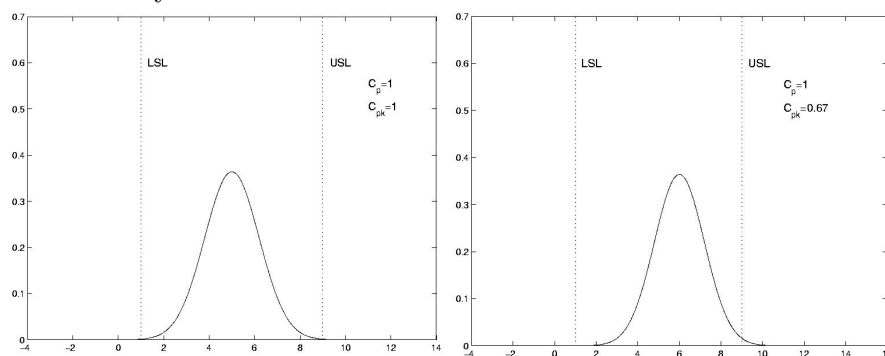
Střední hodnota v těchto vztazích je odhadována jako průměr z výběrových průměrů, tj. $\bar{\bar{X}}$ a σ např. \bar{s} . Ze vzorců pro C_{pk} je patrné, že mění-li se současně variabilita i poloha rozdělení, nemusí dojít ke změně hodnoty C_{pk} , proto je vhodné uvádět současně hodnoty obou koeficientů způsobilosti.

Grafická podoba diagramů je na obrázku. Byl měřen proces se střední hodnotou $\mu = 5$ a zkoumán rozptyl kolem této hodnoty. Regulační meze byly nastaveny na $\mu - 3\sigma$ a $\mu + 3\sigma$, protože v tomto intervalu má proces s normálním rozdělením 99.73% hodnot.

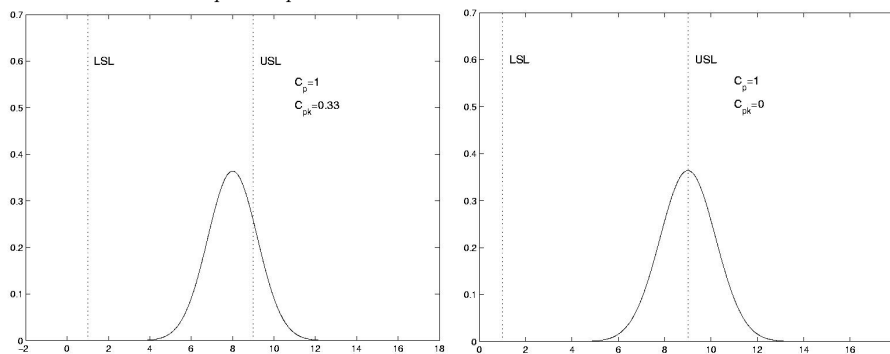
V programech na podporu řízení jakosti je možné se setkat i s indexem způsobilosti C_{pm} (Taguchiho index způsobilosti):

$$C_{pm} = \frac{USL - LSL}{6 \cdot \sqrt{\sigma^2 + (\mu - T)^2}},$$

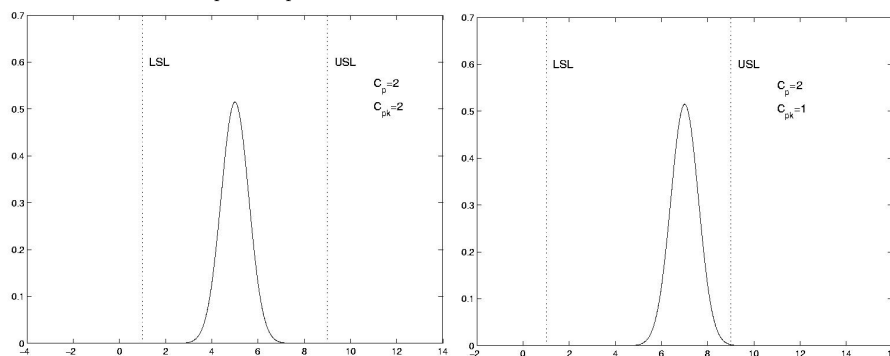
kde T je optimální hodnota, která by měla ležet ve středu tolerančního pole. Tento index tedy zohledňuje nejen variabilitu samotného znaku jakosti ale i rozptyl střední hodnoty znaku kolem optimální hodnoty T .



Různé situace rozdělení hodnot sledovaného znaku jakosti v tolerančním poli a jejich klasifikace pomocí indexu C_p a C_{pk}



Různé situace rozdělení hodnot sledovaného znaku jakosti v tolerančním poli a jejich klasifikace pomocí indexu C_p a C_{pk}



Různé situace rozdělení hodnot sledovaného znaku jakosti v tolerančním poli a jejich klasifikace pomocí indexu C_p a C_{pk}

Příklad

Ve slévárně má být zhotoven výrobek, který obsahuje cílovou hodnotu $T = 15$ jednotek jisté příměsi. Objednatel dále předepsal přípustné dolní a horní tolerance $LSL = 10$ a $USL = 18$ jednotek.

Výrobce se snaží snížit výrobní náklady, příměs je totiž velmi drahá. Bude pro něj výhodné držet obsah příměsi co nejbližší dolní toleranční hranici. Výrobce ví z dřívějších zakázek, že odběratel při testování kvality dodávky zná jediný trik — strategii výpočtu indexu způsobilosti C_{pk} .

Při poradě ředitel, technolog výroby a manažer kvality zvažují vhodnou strategii výroby. Požadavek od odběratele na hodnotu indexu způsobilosti je jako vždy $C_{pk} > 1$.

Výpočet se provádí vztahem

$$C_{pk} = \min \{C_{pU}, C_{pL}\} , \quad (56)$$

$$\text{kde } C_{pU} = \frac{-\mu}{3\sigma} , \quad (57)$$

$$C_{pL} = \frac{\mu -}{3\sigma} .$$

Jelikož i poslední zakázku odběratel bez připomínek převzal, má nyní slévárna prostředky na zakoupení dokonalejšího zařízení. Technolog tak může slíbit, takorčka jakoukoliv disperzi hodnoty příměsi ve výrobcích.

Manažer kvality může demonstrovat, jaký index způsobilosti C_{pk} určí odběratel např. pro následující (hypotetické údaje) výsledků měření:

$$\mathbf{X} = [13,0, 12,5, 12,0, 12,0, 12,0, 12,0, 12,0, 11,5, 11,0]' .$$

Pro výpočet indexu způsobilosti je samozřejmě třeba určit odhad

$$\bar{X} = \hat{\mu} = \frac{\sum_{i=1}^n X_i}{n} = 12,0$$

střední hodnoty μ a

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} = 0,5976$$

disperze σ^2 .

Pro tyto výsledky měření vychází

$$C_{pk} = \min \left\{ \frac{18 - 12}{3 \cdot 0,7731}, \frac{12 - 10}{3 \cdot 0,7731} \right\} = 1,19 > 1,00 .$$

Proces výroby je způsobilý. Tuto pro zákazníka nevýhodnou strategii dodavatele index způsobilosti C_{pk} neodhalí.

11.2.1. Požadavky na způsobilost

Požadavky na způsobilost procesu se většinou vztahují k hodnotě indexu způsobilosti C_{pk} , který charakterizuje reálnou způsobilost procesu udržovat sledovaný znak jakosti v předepsaných tolerančních mezích. Minimální hodnota indexu C_{pk} , při které je proces považován za způsobilý, se s rozvojem technologií zvyšuje. V osmdesátých letech 20. století se ještě vyskytoval požadavek $C_{pk} \geq 1$. V současnosti je proces považován za způsobilý v případě, když hodnota tohoto indexu přesahuje hodnotu $C_{pk} \geq 1,33$.

Proč právě tato hodnota?

Představuje totiž požadavek, aby dosahovaná střední hodnota sledovaného znaku jakosti ležela ve vzdálenosti nejméně 4σ od tolerančních mezí.

Můžeme se také setkat s pojmem *sigma level* a vztahem

$$\sigma_L = \min\{-\mu, \mu-\}/\sigma \quad (58)$$

$$= 3C_{pk} \quad (59)$$

Setkáme-li se pak s požadavkem $\sigma_L > 6$, hovoříme o *six sigma quality*.

Tento přísnější požadavek ($C_{pk} > 2$) na způsobilost procesů zavedla firma Motorola.

Můžeme se setkat i s úpravami vztahů pro výpočet hodnot indexů.

Např. firma Bosch užívá vztah

$$C_g = \frac{0,2(USL - LSL)}{6.\sigma} \quad (60)$$

11.2.2. Odhad podílu zmetků

Podíl zmetků *ppm* vyjadřuje vzhledem k milionu kusů (Part per Million) Pro jednotlivé hodnoty indexu C_p se pak v tabulkách uvádí hodnoty ppm.

například

C_p	0,8	1,0	1,2	1,33	1,4
ppm	16400	2700	320	66	20

Jak tyto hodnoty můžeme určit?

Požadovanou hodnotou indexu způsobilosti C_p je již stanoveno přípustné procento zmetků *NC* (non conforming product).

Podíl *NC* se určí takto:

pro symetrickou toleranci

$$NC = 2\Phi(-3C_p), \quad (61)$$

pro nesymetrickou toleranci

$$NC = \Phi(-3C_{pL}) + \Phi(-3C_{pU}). \quad (62)$$

kde Φ je distribuční funkce standardizovaného normálního rozdělení $N(0, 1)$.

Jak k těmto vztahům dojdeme?

Pravděpodobnost výskytu neshodných výrobků lze pro případ oboustranných tolerančních mezí vyjádřit vztahem

$$P = P(X <) + P(X >)$$

Pro případ splnění normality sledovaného znaku jakosti lze pravděpodobnost výskytu neshodných výrobků odhadnout pomocí vztahu

$$P = \Phi\left(\frac{-\mu}{\sigma}\right) + \Phi\left(\frac{\mu-}{\sigma}\right)$$

Po úpravě dostáváme

$$P = \Phi(-3C_{pL}) + \Phi(-3C_{pU})$$

který můžeme upravit na

$$P = \Phi(-3C_{pk}) + \Phi(3C_{pk} - 6C_p)$$

Ve speciálním případě, kdy střední hodnota sledovaného znaku jakosti leží právě ve středu tolerančního pole a hodnoty indexů C_{pk} a C_p se rovnají, lze pravděpodobnost výskytu neshodných výrobků vyjádřit pomocí vztahu

$$P = 2\Phi(-3C_{pk}) = 2\Phi(-3C_p)$$

Například pro $C_p = 1$ je

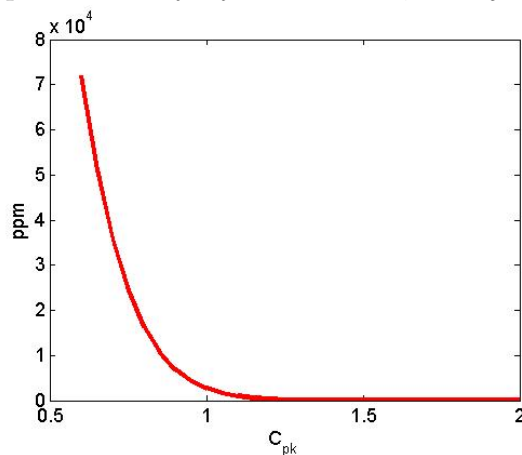
$$NC = 2\Phi(-3 \cdot 1) = 2(1 - \Phi(3)) = 2(1 - 0,99865) = 0,0027.$$

Tedy při $C_p = 1$ je pravděpodobnost výskytu zmetků 0,27% tj. 2700 PPM.

Pro $C_p = 0,8$ je

$$NC = 2\Phi(-3 \cdot 0,8) = 2(1 - \Phi(2,4)) = 2(1 - 0,9918) = 0,0164.$$

Tedy při $C_p = 1$ je pravděpodobnost výskytu zmetků 1,64% tj. 16 400 PPM.



11.2.3. Nejednoznačnost indexů

Bližší analýza uvedených indexů způsobilosti ukazuje, že žádný z nich jednoznačně neurčuje skutečné rozdělení sledovaného znaku jakosti — jedním číslem nelze vyjádřit dva parametry normálního rozdělení.

Nejvíce je to patrné v případě indexu C_p , který nezohledňuje polohu sledovaného znaku jakosti.

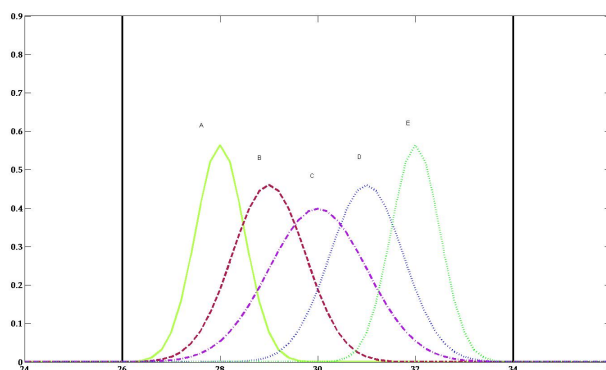
V případě dalších uvedených indexů je tato nejednoznačnost dána tím, že jejich hodnoty závisí jak na poloze, tak na variabilitě sledovaného znaku jakosti.

V případě oboustranné tolerance se u těchto indexů navíc projevuje další nejednoznačnost daná tím, že z jejich hodnoty nelze určit, ke které z tolerančních mezí se proces více přibližuje nebo zda střední hodnota leží nad anebo pod cílovou hodnotou.

V důsledku těchto vlastností je na základě samotných hodnot jednotlivých indexů způsobilosti obtížné usuzovat, v jaké míře dosaženou úroveň ovlivňuje variabilita a v jaké poloze sledovaného znaku. Příklady této nejednoznačnosti budeme ilustrovat na příkladě.

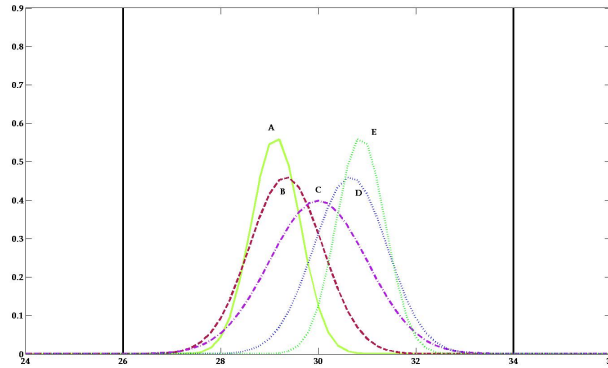
	A	B	C	D	E
μ	28	29	30	31	32
σ	0,5	0,75	1,0	0,75	0,5
C_{pk}	1,33	1,33	1,33	1,33	1,33
C_p	2,67	1,78	1,33	1,78	2,67
C_{pL}	1,33	1,33	1,33	2,22	4,00
C_{pU}	4,00	2,22	1,33	1,33	1,33
$C_{pm} (T = 30)$	0,65	1,07	1,33	1,07	0,65
C_{pmk}	0,32	0,80	1,33	0,80	0,32
ppm	33,05	33,05	66,10	33,05	33,05

Skutečnost, že hodnota indexů závisí jak na střední hodnotě tak variabilitě vedle k tomu, že procesy s rozdílným rozdělením mohou mít stejnou hodnotu některých indexů.



	A	B	C	D	E
μ	29,13	29,34	30,00	30,66	30,87
σ	0,5	0,75	1,0	0,75	0,5
C_{pk}	2,09	1,48	1,33	1,48	2,09
C_p	2,67	1,78	1,33	1,78	2,67
C_{pL}	2,09	1,48	1,33	2,07	3,24
C_{pU}	3,24	2,07	1,33	1,48	2,09
$C_{pm} (T = 30)$	1,33	1,33	1,33	1,33	1,33
C_{pmk}	1,04	1,11	1,33	1,11	1,04
ppm	$1,81 \cdot 10^{-4}$	4,50	66,10	4,50	$1,81 \cdot 10^{-4}$

Tento příklad potvrzuje nezbytnost používání vhodných kombinací indexů a význam grafického zobrazení. Zejména při sledování trendů vývoje způsobilosti procesu je potřeba mít na paměti, že může docházet ke změnám chování procesu, které se nemusí projevit ve změně sledovaného indexu způsobilosti.



11.2.4. Robustnost výrobního procesu

Vedle schopnosti procesu udržet cílovou hodnotu, kdy je $\mu = T$ je žádoucí dostatečná rezerva na konci Gaussovy křivky a krajními body tolerance.

To je velmi významné, neboť jakékoli zhoršení výrobního procesu při dostatečné rezervě nemusí vést ke zmetkovosti.

Uvedená vlastnost, kdy odchylka průměru od cílové hodnoty resp. zvětšení rozptylu nezvýší podíl zmetků, se nazývá robustnost procesu.

Míra robustnosti procesu se popisuje vzorcem

$$R = 3(C_{pk} - 1), \quad (63)$$

kde R je vzdálenost, o kterou by se musel posunout průměr μ od cílové hodnoty T , aby byl proces nezpůsobilý.

Tato vzdálenost se vyjadřuje v násobcích σ .

Např. pro $C_p = 1,33$ je

$$\frac{USL - LSL}{6\sigma} = 1,33 \rightarrow USL - LSL = 1,33 \cdot 6\sigma \sim 8\sigma.$$

Délka tolerančního intervalu je tedy 8σ .

Robustnost při $C_p = 1,33$ je $R = 3(1,33 - 1) = 1$.

Pro $C_p = 1,67$ je

$$\frac{USL - LSL}{6\sigma} = 1,67 \rightarrow USL - LSL = 1,67 \cdot 6\sigma \sim 9,96\sigma.$$

Délka tolerančního intervalu je tedy 10σ .

Robustnost při $C_p = 1,67$ je $R = 3(1,67 - 1) = 2$.

Pro $C_p = 1,0$ je $R = 3(1,0 - 1) = 0$.

To znamená, že sebemenší zhoršení vede ihned k nezpůsobilosti procesu.

$$R = 3(C_{pk} - 1), \quad (64)$$

Vedle robustnosti jsou i další dobré důvody, proč se požaduje index způsobilosti větší než 1.

Předpokládejme, že se automobil skládá z 1000 součástí a každá z nich je vyráběna s $C_{pk} = 1$. To znamená, že podíl NC výrobků, resp. pravděpodobnost, že výrobek bude mimo toleranci, je pouze 0,0027.

Avšak pravděpodobnost, že alespoň jeden z 1000 výrobků bude mimo toleranci je

$$P = 1 - \binom{1000}{0} 0,0027^0 (1 - 0,0027)^{1000} = 0,933$$

a tedy 93,3% pravděpodobnost, že alespoň jeden výrobek bude zmetek.

Při $C_{pk} = 1,33$ dostáváme značně nižší pravděpodobnost

$$P = 1 - \binom{1000}{0} 0,000066^0 (1 - 0,000066)^{1000} = 0,0638$$

11.2.5. Testování C_{pm}

Ze vztahů pro výpočet indexů způsobilosti vyplývá, že k jejich výpočtu potřebujeme teoretické charakteristiky (μ, σ) , které jsou samozřejmě nedostupné. Stanovit lze pouze jejich odhady, a tedy vypočtené indexy způsobilosti představují rovněž pouze odhady. Ty se jako proměnné řídí určitým zákonem pravděpodobnosti, mají určitou střední hodnotu, disperzi, koeficient asymetrie a špičatosti.

Z toho vyplývá, že s vypočtenou hodnotou indexu způsobilosti nelze zacházet jako s konstantou, ale jako s odhadem, pro který lze stanovit konfidenční interval.

Šířka konfidenčního intervalu závisí na počtu měření.

Mají-li náhodné proměnné X_i , $i = 1, \dots, k$ normální rozdělení pravděpodobnosti, pak náhodná veličina $\frac{C_{pm}}{\hat{C}_{pm}}$ má rozdělení pravděpodobnosti

$$\frac{1}{n} \left\{ 1 + \left(\frac{\mu - T}{\sigma} \right)^{-1} \right\} \cdot \chi_{n,\lambda}^2 \quad (65)$$

kde $\chi_{n,\lambda}^2$ je necentrální rozdělení chí-kvadrát s n stupni volnosti a parametrem $\lambda = n \left(\frac{\mu - T}{\sigma} \right)^2$.

Hodnocení způsobilosti v případě nesplnění normality dat

$$C'_p = \frac{USL - LSL}{x_{0.99865} - x_{0.00135}},$$

$$C'_{pk} = \min \left\{ \frac{x_{0.5} - LSL}{x_{0.5} - x_{0.00135}}; \frac{USL - x_{0.5}}{x_{0.99865} - x_{0.5}} \right\}.$$

11.3. Metoda FMEA

Metoda FMEA (Failure Mode and Effect Analysis – Analýza způsobů a důsledků poruch, resp. Analýza možností vzniku vad a jejich následků) představuje týmovou analýzu možností vzniku vad u posuzovaného návrhu, ohodnocení jejich rizika a návrh a realizaci opatření vedoucích ke zlepšení jakosti návrhu. Má induktivní charakter a je jednou z metod plánování a zlepšování jakosti a důležitou součástí přezkoumávání návrhu. Uvádí se, že tato metoda odhalí 70-90% možných neshod.

Metoda FMEA byla vyvinuta v šedesátých letech 20. století v USA a byla původně určena pro analýzu spolehlivosti složitých systémů v kosmickém výzkumu (projekt Apollo) a jaderné energetice. Brzy se začala využívat k prevenci výskytu neshod v dalších oblastech, přičemž k největšímu rozšíření došlo zejména v automobilovém průmyslu. Ford tuto metodu začal používat v r. 1977, koncern Volkswagen od r. 1984.

Mezinárodní normy věnované této metodě rozlišují dvě alternativy: metodu FMEA a metodu FMECA (analýza způsobů, důsledků a kritičnosti poruch). V pojetí norem metoda FMEA nezahrnuje hodnocení rizika možných způsobů poruch (vad vyvolaných určitou příčinou).

U metody FMECA je doplněno hodnocení kritičnosti důsledků poruch a pravděpodobností způsobů poruch a celková kritičnost se na základě těchto dvou kritérií vyhodnocuje v tzv. grafu kritičnosti.

V metodikách automobilového průmyslu se však standardně používá označení FMEA a přitom tyto postupy obsahují hodnocení rizika.

Pro aplikaci metody FMEA hovoří řada argumentů

- představuje systémový přístup k prevenci nejasností,
- snižuje ztráty vyvolané nízkou jakostí výrobků,
- zkracuje dobu řešení vývojových prací,
- optimalizuje návrh a vede ke snížení počtu změn ve fázi realizace (umožňuje dělat věci správně napoprvé),
- umožňuje ohodnotit riziko možných vad a na jeho základě stanovit priority opatření, vedoucích ke zlepšení jakosti návrhu,
- podporuje účelné využívání zdrojů,
- vytváří velice cennou informační databázi o výrobku, využitelnou pro podobné výrobky,
- poskytuje podklady pro zpracování nebo zlepšení plánu jakosti,
- je důležitou součástí kontrolního systému v oblasti tvorby návrhu,
- zlepšuje image a konkurenceschopnost organizace,
- pomáhá zvýšit spokojenost zákazníka,
- náklady vynaložené na její provedení jsou jen zlomkem nákladů, které by mohli vzniknout při výskytu vad.

Používání metody FMEA je doporučováno normami souboru ISO 9000:2000.

FMEA je metodou, kterou je nutno aplikovat v týmu, neboť její výhodou je právě využití znalostí a zkušeností celé řady odborníků. V týmu by měli být zastoupeni pracovníci vývoje, konstrukce, technologie, výroby, zkušeben, útvaru řízení jakosti, servisu, zástupci ekonomického útvaru a zásobování. Zákazníka obvykle zastupují pracovníci marketingu.

11.4. Fáze metody FMEA

Analýza FMEA návrhu výrobku nebo procesu probíhá v těchto fázích:

- a) analýza a hodnocení současného stavu
- b) návrh opatření
- c) hodnocení stavu po realizaci opatření

Průběh analýzy FMEA se zaznamenává do formuláře FMEA.

Vyplněný formulář by neměl být pouhým záznamem o jakosti, ale živým dokumentem dokládajícím soustavnou péči o zlepšování jakosti produkce. Součástí formuláře je podrobná hlavička, v níž jsou specifikovány základní údaje o analyzovaném návrhu, odpovědných pracovnících a času provedení.

11.4.1. FMEA – analýza a hodnocení současného stavu

Hodnocení významu vady

následek	význam vady	hodnocení
nebezpečný - bez výstrahy	vada bez výstrahy ovlivňuje bezpečnost výrobku nebo dodržování zákonných požadavků	10
nebezpečný - s výstrahou	vada ovlivňuje bezpečnost výrobku nebo dodržování zákonných požadavků s výstrahou	9
velmi vážný	nefunkční výrobek se ztrátou hlavní funkce	8
vážný	funkční výrobek se sníženou výkonností, zákazník nespokojen	7
střední	funkční výrobek s nefunkční částí zajišťující pohodlí, zákazník pociťuje nepohodlí	6
nízký	funkční výrobek ale částí zajišťující pohodlí pracují na nižší úrovni, zákazník pociťuje určitou nespokojenost	5
velmi nízký	ozdobné nebo tlumící prvky neodpovídají, vadu zaznamenaná většina zákazníků	4
malý	ozdobné nebo tlumící prvky neodpovídají, vadu zaznamenaná průměrný zákazník	3
velmi malý	ozdobné nebo tlumící prvky neodpovídají, vadu zaznamenaná náročný zákazník	2
žádný	žádný následek	1

Hodnocení očekávaného výskytu vady

pravděpodobnost	možný výskyt vad	hodnocení
velmi vysoká – vada je téměř nevyhnutelná	≥ 1 ze 2	10
	≥ 1 ze 3	9
vysoká – opakované vady	1 z 8	8
	1 z 20	7
střední – občasné vady	1 z 80	6
	1 z 400	5
	1 z 2000	4
nízká – málo vad	≥ 1 z 15000	3
	1 z 15000	2
vzdálená – vada je nepravděpodobná	≤ 1 z 1 500 000	1

Hodnocení odhalitelnosti vady

odhalitelnost	pravděpodobnost odhalení při posuzování návrhu výrobku	hodnocení
absolutně nemožná	posuzování návrhu neodhalí	1
velmi vzdálená	velmi vzdálená možnost	2
vzdálená	vzdálená možnost	3
velmi malá	velmi malá možnost	4
malá	malá možnost	5
průměrná	průměrná možnost	6
mírně nadprůměrná	mírně nadprůměrná možnost	7
vysoká	vysoká možnost	8
velmi vysoká	velmi vysoká možnost	9
téměř jistá	téměř jisté odhalení	10

$$\text{RIZIKOVÉ ČÍSLO} = \text{VÝZNAM} \times \text{VÝSKYT} \times \text{ODHALITELNOST}$$

Hodnocení očekávaného výskytu vady

význam	výskyt	odhalitelnost	charakteristika	opatření
1	1	1	ideální	NE
1	1	10	bezpečně řízený proces	NE
10	1	1	vada se nedostane k zákazníkovi	NE
10	1	10	vada se může dostat k zákazníkovi	ANO
1	10	1	častá vada, snadno odhalitelná, ale drahá	ANO
1	10	10	častá vada, která se může dostat k zákazníkovi	ANO
10	10	1	častá vada velkého významu	ANO
10	10	10	tady není nic v pořádku	ANO

12

Shluková analýza

12.1. Úvod

Z hlediska analýzy dat je vstupem pro shlukování datová matice, výstupem je identifikace shluků, které mohou být různých typů. Při shlukové analýze se zkoumá podobnost objektů, k čemuž slouží míry podobnosti. K dalším problémům, které se ve shlukové analýze řeší, patří kritéria shlukování, stanovení počtu shluků a interpretace výsledků.

Základem vícerozměrné statistické analýzy jsou m -rozměrná pozorování objektů (statistických jednotek, datových jednotek). Počet těchto jednotek je obvykle označován n (rozsah výběru). Objekty mohou být rostliny či živočichové, automobily, hospodářské ukazatele, slova ve větě.

Prvky vektoru pozorování jsou hodnoty statistických znaků neboli proměnných.

Cílem disjunktivního shlukování je vytvořit skupiny (shluky) objektů tak, aby objekt ve shluku byl co nejvíce podobný objektům ve stejném shluku a co nejméně podobný objektům v jiných shlucích. Důležité tedy je zvolit způsob, jakým bude zjišťována podobnost objektů.

Míry podobnosti v ideálním případě nabývají hodnot od nuly pro maximální rozdílnost po jedničku pro totožnost (existují ale i míry s jinými vlastnostmi).

Metody shlukové analýzy jsou obvykle založeny na mírách nepodobnosti, případně vzdálenosti.

Dvojice bodů je obvykle charakterizována jejich vzdáleností, kterou může představovat např. délka úsečky spojující tyto body. Čím je vzdálenost menší, tím jsou si body podobnější. Každou míru podobnosti lze převést na míru nepodobnosti a naopak $D = 1 - S$.

Měr podobnosti existuje velké množství — základním hlediskem pro volbu by měl být typ proměnných.

Vlastní shluková analýza vychází z matice vzdáleností. Její prvky jsou hodnoty vzdáleností (podle zvolené míry vzdáleností) vypočtené pro všechny možné dvojice objektů.

Kromě objektů lze ale také shlukovat proměnné.

12.1.1. Vstupní data

Budeme uvažovat, že vstupní datová matice \mathbf{X} je rozměru $n \times m$. Řádky představují vektory údajů o jednotlivých objektech a sloupce odpovídají jednotlivým proměnným. Prvky se označují x_{ij} .

Jiným typem vstupní datové matice, na jejímž základě se provádí shlukování, je dvourozměrná tabulka sdružených četností (kontingenční tabulka) pro dvě kategoriální proměnné. V tomto případě je rozměr vstupní matice $r \times c$, kde r je počet kategorií řádkové proměnné a c je počet kategorií sloupcové proměnné. Prvky se označují n_{ij} .

Zobrazuje-li tabulka v řádcích čtyři kategorie (1.paluba, 2.paluba, 3.paluba, posádka) a ve sloupcích údaje o nalodění na Titanic — je její rozměr $4 \times$ (počet přístavů).

Při shlukové analýze je třeba vyjít z matice vzdáleností, jejímiž prvky jsou hodnoty charakterizující vztahy mezi všemi dvojicemi objektů, resp. proměnných či kategorií. Tato matice může být dána přímo k dispozici — například jsou dány vzdálenosti mezi městy nebo rozdíly mezi cenou palubního lístku. Nebo tato matice je vypočítána na základě vstupní matice. Podle toho, zda shlukujeme objekty, proměnné či kategorie, má matice vzdáleností některý z rozměrů $n \times n$, $m \times m$, $r \times r$ nebo $c \times c$.

12.2. Rozlišení proměnných

Při analýze statistických dat se vyskytují proměnné různých typů. Jelikož v další části zavedeme pro různé typy proměnných různé míry podobnosti, uvedeme nejprve přehled různých typů proměnných.

Podle typu škály měření rozlišujeme proměnné

- nominální, u kterých můžeme určit, zda jsou různé, nemůžeme však stanovit jejich pořadí (typ absolvované střední školy, druh výrobku)
- ordinální (pořadové), u jejichž hodnot můžeme stanovit pořadí, nemůžeme však určit o kolik je jedna hodnota větší či menší než druhá (úroveň znalostí, důležitost nějakého faktoru, stupeň souhlasu s určitým výrokiem)
- kvantitativní (metrické, numerické)
 - intervalové, u kterých můžeme určit o kolik je jedna hodnota větší než druhá a které mohou nabýt hodnoty 0 (počet dětí, měsíční výdaje na určitý typ zboží)
 - poměrové, u kterých můžeme určit, o kolik i kolikrát je jedna hodnota větší než druhá a které nabývají pouze kladných hodnot (počet členů domácnosti, věk respondenta, cena výrobku)

Zvláštním typem je proměnná dichotomická (alternativní), která nabývá pouze dvou hodnot. Příkladem může být dvojice hodnot kuřák — nekuřák, ekonomicky aktivní — ekonomicky neaktivní.

- dichotomické proměnné dělíme na
 - symetrické, kdy obě kategorie mají stejnou důležitost
 - asymetrické, kdy je jedna kategorie důležitější

Dichotomická proměnná bývá někdy uváděna vedle výše uvedených typů, někteří autoři řadí dichotomické symetrické proměnné k nominální škále a asymetrické ke škále ordinální.

U těchto proměnných se při výpočtech uvažuje, že jde o proměnné binární, které nabývají hodnot 0 a 1.

Kvantitativní proměnné můžeme dělit na

- diskrétní, které nabývají pouze celočíselných hodnot
- spojitě (metrické), které mohou nabývat libovolných hodnot z určitého intervalu (věk respondenta, cena výrobku)

Nominální, ordinální a kvantitativní diskrétní proměnné s malým počtem variant hodnot můžeme souhrnně označit jako kategoriální (varianty hodnot těchto proměnných nazýváme kategoriemi).

12.3. Transformace proměnných

Při analýze dat je vhodné použít transformaci do standardizované škály.

Známe-li odhad střední hodnoty \bar{x} a sm2rodatné odchylky s_x , můžeme tyto výsledky transformovat do škály se střední hodnotou $EY = a$ a s rozptylem $varY = b^2$ takto

$$Y = a + b \frac{X - \bar{x}}{s_x}.$$

Uvedená transformace umožňuje lépe srovnávat výsledky/hodnoty různých proměnných, které nabývají různých hodnot.

Např. tato transformace převede $X \in \{0, 1, 2, \dots, n\}$ na standardizovanou veličinu $Y \in \langle y_0 = a - b\bar{x}/s_x; y_1 = y_0 + bn/s_x \rangle$.

Standardizovaná škála $Y = \frac{X-\mu}{\sigma}$, resp. $Y = \frac{X-\bar{x}}{s_x}$ odvozená z proměnné X měřené na spojitě (metrické) je pro interpretace nevýhodná z několika důvodů

- a) přibližně polovina hodnot je záporná
- b) většinou to nejsou celá čísla
- c) s_x je často vzhledem k \bar{x} velké

Tyto důvody vedou k volbě různých transformací (tak aby Y mělo určitou střední hodnotu a rozptyl) s různými druhy stupnic

- (1) normalizovaná s $EY = 0$, $DY = 1$
- (2) T (W.C. Mc Call) s $EY = 50$, $DY = 100$
- (3) C (J.P. Guilford) s $EY = 5$, $DY = 4$
- (4) stanine (J.P. Guilford) s $EY = 50$, $DY = 3.84$
- (5) sten (A.A. Canfield) s $EY = 5.5$, $DY = 4$
- (6) profile (Graduate Record Examinations) s $EY = 500$, $DY = 100^2$

12.4. Míry podobnosti a míry vzdálenosti

Pro zjišťování podobnosti objektů jsou používány tzv. *míry podobnosti* a také *míry nepodobnosti*.

Pro objekty \mathbf{x}_i a \mathbf{x}_j budeme podobnost zapisovat $S(\mathbf{x}_i, \mathbf{x}_j)$ a zkráceně S_{ij} . Platí, že $S_{ij} = S_{ji}$. Většinou tyto míry nabývají hodnot z intervalu $\langle 0, 1 \rangle$. Pak platí, že $S_{ii} = 1$.

Míru nepodobnosti označujeme $D(\mathbf{x}_i, \mathbf{x}_j)$ a zkráceně D_{ij} . Platí, že $D_{ij} \leq 0$, $D_{ii} = 0$, $D_{ij} = D_{ji}$.

12.4.1. Míry vzdálenosti

V případě kvantitativních dat se pro vyjádření vztahu dvou objektů používají míry vzdálenosti, které jsou založeny na prezentaci objektů v prostoru, jehož souřadnice představují jednotlivé proměnné. Je-li splněna trojúhelníková nerovnost, tj. $D_{ij} + D_{jk} \geq D_{ik}$, $i, j, k = 1, \dots, n$, pak hovoříme o metrice.

K nejznámějším typům vzdáleností patří euklidovská D_E , vážená euklidovská D_{EW} s váhami w_i pro každou i -tou proměnnou, čtvercová D_{ES} , manhattanská (pro binární data Hammingova vzdálenost) D_B , Čebyševova D_C , Minkovského D_M a Lanceyova-Williamsova

(Canberra) D_{LW} .

$$D_E(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^n (x_{i,k} - x_{j,k})^2} = \|\mathbf{x}_i - \mathbf{x}_j\|, \quad (66)$$

$$D_{EW}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^n w_k^2 (x_{i,k} - x_{j,k})^2}, \quad (67)$$

$$D_{ES}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^n (x_{i,k} - x_{j,k})^2}, \quad (68)$$

$$D_B(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^n |x_{i,k} - x_{j,k}|} = |\mathbf{x}_i - \mathbf{x}_j|, \quad (69)$$

$$D_C(\mathbf{x}_i, \mathbf{x}_j) = \max_k (|x_{i,k} - x_{j,k}|), \quad (70)$$

$$D_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt[q]{\sum_{k=1}^n |x_{i,k} - x_{j,k}|^q}, \quad (71)$$

$$D_{LW}(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} \sum_{k=1}^n \frac{|x_{i,k} - x_{j,k}|}{|x_{i,k}| + |x_{j,k}|} & \text{pro } |x_{i,k}| + |x_{j,k}| \neq 0, \\ 0 & \text{pro } |x_{i,k}| + |x_{j,k}| = 0, \end{cases} \quad (72)$$

Jako váhu w_i lze použít převrácenou hodnotu směrodatné odchylky i -té proměnné $\frac{1}{s_i}$ nebo převrácenou hodnotu variačního rozpětí i -té proměnné $\frac{1}{R_i}$.

12.4.2. Míry podobnosti

K nejznámějším typům patří kosinová míra S_K , Jacardův koeficient S_J , Diceův koeficient S_D a Czekanowského koeficient S_C .

$$S_K(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{k=1}^n x_{i,k} x_{j,k}}{\sqrt{\sum_{k=1}^n x_{i,k}^2 \sum_{k=1}^n x_{j,k}^2}}, \quad (73)$$

$$S_J(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{k=1}^n x_{i,k} x_{j,k}}{\sum_{k=1}^n x_{i,k}^2 \sum_{k=1}^n x_{j,k}^2 - \sum_{k=1}^n x_{i,k} x_{j,k}}, \quad (74)$$

$$S_D(\mathbf{x}_i, \mathbf{x}_j) = \frac{2 \sum_{k=1}^n x_{i,k} x_{j,k}}{\sum_{k=1}^n x_{i,k}^2 + \sum_{k=1}^n x_{j,k}^2}, \quad (75)$$

$$S_C(\mathbf{x}_i, \mathbf{x}_j) = \frac{2 \sum_{k=1}^n \min\{x_{i,k}, x_{j,k}\}}{\sum_{k=1}^n (x_{i,k} + x_{j,k})}, \quad (76)$$

$$(77)$$

12.4.3. Příklad

Mějme tři objekty charakterizované dvěma proměnnými. Příslušné vektory jsou $\mathbf{x}_1 = (1, 1)'$, $\mathbf{x}_2 = (3, 3)'$ a $\mathbf{x}_3 = (1, 3)'$. Určete jejich vzdálenosti a koeficienty.

míra	$\mathbf{x}_1 - \mathbf{x}_2$	$\mathbf{x}_2 - \mathbf{x}_3$
čtverc. eukl. míra D_{ES}	8	4
manhatt. míra D_B	4	2
eukl. míra D_E	2,8	2
Čebyšev. vzd. D_C	2	2
Lanceyova-Williamsova D_{LW}	1	0,5
Jaccardův koef. $(1 - S_J)$	0,57	0,25
Czekanowského koef. $(1 - S_C)$	0,5	0,2
Diceův koef. $(1 - S_D)$	0,4	0,14
kosinová míra $(1 - S_K)$	0	0,11

12.5. Algoritmus hierarchického shlukování

Nechť je dána množina objektů x_1, \dots, x_n a předpokládejme, že každý objekt z X je popsán m kvantitativními znaky tak, že si je můžeme představit jako k -rozměrné vektory: $x_i = (x_{i1}, \dots, x_{im})$, kde $x_{ij} \in M_j$ a M_j jsou pro uvažovaná $j = 1, 2, \dots, k$ číselné množiny.

Vyjdeme z toho, že každý objekt množiny X tvoří elementární shluk $S_i = \{x_i\}$, $i = 1, \dots, n$ reprezentovaný k -ticí odpovídající příslušnému x_i . A tak máme na začátku n -prvkovou množinu S shluků

$$S = \{S_1, S_2, \dots, S_n\}.$$

Označme nyní divergenci d mezi i -tým a j -tým shlukem jako d_{ij}

$$d_{ij} = d(S_i, S_j), S_i, S_j \in S$$

a hledejme mezi všemi hodnotami d_{ij} , $i \neq j$ hodnotu nejmenší.

Pak utvořme jediný nový shluk z těch $S_i^*, S_j^* \in S$ pro které je právě uvedené d_{ij} nejmenší. Je-li takových dvojí víc, vybereme jednu z nich. Tento nový shluk budeme reprezentovat například m -ticí aritmetických průměrů odpovídajících souřadnic reprezentantů vybraných shluků S_i^*, S_j^* .

Nový shluk zařadíme mezi shluky z S , když jsme předtím z této množiny odstranili shluky S_i^*, S_j^* . Dostaneme tak novou, $(n - 1)$ prvkovou množinu shluků.

Proces lze vyjádřit stromovým grafem. Například po třetím kroku mohou zůstat shluky $S''' = \{x_1, x_2, x_3, x_4\}$, $S_4 = \{x_4\}$, $S_6 = \{x_6\}$, $S_7 = \{x_7\}$.

Proces můžeme zastavit po určitém počtu p kroků a sledovat interpretovatelnost vytvořených shluků. Takto jsme popsali tzv. hierarchické shlukování.

Výsledek procedury závisí na volbě vztahu pro divergenci d .

Často se volí např. euklidovská metrika. Ta, ale i ostatní divergence, předpokládá jistou homogenitu m uvažovaných proměnných — proto je účelné ještě před procedurou ještě proměnné vhodně transformovat (např. standardizací).

12.5.1. Příklad

Uvažujme data z tříhodnotové stupnice $M = \{0; 0,5; 1\}$: $x_1 = (0; 0,5; 0,5)$, $x_2 = (0,5; 0,5; 0,5)$, $x_3 = (0; 0,5; 0,5)$, $x_4 = (0; 0,5; 0,5)$, $x_5 = (0; 0,5; 0,5)$, $x_6 = (0; 0,5; 0,5)$, $x_7 = (0; 0,5; 0,5)$ a divergenci ve tvaru

$$d(x_i, x_j) = \sum_{k=1}^m (x_{ik} - x_{jk})^2.$$

Matice divergencí (je symetrická) má pak tvar

$$D = \begin{pmatrix} 0; 0,25; 0,25; 0,25; 0,25; 1,25; 0,75; \\ 0; 0,50; 0,50; 0,50; 0,50; 0,50; 0,50; \\ 0; 1,00; 0,50; 1,50; 0,50; \\ 0; 0,50; 1,50; 1,50; \\ 0; 1,00; 1,50 \\ 0; 1,50 \\ 0 \end{pmatrix}$$

Budeme volit ze čtyř možností v prvním řádku matice D a klademe $S'_1 = \{x_1, x_2\}$ a charakteristikou tohoto shluku bude aritmetický průměr souřadnic vektorů x_1 a x_2 , který je $(0,25; 0,50; 0,50)$.

Po tomto kroku se změnění dimenze a první dva řádky matice divergencí, dostaneme tak matici divergencí mezi shluky prvního řádu

$$D^1 = \begin{pmatrix} 0; 0,31; 0,31; 0,31; 0,81; 0,56 \\ 0; 1,00; 0,50; 1,50; 0,50; \\ 0; 0,50; 1,50; 0,50; \\ 0; 1,00; 1,50; \\ 0; 1,50 \\ 0 \end{pmatrix}$$

V prvním řádku matice D^1 zvolíme hodnotu 0,31 jako nejmenší divergenci mezi S' a $\{x_3\}$.

Volíme tedy nový shluk $S''_1 = \{x_1, x_2, x_3\}$ a charakteristikou tohoto shluku bude aritmetický průměr souřadnic vektorů x_1 , x_2 a x_3 , který je $(0,17; 0,67; 0,50)$.

Opět se přepočte matice divergencí, dostaneme tak matici divergencí mezi shluky druhého řádu

$$D^2 = \begin{pmatrix} 0; 0,48; 0,31; 0,97; 0,47 \\ 0; 0,50; 1,50; 1,50; \\ 0; 1,00; 1,50 \\ 0; 1,50 \\ 0 \end{pmatrix}$$

Nejmenší prvek matice divergencí má hodnotu 0,31 a leží ve třetím sloupci — a to znamená, že do shluku S''_1 zařadíme shluk $\{x_5\}$ a dostaneme nový shluk $S'''_1 = \{x_1, x_2, x_3, x_5\}$ s charakteristikou $(0,13; 0,63; 0,38)$.

V iteračním procesu dále pokračujeme a dostáváme

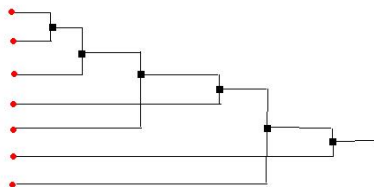
$$D^3 = \begin{pmatrix} 0; 0,43; 0,92; 0,66 \\ 0; 1,50; 1,50; \\ 0; 1,50 \\ 0 \end{pmatrix}$$

Nejmenší prvek matice divergencí má hodnotu 0,43 a leží ve druhém sloupci — a to znamená, že do shluku S'''_1 zařadíme shluk $\{x_4\}$ a dostaneme nový shluk $S''''_1 = \{x_1, x_2, x_3, x_4, x_5\}$ s charakteristikou $(0,10; 0,50; 0,40)$.

Matice divergencí má tvar

$$D^4 = \begin{pmatrix} 0; 0,97; 0,77; \\ 0; 1,50 \\ 0 \end{pmatrix}$$

Do shluku S_1''' zařadíme shluk $\{x_7\}$ a dostaneme nový shluk $S_1^{(5)} = \{x_1, x_2, x_3, x_4, x_5, x_7\}$ s charakteristikou $(0,17; 0,58; 0,50)$.



12.6. Algoritmus nehierarchického shlukování (ploché, flat)

Nehierarchickou shlukovací proceduru dostaneme, pokud zvolíme jistou množinu V vzorovacích objektů $V = \{v_1, v_2, \dots, v_n\}$, kde $v_i = (v_{i1}, \dots, v_{im})$, $v_{ij} \in M_j$, $i = 1, \dots, c$, $j = 1, \dots, m$ — a každý objekt $x_i \in X$ zařadíme do shluku, který je reprezentován některým vzorovacím objektem $v_j \in V$ s co nejmenší divergencí $d_{ij} = d(x_i, v_j)$.

Volíme-li $\varepsilon > 0$, pak za předpokladu, že $d(v_i, v_j) > \varepsilon$ pro $i \neq j$, $i, j = 1, 2, \dots, c$ do j -tého shluku zařadíme všechny objekty $x_i \in X$ pro které je $d(x_i, v_j)$ nejmenší a ještě $d(x_i, v_j) < \varepsilon$ — je-li takových možných vzorovacích objektů více, volíme jeden z možných shluků

$$x_i \in S_j \leftrightarrow d(x_i, v_j) = \min_{v_k \in V} (d(x_i, v_k), \varepsilon).$$

Pak je ale možné, že některé prvky $x \in X$ nebudou nikam zařazeny. Pak je třeba buď upravit vzorovací množinu V nebo zvětšit hodnotu kritéria ε tak, aby každý prvek byl právě v jednom shluku.

Nedostatkem nehierarchického shlukování je někdy dosti subjektivní apriorní volba vzorovacích objektů — proto je důležitá volba hodnoty ε .

12.6.1. Příklad

Uvažujme data z předchozího příkladu k nehierarchickému shlukování a vytvořme množinu vzorovacích objektů $V = \{v_1, v_2, v_3\}$, kde $v_1 = (0; 0; 0)$, $v_2 = (0,5; 0,5; 0,5)$, $v_3 = (1; 1; 1)$. a zvolme shodnou divergenci a zvolme hodnotu $\varepsilon = 0,5$.

Uřídíme $d(v_1, v_2) = d(v_2, v_3) = 0,75$, $d(v_1, v_3) = 3$.

Dále určíme divergence mezi našimi 7 objekty a 3 vzorovacími objekty, viz tabulka.

x	$d(x, v_1)$	$d(x, v_2)$	$d(x, v_3)$	shluk
x_1	0,50	0,25	1,50	2
x_2	0,75	0,00	0,75	2
x_3	1,25	0,50	1,25	—
x_4	0,25	0,50	2,25	1
x_5	0,25	0,50	2,25	1
x_6	1,25	0,50	1,25	—
x_7	2,25	0,50	0,25	3

Prvky x_3 a x_6 nejsou zařazeny do žádného ze tří možných shluků, protože jejich divergence se vzorovými shluky jsou jen rovny zvolenému ε — nejsou menší.

Můžeme tedy buď zvětšit ε např. o 0,1 a pak budou prvky x_3 a x_6 zařazeny do 3 shluku; nebo vytvoříme nový vzorový shluk, např. $v_4 = (1; 0,5; 1)$.

Z tabulky také můžeme odhadnout míru „náležitosti“ objektu $x \in X$ k určitému shluku. Objekt x bude patřit k určitému shluku tím lépe, čím blíže bude ve smyslu zavedení divergence k jeho vzorovacímu objektu.

Za míru náležitosti objektu $x \in X$ k j -tému shluku můžeme tedy považovat

$$\mu(x_i, j) = 1 - \frac{d(x_i, v_j)}{\max_p d(x_i, v_p)} \quad i = 1, \dots, n; j = 1, \dots, c.$$

Hodnoty míry náležitosti pro naši úlohu uvedeme v tabulce

x	$\mu(x, v_1)$	$\mu(x, v_2)$	$\mu(x, v_3)$	shluk
x_1	0,67	0,83	0,00	2
x_2	0,00	1,00	0,00	2
x_3	0,00	0,60	0,00	–
x_4	0,89	0,78	0,00	1
x_5	0,89	0,78	0,00	1
x_6	0,00	0,60	0,00	–
x_7	0,00	0,78	0,89	3

Je vidět, že např. objekty x_4 a x_5 zařazené do 1. shluku mají poměrně vysokou míru náležitosti k 2. shluku. I zde můžeme stanovit hranici pro zařazení nějakého objektu do určitého shluku (míra náležitosti 0.60 a nižší).

Určením $\mu(x_i, v_j)$ jsme zařazení objektu do určitého shluku „ocenili“. To má velký význam pro úvahy o vhodnosti vzorovacích objektů a pro interpretační úvahy.

Míra náležitosti nás přivádí k možnosti realizovat shlukování překrývající se — když dosud jsme uvažovali shlukování disjunktní.

Míra

$$\mu(x_i, j) = 1 - \frac{d(x_i, v_j)}{\max_p d(x_i, v_p)} \quad i = 1, \dots, n; j = 1, \dots, c.$$

je vztažena k určitému objektu nezávisle na zbývajících — a je tedy mezi objekty neporovnatelná. Tuto nepříjemnou vlastnost změníme učením míry následujícím vztahem

$$\mu(x_i, j) = 1 - \frac{d(x_i, v_j)}{\max_{r,p} d(x_r, v_i)} \quad i = 1, \dots, n; j = 1, \dots, c.$$

Hodnoty této nové míry uvedeme v následující tabulce.

Má-li být každý objekt x_i zařazen právě do jednoho shluku, je někdy ještě užitečné požadovat normování všech hodnot po řádcích tak, aby pro každé uvažované x platilo $\sum_j \mu(x_i, j) = 1$.

V tabulce takto normované hodnoty míry uvedeme v závorce.

x	$\mu(x, v_1)$	$\mu(x, v_2)$	$\mu(x, v_3)$	shluk
x_1	0,78 (0,39)	0,89 (0,45)	0,33 (0,16)	2
x_2	0,67 (0,29)	1,00 (0,43)	0,67 (0,29)	2
x_3	0,44 (0,26)	0,78 (0,47)	0,44 (0,26)	1
x_4	0,89 (0,53)	0,78 (0,47)	0,00 (0,00)	1
x_5	0,89 (0,53)	0,78 (0,47)	0,00 (0,00)	1
x_6	0,44 (0,26)	0,78 (0,47)	0,44 (0,26)	2
x_7	0,00 (0,00)	0,78 (0,47)	0,89 (0,53)	3

Dalším předmětem zkoumání je hledání optimálních vzorovacích shluků zcela z experimentálních vah. Tato úloha vede na minimalizaci cenového funkcionálu a tomuto postupu v nehierarchické analýze se pro jeho složitost nebudeme věnovat.

12.7. Velké datové soubory

Základním problémem velkých souborů dat je, že analýza vycházející z matice vzdáleností vypočtených pro všechny dvojice objektů je velmi náročná a to výpočetně a z hlediska paměťových nároků na uložení matice.

Uvádí se, že shlukovací algoritmy založené na mírách vzdáleností fungují efektivně do 16 proměnných.

Jedním z přístupů k řešení tohoto problému je snížení rozměru úlohy (redukce dimenze) na základě analýzy hlavních komponent, kdy jsou lineární kombinací původních proměnných vytvořeny nové (pomocné) proměnné.

Metodě hlavních komponent se budeme věnovat v další kapitole.

Literatura

Půlpán, Z.: *K problematice zpracování empirických šetření v humanitních vědách*. Acedemia, studie AV ČR, číslo 1/2004.

Řezanková, H., Húsek, D., Snášel, V.: *Shluková analýza dat*. PBtisk, Příbram, 2007.

Hebák, P., Hustopecký, J., Pecáková, I., Průša, M., Řezanková, H., Svobodová, A., Vlach, P.: *Vícerozměrné statistické metody (3)*. Informatorium, Praha, 2005.

13

Metoda hlavních komponent

13.1. Úvod

Cílem analýzy hlavních komponent je snížení rozměru úlohy (redukce dimenze), když jsou lineární kombinací původních proměnných vytvořeny nové (pomocné) proměnné.

V našem případě jde o to, jak vytvořit z X_1, X_2, \dots, X_n nějaký menší počet nových náhodných veličin, které by byly v nějakém smyslu co nejlepší náhradou celého vektoru \mathbf{X} . Při konstrukci těchto nových veličin se omezíme na lineární kombinace složek vektoru \mathbf{X} .

Nechť $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ je náhodný vektor s varianční maticí \mathbf{V} . Předpokládejme, že tato matice má právě r , kladných vlastních čísel ($r < n$) a že tato čísla jsou vzájemně různá.

Hledejme takový vektor \mathbf{c} splňující podmínku $\mathbf{c}'\mathbf{c} = 1$, aby náhodná veličina $\mathbf{c}'\mathbf{X}$ měla co největší rozptyl. Protože $\text{Var}(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\mathbf{V}\mathbf{c}$, jde o maximalizaci výrazu $\mathbf{c}'\mathbf{V}\mathbf{c}$ za podmínky $\mathbf{c}'\mathbf{c} = 1$.

Ze statistického hlediska jde o nalezení takové lineární kombinace $\mathbf{c}'\mathbf{X}$, která vyčerpává co největší část variability vektoru \mathbf{X} .

13.1.1. Pomocná tvrzení

Nechť $\mathbf{A}_{n,n}$ je symetrická pozitivně semidefinitní matice s charakteristickými čísly $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$. Položme $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$. Pak existuje matice $\mathbf{U}_{n,n}$ tak, že platí

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}', \quad \mathbf{I} = \mathbf{U}\mathbf{U}' \quad (78)$$

Označme k -tý sloupec matice \mathbf{U} symbolem \mathbf{x}_k , což je charakteristický vektor matice \mathbf{A} , který přísluší číslu λ_k . Ze vztahu $\mathbf{I} = \mathbf{U}\mathbf{U}'$ vyplývá, že vektory $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ jsou ortonormální.

Pak lze vzorec 76 napsat ve tvaru

$$\mathbf{A} = \lambda_1 \mathbf{x}_1 \mathbf{x}_1' + \dots + \lambda_n \mathbf{x}_n \mathbf{x}_n', \quad \mathbf{I} = \mathbf{x}_1 \mathbf{x}_1' + \dots + \mathbf{x}_n \mathbf{x}_n'. \quad (79)$$

Věta 13.1 Pro každý vektor $\mathbf{x} \in \mathbb{R}_n$ splňující podmínku $\mathbf{x}'\mathbf{x} = 1$ platí nerovnost

$$\mathbf{x}'\mathbf{A}\mathbf{x} \leq \lambda_1. \quad (80)$$

Je-li $\mathbf{x} = \mathbf{x}_1$, pak platí rovnost.

13.1.2. Příklad

Najdeme hlavní komponenty na základě údajů o délce, šířce a výšce krunýře určitého druhu želv. Ze souboru 24 exemplářů byla spočtena kovarianční matice

$$\hat{\Sigma} = \begin{pmatrix} 451,39 & 271,17 & 168,70 \\ 271,17 & 171,73 & 103,29 \\ 168,70 & 103,29 & 66,65 \end{pmatrix}.$$

Nejprve najdeme charakteristická čísla vyřešením soustavy

$$\begin{pmatrix} 451,39 - \lambda & 271,17 & 168,70 \\ 271,17 & 171,73 - \lambda & 103,29 \\ 168,70 & 103,29 & 66,65 - \lambda \end{pmatrix} = 0$$

a dostaneme

$$\lambda_1 = 680,40, \lambda_2 = 6,50, \lambda_3 = 2,86.$$

Dosadíme-li postupně hodnoty λ_j do soustavy

$$(\hat{\Sigma} - \lambda_j)\mathbf{l}_j = 0,$$

obdržíme

$$\begin{aligned}\mathbf{l}_1 &= (0,8126, 0,4955, 0,3068)', \\ \mathbf{l}_2 &= (-0,5454, 0,8321, 0,1006)', \\ \mathbf{l}_3 &= (-0,2054, -0,2491, 0,9465)'. \end{aligned}$$

Hlavní komponenty pak jsou

$$\begin{aligned}Y_1 &= 0,81X_1 + 0,50X_2 + 0,31X_3, \\ Y_2 &= -0,54X_1 + 0,83X_2 + 0,10X_3, \\ Y_3 &= -0,21X_1 - 0,25X_2 + 0,95X_3. \end{aligned}$$

Symbol X_1, X_2, X_3 označuje odchylky délky, šířky a výšky krunýře od příslušných průměrů.

Platí $\mathbf{X} = \mathbf{Y}\mathbf{L}'$.

Kovarianční matice Σ_Y hlavních komponent má tvar

$$\Sigma_Y = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ & & \ddots & & \\ 0 & 0 & 0 & \dots & \lambda_n \end{pmatrix}.$$

Výpočet relativní části celkového rozptylu souvisejícího s jednou, dvěma či třemi hlavními komponentami provedeme dle vztahu

$$q(m) = \frac{\sum_{j=1}^m \sigma_{Y_j}^2}{\sum_{j=1}^n \sigma_{Y_j}^2} = \frac{\sum_{j=1}^m \lambda_j}{\sum_{j=1}^n \lambda_j}.$$

Výpočtem dostáváme

$$\begin{aligned}q(1) &= \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = 0,9864, \\ q(2) &= \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = 0,9958, \\ q(3) &= \frac{\lambda_1 + \lambda_2 + \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} = 1. \end{aligned}$$

Lze učinit závěr, že veškerá informace o specifikaci rozměrů krunýře daného druhu želv je obsažena v první hlavní komponentě. Tu lze pak použít při klasifikaci zkoumaných exemplářů.

Literatura

Anděl, J.: Matematická statistika, SNTL, Praha, 1985.

Řezanková, H., Húsek, D., Snášel, V.: *Shluková analýza dat*. PBtisk, Příbram, 2007.