VYSOKÁ ŠKOLA BÁŇSKÁ – TECHNICKÁ UNIVERZITA OSTRAVA
FAKULTA METALURGIE A MATERIÁLOVÉHO INŽENÝRSTVÍ

STUDIJNÍ OPORA

Název opory/předmětu: Chemometrie

Číslo předmětu: 615-0808

Autor/Autoři: Karel Eckschlager, Ervín Kozubek

Katedra: analytické chemie a zkoušení materiálu - 615

Tato studijní opora vznikla v rámci rozvojového projektu Tvorba elektronických studijních opor pro studijní programy FMMI v r. 2008

PŘEDMLUVA

Předkládaný text, který se snaží jednoduchým způsobem přiblížit a ozřejmit studentům základní chemometrické postupy, je v podstatě mírně upraveným opisem prvních čtyř kapitol jediné uváděné literatury [1] ECKSCHLAGER, K. *Chemometrie* [Skripta]. UK, Praha 1991, 156 s., který dostatečně pokrývá požadavky kladené na studenty v oblasti teorie. Z tohoto důvodu již jiná literatura není uváděna. Uvedený text je pouze nutnou teoretickou základnou, která musí být individuálně rozvíjena praktickým použitím běžně dostupných statistických softwarů, což bude z časových důvodů realizováno až v následných inovacích této studijní opory.

OBSAH

1 ÚVOD

Členění kapitoly:

- Úvod

Čas potřebný ke studiu: 120 minut

Cíl: Po prostudování této kapitoly

- pochopíte základní zákonitosti sledování, vytváření a zpracování experimentálních dat

Výklad

Poznávání přírody je umožněno lidskou aktivitou, která směřuje k empirickému získávání poznatků pozorováním nebo měřením a zobecňováním takto získávaných informací v dílčí hypotézy a ucelené teorie. Teorie se též vytváří logickým odvozením z určitých předpokladů, axiomů. Při volbě axiomů má ovšem určitý vliv zkušenost, empirie, event. „redukce“ na exaktnější obory (matematizace fyziky, matematizace a fyzikalizace chemie, chemizace biologie) a vznik interdisciplinárních vědních oborů (fyzikální chemie, matematická chemie, biochemie).

Metodami, které umožňují rozšiřovat vědecké poznání, se zabývá metodologie, což je obor na rozmezí filozofie a jednotlivých specielních věd. Dnes se v metodologii přírodních věd stále více uplatňuje i matematika. Je tedy metodologie přírodních věd poměrně exaktně fundována, přičemž ovšem filozofický způsob myšlení neztrácí nic na svém významu, spíš se pouze „kvantifikuje“.

Jednotlivé vědecké školy a filozofické směry, zabývající se metodologií přírodovědeckého poznání, se liší hlavně tím, jaký způsob poznávání preferují. Filozofický směr, který preferuje shromažďování, klasifikaci a zobecňování empirických poznatků a dat, označujeme jako empirismus; z matematických metod využívá empirismus hlavně statistickou indukci. Směr, který zdůrazňuje prvořadý význam ucelených teorií, se označuje jako racionalismus: ten dnes využívá mnoha matematických oborů. Metody poznávání se ovšem během času vyvíjejí, zdokonalují a jsou pro různé přírodovědní obory poněkud odlišné. V chemii převládá empirické poznávání, často zprostředkované přístroji: důležitým vodítkem pro strategii chemického výzkumu je ovšem teoretická chemie, založená na axiomatických, fyzikálních disciplínách a tzv. matematická chemie, která rozvíjí matematické prostředky pro teorii chemie a představuje alternativní přístup k její fyzikalizaci.

V některých přírodních vědách je hlavním zdrojem poznání pozorování: to se často, např. v biologii, provádí tak, aby neovlivňovalo pozorovaný objekt. V chemii spíše provádíme laboratorní experiment (pokus), tj. pozorování nebo měření za uměle nastaveného počátečního stavu a za přesně řízených podmínek. Pozorování i experiment můžeme zcela obecně pokládat za proces získávání informace, který probíhá ve stochastickém systému. Za stochastický označujeme takový systém, který při opakování pro tentýž vstup poskytuje na výstupu celé pravděpodobnostní rozdělení výsledků. Systém, ve kterém probíhá experiment, je zpravidla sestaven ze subsystémů, ve kterých probíhají jednotlivé dílčí operace celého experimentálního postupu. Tyto subsystémy, jejich pořadí a vzájemná relace jsou pro funkci celého systému nezbytné, ale každá z dílčích operací, které probíhají v jednotlivých subsystémech, přispívá svým podílem k celkové neurčitosti experimentálního výsledku.

Aby byl experiment úspěšný, musí mít jasný cíl a již při jeho navrhování musíme mít určitou předběžnou znalost nebo alespoň představu o objektu nebo ději, který sledujeme. Výsledky pokusu musíme zpracovat (zpravidla matematicky a často za použití počítače), má-li skutečně poskytnout požadovanou informaci. Provádíme-li celou sekvenci pokusů, měl by se každý další pokus plánovat s ohledem na výsledek předešlého. Mělo by se stát běžným takové plánování složitějších pokusů, na němž se podílí experimentátor, teoretik a odborník na zpracování dat. Pokusem vlastně klademe přírodě otázky: její odpovědi pak umožňují popis (v anorganické a organické chemii většinou pomocí chemických vzorců a reakčních schémat), v analytické chemii číselné vyjádření chemického složení, ve fyzikální chemii popis děje matematickým vztahem, event. výklad (interpretaci) jeho průběhu a příp. zjištění jeho příčin. I při tom má diskuze celého týmu různě zaměřených odborníků mimořádný význam.

V metodologii chemického experimentu je dnes zřejmý přesun od dřívější logické analýzy výsledků kvalitativních pokusů k matematickému zpracování naměřených dat, dnes často prováděnému za použití počítače, někdy přímo zapojeného na měřicí přístroj. Proto souvisí metodologie chemie dosti úzce s oborem, označovaným jako chemometrie. Ta je definována jako chemická disciplína, která využívá matematických metod:

1. K volbě optimální experimentální strategie;

2. K získávání maxima relevantních informací z experimentálních výsledků a k jejich prezentaci (uvádění).

Chemometrické zpracování výsledků často usnadní výklad určitého chemického jevu a jeho příčin, zejména tím, že umožňuje získat z dat informace, které nejsou bez takového zpracování patrné. Chemometrie dále umožňuje vytvářet matematické modely experimentálních metodik a kvalifikovat pojmy, artikulované dosud pouze kvalitativně, přičemž často nachází vztahy mezi pojmy, které byly dříve chápány jako nezávislé. Chemometrie také podává pravidla, jak prezentovat výsledky, aby byly zřejmé meze jejich platnosti. Zcela obecně lze k chemometrii zařadit vše, co nám v chemii slouží k tomu, aby se jednotlivá dílčí fakta a experimentální výsledky, ale i různá empirická pravidla a domněnky změnily v informace, resp. v ucelené poznatky.

Chemometrie vznikla koncem šedesátých, začátkem sedmdesátých let minulého století a od té doby se stále vyvíjí. Označení „chemometrie“ použil poprvé S. WOLD v roce 1972. Zatímco pro chemometrii sedmdesátých let, zaměřenou hlavně na fyzikálně-chemické metody a výsledky, je při zpracování dat charakteristické hlavně to, že model byl vždy znám z teorie a tak se určovaly jen hodnoty koeficientů pro daný model (tj. hodnoty látkových, rychlostních a rovnovážných konstant), objevuje se v chemometrii v druhé polovině osmdesátých let zájem - ostatně zřejmý v celé přírodě - o komplikované systémy. Složitost těchto systému je dána nejen velkým počtem subsystémů a složitými relacemi mezi nimi, ale i velkým počtem faktorů, které funkci systému ovlivňují a jsou často ve vzájemné interakci. To vyžaduje zpracování více rozměrných souborů dat, používání faktorové analýzy, multikriteriální rozhodování apod. Poměrně autonomní částí oboru je tzv. analytická chemometrie, která se zabývá otázkami optimální analytické strategie a metodami zpracování analytických dat, dnes pomocí počítače, zapojeného přímo na analytický instrument.
Prof. G. KATEMAN charakterizoval v roce 1988 analytickou chemometrii jako „nehmotnou část analytické chemie“.

Experimentální výsledky, zejména jsou-li dobře matematicky zpracovány, by měly vždy vést k zobecnění, k verifikaci nebo posílení dosavadní hypotézy (domněnky) nebo uznávané teorie, event. k jejímu rozšíření. Výsledek experimentu však může naopak vést k popření, vyvrácení dosavadní hypotézy a vést k návrhu alternativní domněnky.

Proces zobecňování empirických výsledků, tj. indukce vede v chemii k vytváření poznatků, pravidel, hypotéz až event. ke vzniku obecně uznávaných teorií, příp. přírodních zákonů. Pravidla a hypotézy bývají vyjádřeny slovně, vzorcem nebo schématem, tj. kvalitativně; teorie a přírodní zákony jsou zpravidla popsány kvantitativně, např. matematickým vztahem. Indukcí však nelze žádnou domněnku zcela bezpečně dokázat, protože nikdy předem nevíme, zda se nenajde jev, fenomén, který ji vyvrátí. Proto mluvíme spíše o posílení určité hypotézy, jestliže získáme experimentální výsledky, které dané domněnce vyhovují nebo jí alespoň neodporují. Takové „posílení“ je významné zejména tehdy, jestliže nové experimentální výsledky vyhovují dané hypotéze a ostatním alternativním domněnkám odporují. Na druhé straně však možnost empirického popředí, negace, pokládá K. R. POPPER za velmi důležitou vlastnost přírodovědecké hypotézy. V odmítání jakýchkoliv neověřených hypotéz jsou důslední zejména pozitivisté.

Ve fyzikální chemii se někdy vytváří celé teoretické systémy logickým odvozením z několika základních axiomů. Jako axiomy mohou být použity určité konvence nebo empirické hypotézy a celý systém musí být založen na minimálním počtu vzájemně nezávislých předpokladů, mezi nimiž není rozporu (kontradikce). Matematik K. GŐDEL však dokázal, že axiomatický systém nikdy nelze založit na něm samém: abychom dokázali jeho nerozpornost, musíme použít argumentů, procházejících z jeho vnějšku, např. ze široké oblasti lidské zkušenosti. Příkladem axiomatického systému je např. model atomu, založený na BOHROVÝCH postulátech, termodynamika, založená na třech větách termodynamických nebo kvantová mechanika, která je celá vybudována na pěti axiomech apod.

Induktivně vytvořený nebo axiomaticky vybudovaný teoretický aparát umožňuje nejen vysvětlovat jevy, ale také je předvídat. Umí často předpovědět, jaký vliv mají jednotlivé faktory na průběh těchto jevů, tj. umožňuje deduktivní usuzování z obecné teorie na specielní

případy. Indukce a dedukce se vzájemně doplňují při získávání nových chemických poznatků tak, že se indukcí z výsledků experimentů vytvoří hypotéza, která usnadňuje plánování dalších účelných pokusů dedukcí, předvídáním optimálních podmínek experimentu, umožňuje předem odhadnout potřebnou přesnost měření, účelný počet opakování jednotlivých měření apod. Teoretickým základem pro dedukci v chemii je dnes jednak teoretická (fyzikální) chemie, jednak matematická chemie, založená na formálním (matematickém) přístupu k racionalizaci a unifikaci základních chemických pojmů. Oba tyto přístupy byly ovšem v podstatě vytvořeny indukcí z empirických poznatků anorganické a organické, případně obecné chemie.

Součástí vytváření kvantitativně formulovaných teorií z kvalitativně vyjádřených (artikulovaných) hypotéz je i kvantifikace pojmů (konceptů). Tento proces většinou probíhá tak, že se pojmy vytvářejí na základě zkušenosti zprvu jako názvy (termíny) v rámci určité hypotéky a teprve při vytváření teorie jsou definovány matematicky. Kvalifikace pojmů, tj. matematické vyjádření definic, umožní často najít souvislost mezi pojmy, které byly do té doby chápány jako spolu vzájemně nesouvisející. To má význam hlavně pro rozvoj mezioborových disciplín. Někdy se nepodaří zobecnit experimentální výsledky, ale je možné vytvořit matematický model dané, empiricky zjištěné závislosti. Takový model, i když jev nevysvětluje, alespoň jej kvantitativně popisuje: to někdy usnadní pozdější výklad a často umožňuje předvídání určitého jevu nebo průběhu nějaké závislosti, což je užitečné pro plánování experimentů, které mají přinést další poznatky o jevu, popsaném matematickým modelem. Každý model má ovšem omezenou oblast platnosti a jakákoliv extrapolace mimo tuto oblast je velmi riskantní.

Konzistentní soubor teorií, vytvořených v různých vědních oborech, zejména mají-li obecnější světonázorový význam, označujeme jako paradigma. Paradigma zůstává v platnosti zpravidla po několik generací: nové paradigma zastávají zprvu jen některé vědecké školy a stává se, že určitá škola zanikne, nedokáže-li přijmout nové názory. Takovou změnu paradigmatu způsobilo v předminulém století poznání, že ohromná rozmanitost různých forem neživé i živé hmoty na zemi i ve vesmíru je výsledkem kombinací omezeného počtu prvků, nebo v nedávné době změnily celou fyziku teorie relativity, vlnová a kvantová teorie.

Ukázaly, kde končí meze použitelnosti klasických pojmů fyziky, objasnily jednotu času a prostoru, hmoty a energie, setrvačnosti a gravitace a umožnily nám pochopit, že elektřina, magnetizmus a světlo jsou jen různé projevy téhož fyzikálního působení. Změna paradigmatu představuje vždy velký přelom ve vědě, v lidském myšlení a jednání. Často zahajuje novou éru vědeckého a technického rozvoje.

Věda je přes veškerou snahu po objektivnosti lidským dílem a nese stopy osob, které ji rozvíjejí, a to nejen v tom jak řeší problémy, ale i v tom, jakými metodami pracují. Aby empirické získávání chemické informace plnilo skutečně svou funkci v procesu rozšiřování poznatků, aby vzájemné doplňování induktivního a deduktivního přístupu umožňovalo konvergenci teorie ke skutečnosti, je účelné, aby každý, kdo se na tomto procesu chce podílet:

1. Seznámil se alespoň s nejdůležitějšími logickými základy chemometrie a naučil se používat chemometrické metody i při běžné laboratorní práci, přičemž by měl ovládat používání výpočetní techniky;
2. Naučil se využívat co nejvíce týmové práce, na níž se podílejí odborníci různého zaměření;
3. Zvykl si na skutečnost, že se věda i její metody neustále vyvíjejí a že je nutné tento vývoj neustále sledovat a přizpůsobovat mu svou vlastní práci.

2 TEORIE PRAVDĚPODOBNOSTI

Členění kapitoly:

- Základy teorie množin
- Základy teorie pravděpodobnosti
- Náhodná veličina
- Centrální limitní věta

Čas potřebný ke studiu: 600 minut

Cíl: Po prostudování této kapitoly

- pochopíte základní zákonitosti teorie množin
- pochopíte základní zákonitosti teorie pravděpodobnosti
- seznámíte se s pojmy pravděpodobnost, náhodná veličina, distribuční funkce, očekávána hodnota, rozptyl, směrodatná odchylka, koeficient šikmosti, koeficient špičatosti apod.
- seznámíte se s některými základními typy rozdělení pravděpodobností

Výklad

Systém, ve kterém probíhá chemický experiment, je vždy systémem difúzním, tj. jeho výstup (získaná informace) má jistou neurčitost. Tato neurčitost může mít různé složky: u výsledků měření jsou to např. chyby, v případě kvalitativních experimentů nedokonalá rozlišitelnost jednotlivých výsledků apod. Vznik neurčitosti si vysvětlujeme tak, že se během provádění pokusu v různé míře uplatňují různé nahodilé okolnosti: proto nejčastěji popisujeme neurčitost experimentálně zjištěné informace pomocí pojmů teorie pravděpodobnosti. Jen málokdy můžeme zjistit příčinu této neurčitosti, spíše jde většinou o souhru mnoha příčin, z nichž některé ani neznáme. Proto také tyto příčiny ani nehledáme, ale pokládáme experimentální systém za stochastický, tj. takový, že při opakování pro tentýž vstup dostáváme na výstupu soubor hodnot řídící se určitým pravděpodobnostním rozdělením.

Pro různé rozhodování, které provádíme na základě experimentálně zjištěných informací jsou různé složky neurčitosti různě nepříznivé. Je zcela nerealistické chtít úplně odstranit veškerou neurčitost; proto hledáme takový experimentální postup, který poskytuje výsledky, nezatížené tou složkou neurčitosti, která by nejvíce ohrožovala správnost našeho rozhodování. Vždycky chceme získat výsledky relevantní pro daný problém. Abychom získali relevantní výsledky, musíme umět kvantitativně vyjádřit celkovou neurčitost, klasifikovat její jednotlivé složky a najít faktory, které vznik neurčitosti nejvíce ovlivňují. Přitom se používá teorie pravděpodobnosti, matematická statistika, teorie chyb, informace systému a event. i další matematické obory.

Pravděpodobnost je definována axiomaticky za použití teorie množin a relevanci výsledků pro určitý problém lze kvalifikovat pomocí teorie tzv. neostrých („fuzzy“) množin. Proto zde uvedeme hlavní pojmy z teorie „klasických“ i neostrých množin.

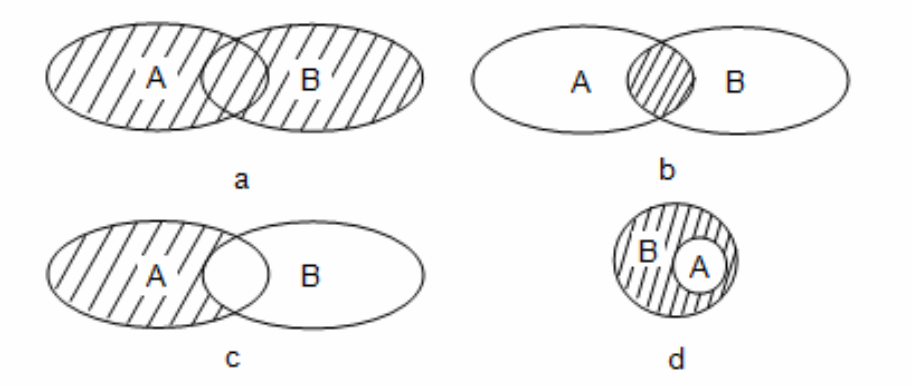
2.1 Základy teorie množin

Moderní matematika vychází z obecné teorie množin. Množina je soubor vzájemně rozlišitelných prvků s určitou vlastností, která je zařazuje do této množiny. Množiny jsou buď ostré: prvek x do množiny A buď určitě patří, $x \in A$, nebo do ní nepatří, $x \notin A$; nebo jsou neostré (fuzzy), kdy prvek do množiny patří více nebo méně. Příslušnost prvku x do neostré množiny A je charakterizována hodnotou funkce příslušnosti $m_A(x) \in \langle 0,1 \rangle$. Funkce příslušnosti nabývá různých hodnot podle okolností, takže některý prvek x patří do fuzzy-množiny A více za jistých okolností než za jiných. Přitom vždy platí, že $m_{\bar{A}}(x) = 1 - m_A(x)$, kde $m_{\bar{A}}(x)$ je hodnota funkce nepříslušnosti prvku x do fuzzy-množiny A. Prázdnou množinu, která neobsahuje žádný prvek, označujeme $\varnothing$.

Řešený příklad:

Máme hrst červených a žlutých kuliček; není těžké je roztřídit do dvou ostrých množin. Máme-li však hrst kuliček různých barevných odstínů od žluté přes oranžovou do červené, budou množiny červených (Č) a žlutých (Ž) kuliček neostré: čistě žluté (ž) nebo červené (č) kuličky zařadíme bez váhání: $m_{Č}(č) = 1$, $m_{Ž}(č) = 0$, ale pro oranžovou kuličku (o) bude $m_{Č}(o) \in (0,1)$; $m_{Ž}(o) = 1 - m_{Č}(o)$ podle toho, zda má odstín více do červena nebo do žluta.

Dalším důležitým pojmem je podmnožina. Množina A je podmožinou množiny B, $A \subset B$, je-li každý prvek $x \in A$ také prvkem množiny B, tj. $x \in B$. Existuje-li prvek $y \in B$, $y \notin A$, pak $A \subset B$, $A \neq B$ a množina A je vlastní podmnožinou B. Často je neostrá množina A podmnožinou ostré množiny B, $A \subseteq B$. Grafy, které znázorňují vztahy mezi množinami, označujeme jako VENNOVY diagramy. Na obr. 1 jsou uvedeny pro nejčastější případy vztahů mezi množinami.



Obr. 1: VENNŮV diagram: a.) $A \cup B$ b.) $A \cap B \neq 0$ c.) $A - B$ d.) $A \subset B$

Pro sjednocení, průnik, rozdíl a doplněk ostrých a neostrých množin platí:

Sjednocení množin:	$A \cup B$	obr. 1a
ostré množiny:	$x \in A$ nebo $x \in B$	
fuzzy množiny:	$m_{A \cup B}(x) = \{m_A(x), m_B(x)\}$	
Průnik množin:	$A \cap B \neq \varnothing$	obr. 1b
ostré množiny:	$x \in A$ a současně $x \in B$	
fuzzy množiny:	$m_{A \cap B}(x) = \min \{m_A(x), m_B(x)\}$	
Rozdíl množin:	$A - B$; $A \cap B \neq \varnothing$	obr. 1c
ostré množiny:	$x \in A$; $x \notin A \cap B$	
fuzzy množiny:	$m_{A-B}(x) = m_A(x) - m_{A \cap B}(x)$	
Doplněk množiny:	$\overline{A} = B - A$ pro $A \subset B$	$\overline{A} = B - A$ z obr. 1d
ostré množiny:	$x \notin A$, $x \in B$	
fuzzy množiny:	$m_{\overline{A}}(x) = m_B(x) - m_A(x)$	

2.2 Základy teorie pravděpodobnosti

Základem pro pravděpodobnostní úvahy je představa náhodného pokusu a stabilita relativní četnosti, tj. fenomén statistické regularity. Náhodný experiment je takový, jehož výsledky podléhají náhodě a nelze je předem předpovědět. Výsledky náhodného experimentu při mnohonásobném opakování však nejsou chaotické: existuje totiž fenomén (jev) statistické regularity, že totiž relativní četnost výskytu jednotlivých výsledků při daném uspořádání náhodného pokusu se blíží konstantě. Je to jeden ze základních přírodních zákonů; matematická statistika pak indukční cestou vyvozuje závěry, kterými lze zevšeobecnit poznatky, získané ze souboru výsledků náhodných experimentů.

2.2.1 Pravděpodobnost

Definice pravděpodobnosti se během času vyvíjela. Uvedeme si zde tři nejdůležitější.

A. **Matematická** definice pravděpodobnosti:

Problémy teorie pravděpodobnosti se v této definici formulují s ohledem k množině všech možných výsledků daného náhodného experimentu. Taková množina se nazývá výběrový prostor a označuje S; její prvky, tj. jednotlivé výsledky jsou elementární jevy. Každá podmnožina A_S výběrového prostoru S je jev, prázdná množina $\varnothing$ je *nemožný jev* a celý výběrový prostor S je *jistý jev*.

Množinová funkce P je definována pro třídu všech jevů A_S se nazývá pravděpodobnostní míra a P(A) je pravděpodobnost jevu A, splňuje-li tyto axiomy:

(1) $P(A) \geq 0$ pro všechna $A \in A_S$

(2) $P(S) = 1$ pravděpodobnost jistého jevu je jedna

(3) $P(A \cup B) = P(A) + P(B)$ pro disjunktní A, B, tj. pro $A \cap B = \varnothing$

Pravděpodobnostní prostor (S, A_S, P) představuje matematický model náhodného pokusu.

Z axiomů (1) až (3) vyplývají např. tyto vlastnosti pravděpodobnosti:

(a) $P(\overline{A}) = 1 - P(A)$

(b) $P(\varnothing) = 0$

(c) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ pro $A \cap B \neq \varnothing$

(d) $P(A \cap B) = P(A) \,.\, P(B)$, jsou-li jevy A, B stochasticky nezávislé.

Tato definice pravděpodobnosti předpokládá, že jevové pole obsahuje konečně mnoho náhodných jevů A_S. Speciálním případem tohoto konečného pravděpodobnostního pole je tzv. klasické pravděpodobnostní pole, jehož elementární jevy jsou stejně pravděpodobné. Potom pravděpodobnost jevu A je

$$P(A) = \frac{n(A)}{N} \qquad (2.1)$$

kde n(A) je počet příznivých případů, tj. případů, kdy nastane jev A a N je počet všech možných případů.

Řešený příklad:

Při házení kostkou je množina elementárních jevů, tj. výběrový prostor S = {1,2,3,4,5,6}. Elementární jev, tj. že padne číslo i, pro i = 1,2,3,4,5,6, má pravděpodobnost P(i) = 1/6. Náhodný jev, že padne liché číslo A = {1}∪{3}∪{5}, n(A) = 3, má pak pravděpodobnost P(A) = 3/6 =1/2.

Výpočet pravděpodobností v klasických pravděpodobnostních polích se zakládá na kombinatorických úvahách. Také některé chemometrické problémy, zejména určení informační obsažnosti výsledků kvalitativních experimentů, lze převést na tento případ. Ke

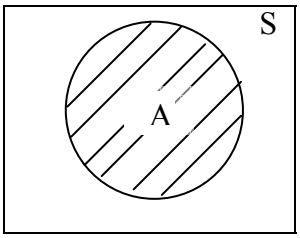
studiu obecnějších úloh, kdy je nutno uvažovat nekonečně mnoho možných jevů, slouží KOLMOGOROVOVA axiomatická teorie pravděpodobnosti, která podává nejdokonalejší definici pravděpodobnosti. Protože vyžaduje podrobnější znalosti teorie množin a základní znalosti teorie míry, nebudeme ji zde uvádět.

B. **Geometrická** definice pravděpodobnosti:

Geometrická definice pravděpodobnosti se používá tehdy, můžeme-li náhodné jevy zobrazit geometricky na přímce, v rovině nebo v prostoru. To znamená, že množina elementárních jevů má nekonečný počet, prvků, vytvářejících určitý geometrický útvar (úsečku, obrazec, oblast), který je kompaktní, tj. omezený a uzavřený. Tento útvar označíme písmenem S. Nechť existuje konečné kladné číslo $\mu(S)$, které je mírou útvaru S. Uvažujeme náhodný jev A, který je podmnožinou S, tj. $A \subset S$, a jehož míra je $\mu(A)$. Potom je geometrická pravděpodobnost jevu A dána zlomkem

$$P(A) = \frac{\mu(A)}{\mu(S)}$$

Mírou intervalu na přímce je jeho délka, mírou oblasti v rovině její plocha a mírou oblasti v prostoru její objem. Při výpočtech různých případů geometrické pravděpodobnosti používáme často grafů, např. VENNOVY diagramy. Na obr. 2. je znázorněna množina elementárních jevů, tj. výběrový prostor S a jev A je vyšrafován.



Obr. 2: Množina elementárních jevů, tj. výběrový prostor, S a jev A

C. **Statistická** definice pravděpodobnosti:

Statistická (VON MISESSOVA) definice pravděpodobnosti nevyžaduje ani apriorní znalosti objektivních vlastností zkoumaného náhodného jevu, ani konečnost počtu elementárních jevů. Je založena na mnohonásobném opakování náhodného pokusu. Provedeme-li n - krát náhodný pokus a v této sérii n pokusů nastal jev A celkem n(A) - krát, potom

$$P(A) = \lim_{n \to \infty} \frac{n(A)}{n}$$

tj. relativní četnost jevu A konverguje k pravděpodobnosti pro velký počet provedených náhodných pokusů.

Poznámka: *Tato konvergence relativní četnosti k pravděpodobnosti má poněkud jiný charakter než konvergence posloupnosti k limitě v matematickém smyslu: jde o konvergenci ve smyslu pravděpodobnostním.*

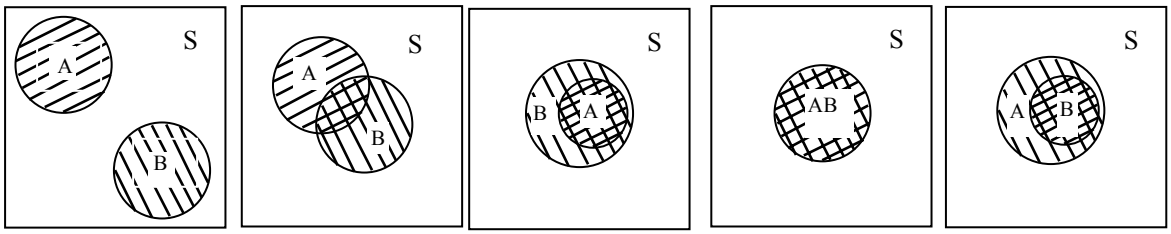
Klasická, geometrická a statistická definice pravděpodobnosti jsou speciálními případy KOLMOGOROVOVY definice.

2.2.2 Podmíněná pravděpodobnost

Je dán pravděpodobnostní prostor (S, A_S, P) a jevy A, B v A_S, přičemž $A \cap B \neq \varnothing$, $P(A) > 0$, pak podmíněná pravděpodobnost funkce

$$P(A \mid B) = \frac{P(A \cap B)}{P(A)}$$

přiřazuje jevu B podmíněnou pravděpodobnost, že nastane B, když nastal jev A. V chemometrii jsou časté případy, kdy $A \subset B$, $A = B$ nebo $B \subset A$. VENNOVY diagramy, které znázorňují různé případy podmíněné pravděpodobnosti, jsou uvedeny na obr. 3.



Obr. 3: VENNOVY diagramy pro různé případy podmíněné pravděpodobnosti.

a. $A \cap B = \varnothing$, $P(B \mid A) = P(A \mid B) = 0$

b. $A \cap B \neq \varnothing$, $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$, $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

c. $A \subset B$; $P(A \mid B) = \frac{P(A)}{P(B)}$; $P(B \mid A) = 1$

d. $A = B$, $P(A \mid B) = P(B \mid A) = 1$

e. $B \subset A$, $P(A \mid B) = 1$, $P(B \mid A) = \frac{P(B)}{P(A)}$

Pro $A = B$ je podmíněná pravděpodobnost $P(A \mid B) = P(B \mid A) = 1$; je jisté že nastane-li A, nastane i B nebo naopak.

BAYESŮV vztah

$$P(A_1 \mid B) = \frac{P(A_1) \cdot P(B \mid A_1)}{\sum_{i=1}^{n} P(A_i) \cdot P(B \mid A_i)} \qquad (2.5)$$

umožňuje vypočítat podmíněnou pravděpodobnost $P(A \mid B)$ z hodnot $P(A)$ a $P(B \mid A)$. Pro případ, kdy můžeme rozlišit příčinu a následek, lze BAYESŮV vztah pokládat za teorém pravděpodobnosti příčin: deterministicky příčinný vztah je pak charakterizován podmíněnými

pravděpodobnostmi $P(A|B) = P(B|A) = 1$. Podmíněná pravděpodobnost má velký význam pro teorii kvalitativního experimentu: charakterizuje např. pravděpodobnost $P(A_i|B)$, že platí hypotéza A_i, jestliže při pokusu, který měl za úkol tuto hypotézu posílit, nastal jev B. Taková hypotéza může být např., že je v analyzovaném vzorku prvek A_i, je-li v emisním spektru čára B, tj. čára o vlnové délce, která odpovídá nejen prvku A_i, ale event. i dalším prvkům.

2.3 Náhodná veličina

Za náhodnou veličinu označujeme takovou veličinu, která působením náhodných vlivů nabývá různých hodnot buď se stejnou nebo s různými pravděpodobnostmi. Tak např. při náhodném pokusu házení jednou kostkou dostáváme pro všechny možné výsledky i = 1,2,.....,6 stejnou pravděpodobnost P(i) = $\frac{1}{6}$; při házení dvěma kostkami dostaneme jednotlivé výsledky, tj. součet na obou kostkách, s různou pravděpodobností: např. součet i = 2 můžeme dostat jen tehdy, padne-li na obou kostkách 1; součet i = 7 může padnout jako 1 + 6, 2 + 5 nebo 3 + 4. Protože počet všech možností při házení dvěma kostkami se rovná počtu variací s opakováním, tj. N = $V_2(6) = 6^2 = 36$, je P(2) = $\frac{1}{36}$, P(7) = $\frac{3}{36} = \frac{1}{12}$. Házení jednou nebo více kostkami je náhodný pokus, ale rozdělení pravděpodobnosti pro jednotlivé výsledky závisí na tom, s kolika kostkami házíme.

2.3.1 Jednorozměrná náhodná veličina

Zobrazme si výběrový prostor S do množiny reálných čísel R_1. To lze provést pomocí náhodné veličiny ξ_S, což je reálná funkce definovaná na S tak, že pro každé reálné číslo x patří jev $A_x = \{S, \xi_S \leq x\}$ do A_S. Ke každé náhodné veličině potřebujeme popisující funkci, která udává rozdělení pravděpodobnosti: nejčastější je kumulativní distribuční funkce a pravděpodobnostní funkce pro nespojité (diskrétní) veličiny, resp. hustota pravděpodobnosti pro spojité náhodné veličiny.

Distribuční funkce náhodné veličiny ξ je funkce $F(x) = P(\xi = x)$, definovaná na R_1, která má tyto vlastnosti:

(1) $\lim_{x \to -\infty} F(x) = 0$; $\lim_{x \to +\infty} F(x) = 1$

(2) F(x) je neklesající funkce;

(3) F(x) je všude zprava spojitá.

O distribuční funkci dále platí:

(a) $P(x_1 < \xi \leq x_2) = F(x_2) - F(x_1)$

(b) $P(\xi > x) = 1 - F(x)$

(c) $P(\xi = x) = F(x) - F(x - 0)$; je-li $F(x)$ spojitá v bodě x, potom $P(\xi = x) = 0$; má-li $F(x)$ v bodě $\xi = x$ skok, představuje $P(\xi = x)$ velikost tohoto skoku.

Diskrétní nespojitá náhodná veličina je taková, jejíž distribuční funkce je schodovitá, skoky v bodech x_i mají velikost $P(\xi = x_i)$. Distribuční funkce je pak

$$F(x) = \sum_{\substack{i \\ x_i \leq x}} P(\xi = x_i)$$

Funkci $F(x_i) = P(\xi = x_i)$ označujeme jako pravděpodobnostní funkci nespojité náhodné veličiny.

Spojitá náhodná veličina je taková, jejíž distribuční funkce je spojitá na R_1 a má derivaci $\frac{dF(x)}{dx} = p(x)$ všude (nebo skoro všude). Distribuční funkce pro spojitou náhodnou veličinu je

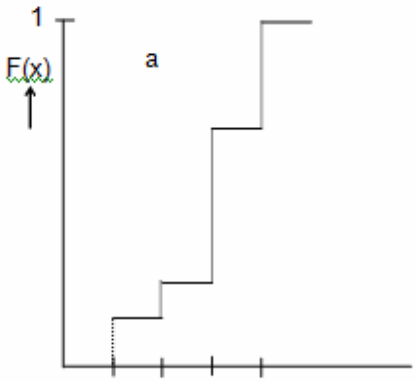
$$F(x) = \int_{-\infty}^{x} p(z)dz$$

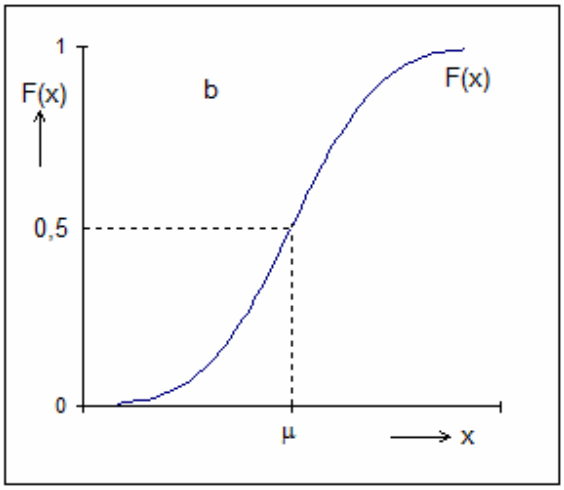
pro $p(z) \geq 0$; přitom platí, že $\int_{-\infty}^{\infty} p(z)dz = 1$.

Hustota pravděpodobnosti je obdobou pravděpodobnostní funkce pro diskrétní náhodnou veličinu; znázorňuje závislost pravděpodobnostního elementu

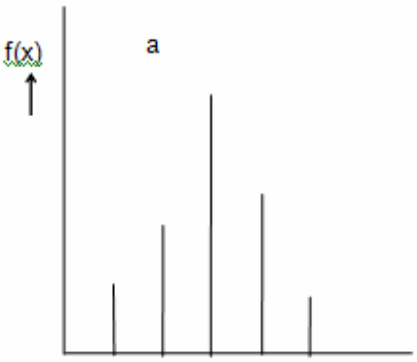
$$\int_{x}^{x+\Delta x} p(z)dz$$

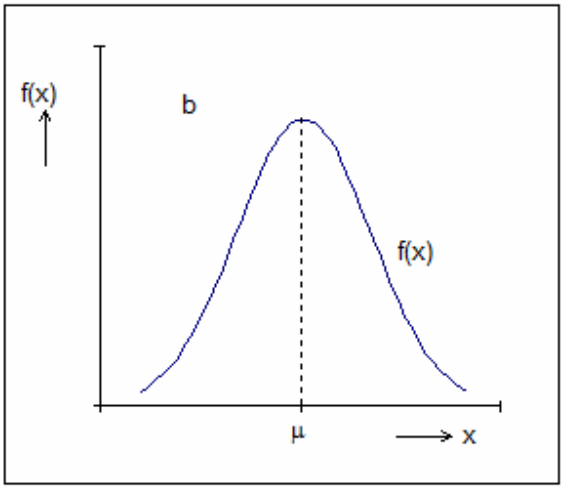
na hodnotě x. Distribuční funkci diskrétní (nespojité) a spojité náhodné veličiny znázorňuje obr. 4.a,b; pravděpodobnostní funkci nespojité náhodné veličiny znázorňuje obr. 5.a; hustotu pravděpodobnosti spojité náhodné veličiny obr. 5.b.





Obr. 4: Distribuční funkce a: nespojité; b: spojité náhodné veličiny





Obr. 5. a: Pravděpodobnostní funkce nespojité náhodné veličiny; b: Hustota pravděpodobnosti spojité náhodné veličiny.

Distribuční funkce představuje model náhodného experimentu, přičemž tento model nám umožňuje výpočet pravděpodobností pro různé hodnoty x, resp. pro různé intervaly hodnot x. Počítání pomocí vztahů (2.6) a (2.7) je však pro praxi poněkud těžkopádné. Každé rozdělení pravděpodobnosti však můžeme charakterizovat určitými čísly, v nichž je zhuštěna informace o vlastnostech tohoto rozdělení. Tyto charakteristiky rozdělení označujeme jako

momenty rozdělení; jde-li o rozdělení výsledků měření charakterizují momenty metrologické charakteristiky vlastností výsledků.

Moment r-tého stupně je

$$m_r = \sum_i x_i^r f(x_i) \qquad \text{pro diskrétní náhodnou veličinu} \qquad (2.8a)$$

$$m_r = \int x^r p(x)dx \qquad \text{pro spojitou náhodnou veličinu} \qquad (2.8a)$$

Vedle těchto obecných momentů zavádíme někdy tzv. centrální momenty r - tého stupně, tj. momenty

$$m(\mu)_r = \sum_i (x_i - \mu)^r f(x_i) \qquad \text{pro diskrétní náhodnou veličinu} \qquad (2.8b)$$

$$m(\mu)_r = \int (x - \mu)^r p(x)dx \qquad \text{pro spojitou náhodnou veličinu} \qquad (2.8b)$$

Nejužívanější je první moment $m_1 = E[\xi]$, tj. očekávaná hodnota rozdělení a druhý centrální moment $m(\mu)_2 = V[\xi]$, tj. rozptyl pravděpodobnostního rozdělení.

Očekávaná hodnota náhodné veličiny μ je číslo, dané jako

$$E[\xi] = \sum_i x_i \cdot f(x_i) \qquad \text{pro diskrétní náhodnou veličinu} \qquad (2.9a)$$

$$E[\xi] = \int x \cdot p(x)dx \qquad \text{pro spojitou náhodnou veličinu} \qquad (2.9a)$$

Očekávaná hodnota charakterizuje polohu rozdělení na ose reálných čísel. Jako metrologická charakteristika představuje výsledek měření nebo opakovaných kvantitativních pokusů.

Rozptyl náhodné veličiny σ^2 je dán jako

$$V[\xi] = \sum_i (x_i - \mu)^2 f(x_i) \qquad \text{pro diskrétní náhodnou veličinu} \qquad (2.9b)$$

$$V[\xi] = \int (x-\mu)^2 p(x)dx \qquad \text{pro spojitou náhodnou veličinu} \qquad (2.9b)$$

Druhá odmocnina rozptylu σ je tzv. směrodatná odchylka, která charakterizuje rozptýlení jednotlivých výsledků okolo očekávané hodnoty; jako metrologická charakteristika určuje přesnost výsledků.

Třetí a čtvrtý centrální moment mají význam jako charakteristiky tvaru rozdělení. Třetí moment, resp. poměr

$$M_3 = \frac{m(\mu)_3}{\sigma^3} \qquad (2.9c)$$

je tzv. koeficient šikmosti. Pro souměrné rozdělení $M_3 = 0$. Čtvrtý moment, resp. koeficient špičatosti („exces“)

$$M_4 = \frac{m(\mu)_4}{\sigma^4} \qquad (2.9d)$$

charakterizuje špičatost rozdělení: čím je rozdělení špičatější, tím větší je jeho hodnota M_4.

Při zpracování experimentálních dat je ovšem nejdůležitější očekávaná hodnota $E[\xi]$ a rozptyl $V[\xi]$. Platí pro ně tato pravidla:

$E[c] = c$; c = konst.	$V[c] = 0$; c = konst.
$E[c\xi] = c.E[\xi]$	$V[c\xi] = c^2.V[\xi]$
$E[\xi \pm \eta] = E[\xi] \pm E[\eta]$	$V[\xi \pm \eta] = V[\xi] + [\eta]$, jsou-li ξ a η nezávislé.

Očekávané hodnoty se pro součet náhodných veličin sečítají, pro rozdíl odečítají; rozptyly se sečítají pro součet i pro rozdíl náhodných veličin. Pro závislé ξ a η závisí rozptyl jejich součtu nebo rozdílu i na „síle“ jejich vzájemné závislosti.

2.3.2 Dvojrozměrná náhodná veličina

Máme dvojici náhodných veličin ξ, η definovaných na S; sdružená distribuční funkce je funkce, definovaná na dvojrozměrném prostoru reálných čísel R_2 jako

$$F(x,z) = P(\xi \le x; \eta \le y)$$

která je pro každou dvojici (x,y) pravděpodobností současného výskytu obou jevů. Má tyto vlastnosti:

(1) $\lim_{x\to-\infty} F(x,y) = F(-\infty,y) = 0$; $\lim_{y\to-\infty} F(x,y) = F(x,-\infty) = 0$

(2) $\lim_{\substack{x\to\infty \\ y\to\infty}} F(x,y) = F(\infty,\infty) = 1$

(3) $F(x,y) + F(x+k,y+h) - F(x+k,y) - F(x,y+h) \ge 0,\ k \ge 0,\ h \ge 0$

(4) F(x,y) je zprava spojitá pro každou proměnnou.

Je-li známa sdružená distribuční funkce dvou náhodných veličin, jsou individuální distribuční funkce dány vztahy

$$F(x) = F(x,\infty)$$

$$F(y) = F(\infty,y)$$

a nazývají se marginální distribuční funkce.

O sdružené distribuční funkci platí, že

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} p(z_x, z_y)\,dz_x\,dz_y$$

kde $p(z_x,z_y)$ je sdružená hustota pravděpodobnosti a marginální hustoty pravděpodobnosti jsou $p(z_x,\infty)$ a $p(z_y,\infty)$, tedy hustoty pravděpodobnosti individuálních náhodných veličin ξ a η.

Náhodné veličiny ξ, η jsou stochasticky nezávislé, jestliže pro $p(x,y) \in R_2$ platí, že pro všechny dvojice (x,y) je

$$F(x,y) = F(x).F(y)$$

$$p(x,y) = p(x).p(y)$$

Podmíněná hustota pravděpodobnosti je pak

$$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

a platí, že $p(x,y) = p(x \mid y).\ p(y)$, $p(y) > 0$.

2.3.3 Některá rozdělení pravděpodobnosti

Zde si ukážeme rozdělení, důležitá pro náš další výklad a uvedeme jejich očekávané hodnoty a rozptyl, příp. jejich další vlastnosti. Napřed uvedeme dvě diskrétní rozdělení, pak několik spojitých.

Binomické rozdělení Bi(p,n): Provádíme n pozorování, přičemž se v každé z nich může, ale nemusí objevit jev A. Přitom je pravděpodobnost výskytu A ve všech nezávislých pozorováních stejná a rovna p. Počet výskytu A těchto n pozorováních je diskrétní náhodná veličina, která nabývá hodnot
x = 0,1,2,.....,n. Pravděpodobnost

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

a distribuční funkce $F(x) = \sum_x P_n(x)$. Binomický koeficient je pro $x \le n$, x, n celá, nezáporná, dán jako výraz $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ a dá se snadno určit pomocí tzv. Pascalova trojúhelníku.

Očekávaná hodnota pro Bi(p,n) je $\mu = n \cdot p$ a rozptyl je $\sigma^2 = n \cdot p(1-p)$. Binomické rozdělení má význam v teorii chyb, zejména tehdy, uvažujeme-li o pravděpodobnosti vzniku maximální chyby.

POISSONOVO rozdělení Po(λ) je rozdělením málo četných diskrétních jevů; uplatňuje se např. při hodnocení radiometrických měření a analýz, při počítání kolonií bakterií nebo krvinek apod.

Pravděpodobnost

$$P_\lambda(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

a distribuční funkce $F(x) = \sum_x P_1(x)$. Očekávaná hodnota pro POISSONOVO rozdělení je $\mu = 1$ a jeho rozptyl je $\sigma^2 = \lambda$. Pro hodnotu $\lambda \geq 12$ můžeme POISSONOVO rozdělení aproximovat normálním $N(\lambda,\lambda)$.

Ze spojitých jsou pro chemometrii důležitá tato rozdělení:

Uniformní (rektangulární) $U(x_1,x_2)$, které má hustotu pravděpodobnosti

$$p(x) = \begin{cases} = \dfrac{1}{x_2 - x_1} \text{ pro } x \in \langle x_1, x_2 \rangle \\ = 0 \text{ pro } x \notin \langle x_1, x_2 \rangle \end{cases}$$

Je vhodným modelem rozdělení spojité náhodné veličiny, kdy víme jen to, že nabývá hodnot od x_1 do x_2 se stejnou pravděpodobností, ale mimo tuto oblast nemůže vzniknout žádná jiná hodnota (např. koncentrace od 0 do 100%). Očekávaná hodnota $\mu = \frac{1}{2}(x_1 + x_2)$ a rozptyl $\sigma^2 = \frac{1}{12}(x_1 - x_2)^2$.

Normální (GAUSSOVO) rozdělení $N(\mu,\sigma^2)$ je pro chemometrii nejdůležitějším rozdělením. Má hustotu pravděpodobnosti

$$p(x) = \frac{1}{6\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]; \qquad -\infty < x < +\infty$$

Transformací $z = \frac{x-\mu}{\sigma}$, což je tzv. normovaná náhodná veličina, dostáváme hustotu pravděpodobnosti normálního rozdělení N(0,1), tj.

$$\varphi(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z^2\right)$$

Distribuční funkce $\varnothing(z) = \int_0^{z>0}\varphi(z)dz$ je tabelována (viz. tab. 2 [1]). Určení $F(z) = \int_{-\infty}^{z}\varphi(z)dz$, resp. souměrného $F(-z,+z) = \int_{-z}^{+z}\varphi(z)dz$ z tabelovaných hodnot je velmi snadné. Očekávaná hodnota μ a rozptyl σ^2 normálního rozdělení jsou navzájem nezávislé, koeficient šikmosti $M_3 = 0$ a exces $M_4 = 3$. Modus, tj. hodnota x, pro kterou nabývá hustota pravděpodobnosti p(x) maximální hodnotu, je pro normální rozdělení $x_M = \mu$.

Normální rozdělení je vhodné jako teoreticky zdůvodnitelný a v praxi osvědčený model pravděpodobnostního rozdělení pro mnohá fyzikální měření, pro výsledky chemických analýz středních a vyšších obsahů apod. Ostatně, mnohé metody statistické indukce, založené na normálním rozdělení, jsou robusní, tj. poskytují přibližně správné výsledky i v případě, že skutečné rozdělení se poněkud odchyluje od přesně normálního.

Od normálního rozdělení jsou odvozena dvě další, pro teorii experimentu důležitá rozdělení, a to logaritmicko-normální a zkomolené („useknuté“) normální rozdělení.

Logaritmicko-normální (lognormální) rozdělení LN (μ,σ^2) je takové rozdělení, kdy nikoliv výsledky, ale jejich logaritmy jsou rozděleny normálně. Má hustotu pravděpodobnosti

$$p(x) = \begin{cases} = 0 & x \le 0 \\ = \dfrac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right] & x > 0 \end{cases}$$

Očekávaná hodnota $E[\xi] = \exp(\mu + \frac{1}{2}\sigma^2)$ a rozptyl $V[\xi] = \exp(2\mu + \sigma^2)(e^{\sigma^2} - 1)$ nejsou ovšem vzájemně nezávislé. Modus lognormálního rozdělení $x_M = \exp(\mu + \sigma^2)$, tj. $x_M = E[\xi]$.

Pro praktické účely často velmi dobře vyhovuje posunuté logaritmicko-normální rozdělení $LN(\mu^2; x_0)$ s hustotou pravděpodobnosti

$$p(x) = \begin{cases} = 0 & x \le x_0 \\ = \dfrac{1}{(x-x_0)\sigma\sqrt{2\pi}} \exp\left[-\dfrac{1}{2}\left(\dfrac{\ln(x-x_0)-\mu}{\sigma}\right)^2\right] & x > x_0 \end{cases}$$

které má počátek v $x_0 \ge 0$. Pak je výhodné definovat $\mu = \ln k.x_0$, kde k je parametr asymetrie. Očekávaná hodnota $E[\xi] = x_0 + \exp(\mu + \frac{1}{2}\sigma^2)$, rozptyl je $V[\xi] = k.x_0.\exp(\sigma^2)(\exp(\sigma^2) - 1)$. Posunuté lognormální rozdělení předpokládáme u rozdělení výsledků spektrální analýzy, což lze vysvětlit tím, že SCHEIBEHO-LOMAKINŮV vztah (závislost mezi tokem záření a koncentrací emitujícího prvku) je logaritmický.

<u>Zkomolené („useknuté“) normální rozdělení</u> $TN(\mu,\sigma^2;x_0)$, tj. normální rozdělení „useknuté“ v bodě x_0 má hustotu pravděpodobnosti

$$p_T(x) = \begin{cases} = \dfrac{\exp\left[-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right]}{[1-F(x_0)]\sigma\sqrt{2\pi}} & x > x_0 \end{cases}$$

$= 0 \qquad\qquad x \le x_0$

kde $F(x_0)$ je distribuční funkce normálního rozdělení pro x_0. Tvar „useknutého“ normálního rozdělení má očekávanou hodnotu

$E[\xi] = \mu + \frac{p(z_0)}{1-F(z_0)} \cdot \sigma$, kde $p(z_0)$ je hustota pravděpodobnosti normálního rozdělení pro normovanou náhodnou veličinu $z_0 = \frac{x_0-\mu}{\sigma}$ a $F(z_0)$ je jeho distribuční funkce. Toto rozdělení dobře vystihuje distribuci výsledků stopových analýz blízko meze důkazu.

Při testování vlastností normálně rozdělených výsledků se používají další rozdělení: Studentovo t-rozdělení, F a χ^2 - rozdělení. Jejich použitím se budeme zabývat až v odstavci o statické indukci; zde si podáme jenom jejich zcela stručnou charakteristiku.

Studentovo t-rozdělení (*) je rozdělením veličiny

$$t = \frac{|\bar{x}-\mu|}{\sigma}\sqrt{n}$$

a závisí na počtu stupňů volnosti $f = n - 1$, kde n je počet výsledků, z nichž je určen průměr $\bar{x}$. Je symetrické, velmi podobné normálnímu rozdělení, ale je méně špičaté; pro $f \to \infty$ přechází v normální, $t = z$.

(*) *Spojité t-rozdělení označujeme často jako „Studentovo“, poněvadž je počátkem 20. století publikoval W. GOSSET pod pseudonymem „Student“ v době, kdy ještě sám studoval.*

F-rozdělení je rozdělení poměru

$$F = \frac{s_1^2}{s_2^2},\ s_1^2 \ge s_2^2$$

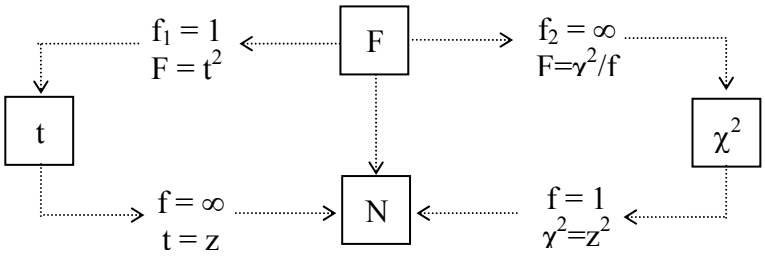
kde s_2 je odhad rozptylu (viz. odst. 4.1.1). Závisí na dvou počtech stupňů volnosti, $f_1 = n_1 - 1$; $f_2 = n_2 - 1$, kde n_1, n_2 je počet výsledků, z nichž je určen odhad rozptylů s_1^2, s_2^2. F-rozdělení je nesymetrické, ale pro $f_1 = 1$ a $f_2 = \infty$ přechází v normální, $F = z^2$.

Rozdělení χ^2 (čti: chý kvadrát) je rozdělením výrazu

$$\chi^2 = \sum_i \left(\frac{x_i - \bar{x}}{\sigma} \right)^2$$

kde x_i má normální rozdělení. Je to nesymetrické rozdělení, závislé na počtu stupňů volnosti $f = n - 1$; pro $f = 1$ přechází v normální rozdělení, $\chi^2 = z^2$.

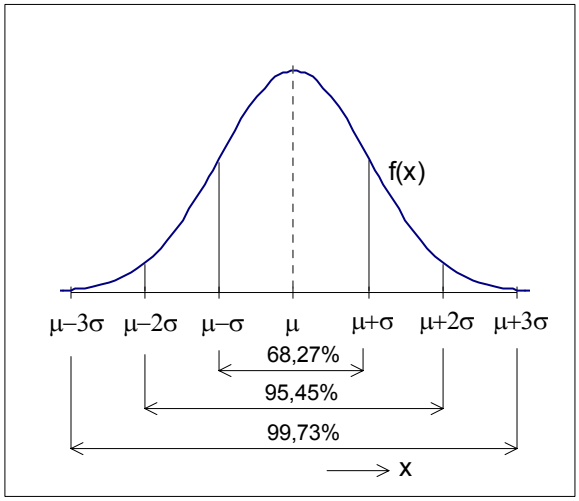
Testovací rozdělení pro jisté limitní podmínky přecházejí v normální rozdělení tak,



jak to ukazuje toto schéma:

Toto schéma také ukazuje význam normálního rozdělení jakožto rozdělení limitního pro F, t a χ^2-rozdělení. Také POISSONOVO a binomické rozdělení za určitých podmínek přecházejí v normální rozdělení.

Normální rozdělení ukazuje obr. 6.



Obr. 6: Normální rozdělení

2.4 Centrální limitní věta

Rozdělení náhodné veličiny ξ, která je součtem n nezávislých náhodných veličin ξ_i, i = 1,2.....,n, tj.

$$\xi = \sum_{i=1}^{n} \xi_i$$

přičemž jednotlivé veličiny ξ_i mají očekávanou hodnotu μ_i a rozptyl σ_i^2, se v limitě blíží normálnímu rozdělení s parametry

$$\mu = \sum_{i=1}^{n} \mu_i \; ; \qquad \sigma^2 = \sum_{i=1}^{n} \sigma_i^2 \; .$$

Nejjednodušší je případ, kdy mají všechny veličiny ξ_i stejné rozdělení. Tak např. mají-li binomické rozdělení Bi(p,n), má ξ pro $n \to \infty$ normální rozdělení N(p,p(1-p)); mají-li POISSONOVO rozdělení Po(λ), má v limitě ξ rozdělení N(λ,λ).

Pro teorii experimentu je důležité, že centrální limitní věta platí i pro průměr: to znamená, že náhodná veličina

$$\xi = \frac{1}{n} \sum_{i=1}^{n} \xi_i$$

má v limitě normální rozdělení $N(\mu, \frac{\sigma^2}{n})$. Z toho plyne, že průměr n - kráte opakovaných měření se stejnou očekávanou hodnotou μ a s rozptylem σ^2 má v limitě vždy normální rozdělení s parametry μ a σ^2/n, resp. se směrodatnou odchylkou $\sigma/\sqrt{n}$, a to bez ohledu na rozdělení jednotlivých ξ_i. Platnost centrální limitní věty je důvodem, proč při zpracování výsledků měření a chemických analýz zpravidla používáme jako konečný výsledek aritmetický průměr

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

a předpokládáme jeho normální rozdělení se směrodatnou odchylkou $\sigma/\sqrt{n}$.

3 TEORIE CHYB

Členění kapitoly:

- **Klasifikace chyb**
- **Zákon hromadění (šíření) chyb**
- **Hromadění chyb chemických experimentů**

Čas potřebný ke studiu: 360 minut

Cíl: Po prostudování této kapitoly

- **pochopíte základní zákonitosti teorie chyb**
- **seznámíte se základními typy chyb**
- **seznámíte se s pojmy přesnost a správnost**

Výklad

Systém, ve kterém probíhá chemický experiment, je vždy systémem difúzním, tj. jeho výstup (získaná informace) má jistou neurčitost. Tato neurčitost může mít různé složky: u výsledků měření jsou to např. chyby, v případě kvalitativních experimentů nedokonalá rozlišitelnost jednotlivých výsledků apod. Vznik neurčitosti si vysvětlujeme tak, že se během provádění pokusu v různé míře uplatňují různé nahodilé okolnosti: proto nejčastěji popisujeme neurčitost experimentálně zjištěné informace pomocí pojmů teorie pravděpodobnosti. Jen málokdy můžeme zjistit příčinu této neurčitosti, spíše jde většinou o souhru mnoha příčin, z nichž některé ani neznáme. Proto také tyto příčiny ani nehledáme, ale pokládáme experimentální systém za stochastický, tj. takový, že při opakování pro tentýž vstup dostáváme na výstupu soubor hodnot řídící se určitým pravděpodobnostním rozdělením.

Výsledkem experimentu je údaj, který může být veličinou kvalitní (nominální nebo ordinální) nebo kvantitativní (kardinální). Veličiny nominální lze označit názvem, symbolem, vzorcem apod., ale nelze je seřadit; ordinální můžeme seřadit nebo zařadit do různých úrovní, aniž lze jednotlivým úrovním přiřadit číselnou hodnotu. Kvantitativní (kardinální) veličiny můžeme vyjádřit číselně v určitých jednotkách; toto číselné vyjádření se provádí měřením.

Výsledek měření může být zatížen chybou. Chyba se vyjadřuje jako rozdíl mezi experimentálně zjištěnou x a skutečnou hodnotou X měřené veličiny, tj. jako

$$d = (x - X) \tag{3.1}$$

Vedle absolutní chyby podle (3.1) určujeme často i chybu relativní; tj.

$$e = \frac{d}{X} = \frac{x - X}{X} \tag{3.2a}$$

kterou někdy vyjadřujeme v procentech jako

$$e(\%) = 100\ \frac{d}{X} \qquad (3.2b)$$

Protože zpravidla neznáme skutečnou hodnotu X, slouží absolutní chyba d podle (3.1) a relativní chyba e, resp. e(%) podle (3.2) spíše jako teoretický pojem než jako metrologická charakteristika. Takovou charakteristikou je odchylka

$$\Delta = (x_i - \bar{x}) \qquad (3.3)$$

kde x_i je výsledek i-tého měření a

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\ , \qquad i = 1,\ldots\ldots, n$$

je aritmetický průměr. Někdy můžeme absolutní chybu odhadnout úvahou: např. odečítáme-li na stupnici přístroje, bude d = 0,5 dílku, při odměřování roztoku je chyba dána s tolerancí podle cejchovních předpisů a možnou chybou paralaxy odečítání apod.

3.1 Klasifikace chyb

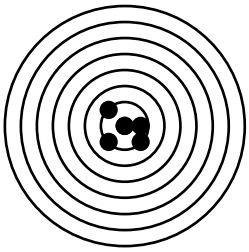
Chyby, kterými mohou být zatíženy výsledky měření nebo chemických kvantitativních analýz jsou:

1. Chyby náhodné: vyskytují se nepravidelně, mohou mít kladné i záporné znaménko a bývají malé takže nezkreslují výsledky oproti skutečné hodnotě; způsobují jen, že se výsledky opakovaných měření mezi sebou poněkud liší.

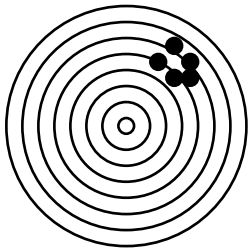
2. Chyby soustavné (systematické): mají pravidelný charakter a zkreslují výsledky vždy v určitém směru; vznikají buď použitím nesprávné experimentální techniky nebo nesprávným používáním správné laboratorní techniky.

3. Chyby hrubé vznikají jako důsledek nedopatření nebo malé pečlivosti pracovníka; bývají jimi zatíženy jednotlivé výsledky z celého souboru dat.

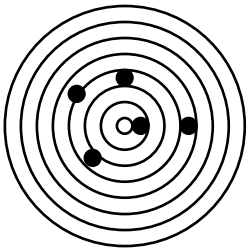
Podle toho, jaké chyby zatěžují výsledky, rozlišujeme jejich přesnost a správnost.



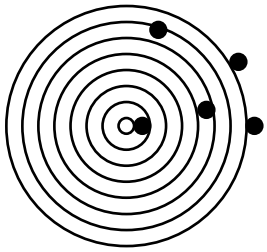
Správné a přesné



Nesprávné ale přesné



Správné ale nepřesné



Nesprávné a navíc nepřesné

Obr. 7: Přesnost a správnost

Správné jsou takové výsledky, které se v průměru dobře shodují se skutečnou hodnotou , tj. jsou zatíženy jen náhodnými chybami a přesné jsou výsledky, které se vzájemně dobře shodují, ale mohou se od skutečné hodnoty lišit o soustavnou chybu. V praxi nelze vždy jednotlivé druhy chyb bezpečně odlišit, poněvadž je mezi nimi spojitý přechod. Tak např. náhodné a hrubé chyby se liší jen velikostí a soustavné a náhodné jen pravidelností výskytu v určitém směru.

Matematickým modelem správných výsledků, zatížených jenom náhodnými chybami ε_i je vztah

$$x_i = x_0 + \varepsilon_i \qquad (3.4)$$

kde x_0 je hodnota, nezatížená náhodnou chybou. Pro podmínku

$$\sum_{i=1}^{n} \varepsilon_i = 0, \qquad i = 1,.....,n$$

vede součet n rovnic (3.4) k výrazu

$$\sum_{i=1}^{n} x_i = n \,.\, x_0$$

a hodnotu x_0 určíme jako

$$x_0 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

tj. jako aritmetický průměr. Jsou-li však výsledky zatíženy soustavnou chybou δ, je

$$x_i = x_0 + \delta + \varepsilon_i$$

a aritmetický průměr $\bar{x} = (x_0 + \delta)$. Aritmetický průměr tedy kompenzuje pouze náhodné, nikoliv však soustavné chyby. Má-li výsledek x_i normální rozdělení $N(x_0,\sigma^2)$, má náhodná chyba ε rozdělení $N(0, \sigma^2)$ a soustavná chyba $\delta = (X - x_0)$, kde X je správná (skutečná) hodnota měřené veličiny. Jednotlivé veličiny ukazuje např. obr. 7 v literatuře [1].

Chyby jsou závažnou, ale nikoliv jedinou složkou neurčitosti výsledků měření. Některé problémy, které nemohla vyřešit teorie chyb, založená pouze na pravděpodobnostním přístupu, vyřešila v poslední době teorie informace.

3.2 Zákon hromadění (šíření) chyb

Počítáme-li veličinu $x = f(x_1, .. , f_n)$ z hodnot $x_1, .. x_n$ získaných měřením, hromadí se chyby těchto hodnot v chybě konečné veličiny x. Jsou-li hodnoty x_i zatíženy absolutní chybou d_i, resp. relativní chybou e_i, platí tzv. zákon hromadění (šíření) chyb. Podle tohoto zákona je pro $x = f(x_1, .., x_n)$ kde $x_1, x_2 .. x_n$ jsou vzájemně nezávislé výsledky měření, je absolutní chyba veličiny x dána vztahem

$$d = \left(\frac{\partial f(x_1,...,x_n)}{\partial x_1}\right) d_1 + ... + \left(\frac{\partial f(x_1,...,x_n)}{\partial x_n}\right) d_n \qquad (3.6)$$

a její relativní chyba vztahem

$$e = \left(\frac{\partial \ln f(x_1,...,x_n)}{\partial x_1}\right) d_1 + ... + \left(\frac{\partial \ln f(x_1,...,x_n)}{\partial x_n}\right) d_n \qquad (3.7)$$

Pro nejčastější případy platí:

Součet nebo rozdíl: $x = x_1 \pm x_2$

$$d = d_1 + d_2 \qquad (3.8)$$

tj. sečítají se absolutní chyby, a to i pro rozdíl;

Součin nebo podíl: $x = x_1 . x^{\pm 1}$

$$e = e_1 + e_2 \qquad (3.9)$$

tj. sečítají se relativní chyby. Chyby se tedy vždy sečítají, i když jde o rozdíl nebo podíl; viz též pravidla o očekávané hodnotě a rozptylu v odst. 2.3.1. Vztahy (3.8) a (3.9) platí pouze pro případ nezávislých hodnot x_i: pro korelované hodnoty x_i platí rozšířený zákon hromadění chyb.

Při výpočtu podle zákona hromadění chyb určujeme vždy maximální chybu; tam, kde je $x = f(x_1, .., x_n)$, je pro větší n pravděpodobnost, že se všechny dílčí chyby sečtou, poměrně malá. Pravděpodobnost, že se všechny chyby sečtou nebo do jaké míry se kompenzují, je dána jako $P_n(x)$ podle (2.10) pro binomické rozdělení. Pravděpodobnost, že vznikne maximální chyba, je

$$P_n = \left(\frac{1}{2}\right)^{n-1}$$

a rychle klesá se stoupajícím počtem n veličin, z nichž se určuje konečný výsledek. Zákon šíření chyb platí i pro rozptyly (viz. odst. 2.3.1); pak ale je výsledkem opět rozptyl (a ne nějaký „maximální“ rozptyl).

Hlavní praktický význam zákona šíření chyb není ve výpočtu skutečných hodnot chyb, ale v hledání podmínek, za nichž má určitý experimentální postup minimální chybu výsledku, tj. umožňuje provádění analýzy chybové stránky určitého postupu.

Řešený příklad:

Máme-li navážit 50 mg, je relativní chyba diferenčního vážení e(%) = (2.0,2.100)/50 = 0,80%; naváží-me-li 500 mg, rozpustíme, zředíme na 250 ml a pipetujeme 25 ml, je relativní chyba e(%) = (2.0,2.100)/500 + 0,29.100/250 + 0,056.100/25 = 0,42%; a to ještě taková chyba vznikne s pravděpodobností pouze $P_n = 0,5^3 = 0,125$, takže s pravděpodobností P = (1 - 0,125) = 0,875 můžeme očekávat, že skutečná relativní chyba bude menší než vypočtená 0,42%, protože se některé dílčí chyby vzájemně vykompenzují.

3.3 Hromadění chyb chemických experimentů

Při provádění analýzy chybové stránky experimentů, prováděných v chemické laboratoři, se uplatňují i konstanty, nezbytné pro interpretaci výsledků. Jsou to např. ekvivalenty v odměrné nebo přepočítací faktory ve vážkové analýze, rovnovážné a rychlostní konstanty apod. V matematickém smyslu ovšem nejde o skutečné konstanty; jsou to však hodnoty, experimentálně určené daleko přesněji nežli ostatní experimentální data a tak s nimi při dosazování do vztahů, vyplývajících ze zákona hromadění chyb, zacházíme jako s konstantami. Pak má na konečnou chybu vliv nikoliv jejich přesnost, ale pouze jejich hodnota.

Řešený příklad:

Vyvažujeme-li při gravimetrickém stanovení hliníku vyžíhaný Al_2O_3 (přepočítávací faktor f_1 = 0,529) a dopouštíme se chyby vážení 0,2 mg, je absolutní chyba určení hmotnosti hliníku d_1 = 0,2.0,529 = 0,1058 mg Al. Vyvažujeme-li oxinát hlinitý (přepočítávací faktor f_2 = 0,0578), je absolutní chyba stanovení hmotnosti hliníku d_2 = 0,2 . 0,0578 = 0,01156 mg Al. Zdálo by se tedy, že metoda s menším přepočítávacím faktorem je pro stanovení vždy výhodnější. Určeme si však celkovou chybu výsledku analýzy. Máme-li určit vážkově 1% Al amoniakální metodou, pak z navážky 1 g dostaneme 0,0189 g Al_2O_3. Relativní chyba stanovení, vážíme-li diferenčně, je e(%) = (2.0,2.100)/1000 + (2.0,2.100)/18,9 = 2,15 %; vyloučíme-li však hliník oxinem, vyvažujeme z navážky 0,600 g asi 0,104 g oxinátu, takže e(%) = (2.0,2.100)/600 + (2.0,2.100)/104 = 0,45%. Analyzujeme-li však vzorek s obsahem 60% Al, musíme navážit 0,176 g vzorku, abychom vyvažovali 0,20 g Al_2O_3 a e(%) = (2.0,2.100)/176 + (2.0,2.100)/200 = 0,43%, kdežto při oxinátové metodě smíme navážit jen 19,3 mg vzorku, abychom nedostali více než 200 mg oxinátu (větší množství se obtížně suší) a relativní chyba celého stanovení e(%) = (2.0,2.100)/19,3 + (2.0,2.100)/200 = 2,27%. Můžeme ovšem navážit 193 mg

vzorku, rozpustit, doplnit na 250 ml a pipetovat k analýze 25 ml; pak bude relativní chyba e(%) = 0,75%, tedy menší, ale stále výrazně větší než při použití amoniakální metody. Uvedený příklad zároveň dokládá známou zkušenost, že metoda, vhodná např. pro stanovení malého obsahu určité složky může být zcela nevhodná pro stanovení velkého obsahu téže složky.

Ve výpočtech maximální relativní chyby za použití zákona hromadění chyb není ovšem pamatováno na možné chyby postupu. Někdy je účelné odvodit chybovou funkci, tj. závislost konečné chyby na chybách dílčích výsledků a na hodnotách konstant. Pro některé „standardní“ postupy, např. analytické metody, jsou tyto chybové funkce publikovány a při analýze chybové stránky takové analytické metody stačí dosadit do těchto funkcí odpovídající hodnoty. Použití takové chybové funkce si ukážeme na příkladu.

Řešený příklad:

Pro chelatometrickou titraci uvádí R. Přibil tzv. ukazatel chyby

$$p = 1 - U/c + \alpha_H /UK,$$

z něhož se určí relativní chyba jako e(%) = (p - 1)100; zde je U citlivost indikátoru, kterou určíme pokusně, c je koncentrace titrovaného kovu, K je konstanta stability komplexu kovu s chelatonem a α_H je koeficient vedlejší reakce s H^+ - ionty. Určeme podle této chybové funkce relativní chybu titrace 20 ml 0,01 M roztoku Mg^{2+} roztokem 0,01 M chelatonu 3 na eriochromčerň T při pH = 9,5.

Pokusně byla určena hodnota $U = 5.10^{-6}$, z tabulek najdeme log K = 8,69 a pro pH = 9,5 je log α_H = 0,83. Při konci titrace je c_M = (0,01.20)/40 = 0,005. Pak je p = 1 - 0,001 + 0,00275 = 1,00175 a relativní chyba titrace e(%) = (p - 1)100 = (1,00175 - 1)100 = 0,175%.

4 MATEMATICKÁ STATISTIKA

Členění kapitoly:

- **Statistický odhad (bodový, intervalový)**
- **Statistické testování**
- **Závislost dvou proměnných**
- **Vícenásobná lineární regrese**

Čas potřebný ke studiu: 1200 minut

Cíl: Po prostudování této kapitoly

- **pochopíte základní zákonitosti statistického odhadu a statistického testování**
- **seznámíte se s lineární, nelineární a vícenásobnou lineární regresi**

Výklad

Zatímco se matematická statistika uplatňuje v teoretické chemii jako prostředek vytváření teoretických systémů (např. statistická termodynamika), které umožňují provádět dedukci, je pro chemometrii důležitým úkolem statistiky indukce, prováděná na základě odhadů neznámých parametrů, získaných z náhodných výběrů a podle výsledků testování hypotéz o těchto odhadech. Důležitým úkolem matematické statistiky je vyšetřování závislosti dvou veličin (regrese a korelace).

Určujeme-li hodnotu náhodné veličiny v řadě n nezávislých opakování experimentů, dostaneme náhodný výběr $(x_1,.....x_n)$, kde x_i, i = 1,2, ... n jsou realizace náhodné veličiny ξ s distribuční funkcí F(x). Takový náhodný výběr musí splňovat dvě základní podmínky:

1. Musí být reprezentativní, tj. každý prvek základního souboru musí mít stejnou pravděpodobnost, že se dostane do výběru;

2. Jednotlivé prvky výběru musí být vzájemně nezávislé.

Přitom není důležité, zda základní soubor, z něhož výběr provádíme, je reálný (např. bedna součástek stejného druhu) nebo hypotetický (soubor všech možných výsledků určitého měření, provedených na daném objektu).

Zatímco charakteristika polohy μ a rozptyl σ^2 základního souboru jsou pevné hodnoty, dané vztahy (2.8.a,b) a (2.9.a,b), jsou výběrové charakteristiky (výběrové „statistiky“), např. výběrový průměr nebo výběrový rozptyl, náhodné veličiny a mají své určité pravděpodobnostní rozdělení. Je zřejmé, že odhad parametrů základního souboru pomocí výběrových statistik je tím přesnější, čím větší je rozsah výběru, tj. počet výsledků n.

4.1 Statistický odhad

Protože je statistický odhad náhodná veličina, může se jeho hodnota, určená z náhodného výběru rozsahu n, lišit od skutečné hodnoty odhadovaného parametru. K posuzování vlastností odhadů slouží tato kriteria:

1. Konzistentnost: Odhad je konzistentní, jestliže s rostoucím rozsahem výběru se zmenšuje rozdíl mezi odhadem a skutečnou hodnotou parametru.

2. Nestrannost: Odhad je nestranný, když při opakovaných výběrech kolísá jeho hodnota symetricky, stejně na obě strany kolem teoretické hodnoty parametru. Takový nestranný odhad ani při malém n soustavně nepodhodnocuje ani nenadhodnocuje odhadovaný parametr.

3. Vydatnost: Odhad je vydatný, když se jeho rozptyl okolo skutečné hodnoty parametru rychle zmenšuje s rostoucím rozsahem výběru n.

4. Robustnost: Odhad je robustní, není-li příliš závislý na malých odchylkách od předpokládaného rozdělení pravděpodobnosti.

Dále se budeme zabývat odhady parametru μ a σ^2 pro výběr z normálně rozděleného základního souboru a ukážeme si na jejich vlastnostech některá obecně platná pravidla.

4.1.1 Bodové odhady

Odhady, které jsou dány jediným číslem, označujeme jako bodové. Bodovým odhadem parametru μ z náhodného výběru $(x_1, ..., x_n)$ může být libovolná, náhodně zvolená hodnota x_i. Takový odhad ale určitě nebude vydatný ani robusní a nemusí být ani nestranný. Lepším odhadem bude medián, tj. střední hodnota z výběru seřazeného podle velikosti, kdy $x_1 \leq x_2 \leq .. \leq x_n$; pro liché n je medián

$$\tilde{x}=x_{\frac{n}{2}+\frac{1}{2}} \qquad (4.1a)$$

a pro n sudé je

$$\tilde{x}=\frac{1}{2}(x_{\frac{n}{2}}+x_{\frac{n}{2}+1}) \qquad (4.1b)$$

Tak např. pro n = 7 je medián $x_{3,5+0,5} = x_4$, tedy čtvrtý výsledek nebo pro n = 8 je to průměr ze čtvrtého a pátého výsledku. Medián je nestranný a robustní odhad, ale není příliš vydatný; mimoto při jeho určení z velkého souboru dat zcela ztrácíme informace, obsažené ve všech mimo jediného, resp. dvou výsledcích, protože je do výpočtu mediánu vůbec nezahrnujeme. Vydatnost mediánu závisí na rozsahu výběru a je pro sudé hodnoty n větší než pro sousední liché.

Nejčastěji používaným odhadem očekávané hodnoty μ jsou různé průměry. Nejběžnější je aritmetický průměr

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

Je nestranným, velmi vydatným odhadem očekávané hodnoty, který skoro úplně odstraňuje vliv náhodných chyb; není však, zejména pro menší n, příliš robustní. Malou robustnost průměru můžeme obejít několikerým způsobem:

1. Jako odhad μ počítáme průměr $\bar{x}$ a zároveň medián $\tilde{x}$: je-li rozdíl mezi nimi malý, uvádíme jako konečný výsledek průměr. Je-li však tento rozdíl velký, je výhodné se přesvědčit, zda krajní hodnoty souboru nejsou odlehlé (viz odst. 4.2) a příp. je vyloučit. Souhlasí-li pak $\bar{x}$ a $\tilde{x}$, uvádíme tento nový průměr, vypočtený po vyloučení krajních odlehlých hodnot, jako konečný výsledek.

2. Počítáme průměr redukovaného souboru jako

$$\bar{x}_{red}=\frac{1}{n-2k}\sum_{i=k}^{n-k}x_i \qquad (4.3)$$

nebo tzv. winsorizovaný průměr, kdy seřazený soubor $x_1,x_2 .. x_{n-1},x_n$ nahradíme souborem $x_2,x_2,x_3 .. x_{n-2},x_{n-1},x_{n-1}$ a počítáme průměr

$$\bar{x}_{win} = \frac{1}{n}\sum_{i=2}^{n-1} x_i \qquad (4.4)$$

Soubor někdy redukujeme nebo winsorizujeme pouze jednostranně, např. tak, že určíme průměr i medián, a pokud se neshodují, vypouštíme nebo nahradíme nejbližším ten krajní výsledek, který má větší odchylku od mediánu. Průměry (4.3) a (4.4) patří mezi tzv. robustní statistiky, kterých dnes známe velké množství; v chemometrické praxi se však uplatňují hlavně ty, které zde byly uvedeny. Je zřejmé, že uvedené robustní odhady jsou zaměřeny hlavně na odstranění nesymetrie rozdělení.

Jiným odhadem μ je tzv. vážený průměr

$$\bar{x}_w = \frac{\sum_{i=1}^{n} x_i \cdot w_i}{\sum_{i=1}^{n} w_i}, \qquad i = 1,2, .. n \qquad (4.5)$$

který umožňuje jednotlivým výsledkům x_i přiřadit různou „váhu“ w_i; tak např. výsledkům, více vzdáleným od mediánu, přiřadíme menší váhu než těm, které jsou mu blíže. Pro $w_i = 1$ pro všechna i přechází $\bar{x}_w$ v aritmetický průměr $\bar{x}$ podle vztahu (4.2).

Odhadem očekávané hodnoty lognormálního rozdělení je geometrický průměr

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \ldots x_n} \qquad (4.6a)$$

Často určujeme jeho logaritmus

$$\log \bar{x}_g = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

a ten pak odlogaritmujeme.

Konsistentním, nestranným a dosti vydatným odhadem parametru σ je směrodatná odchylka

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{f}\sum_{i=1}^{n}\Delta_i^2} \qquad (4.7a)$$

kde počet stupňů volnosti f = n - 1 (viz pozn. *) a Δ_i je odchylka i-tého výsledku od průměru. Počet stupňů volnosti f vyjadřuje počet nezávislých hodnot, z nichž je odhad počítán: průměr $\bar{x}$ není nezávislý na hodnotách x_i - byl z nich počítán - a každá hodnota x_i se na hodnotě průměru podílí jednou n-tinou své hodnoty, takže nezávislých hodnot je $f = n - \frac{1}{n}n = n - 1$. Obecně platí, že počet stupňů volnosti f = n - k, kde k je počet parametrů, určených z výsledků x_i, od nichž počítáme odchylku. Odhad podle (4.7a) je vhodný pro soubor paralelních měření, pokud jejich počet n není příliš malý. Pokud máme odhadnout směrodatnou odchylku jako metrologickou charakteristiku měřící metody, analytické metody apod., je vhodnější následující postup: Každý z k různých vzorků měříme n_j - krát, j = 1,2..., k a dostaneme matici výsledků [$|x_{ij}|$], kde i značí opakování a j vzorek. Provedeme „centrováni“ tak, že vypočteme průměry pro všechny vzorky a odchylky $\Delta_{ij} = (x_{ij} - \bar{x}_j)$. Odhad směrodatné odchylky je pak

$$s = \sqrt{\frac{1}{f}\sum_{j=1}^{k}\sum_{i=1}^{n}\Delta_{ij}^2} = \sqrt{\frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \qquad (4.7b)$$

kde N je počet všech výsledků

$N = \sum_{j=1}^{k} n_j$. Je-li počet měření, provedených na všech vzorcích stejný, je počet stupňů volnosti f = k(n - 1)

Poznámka (*): Kdybychom v (4.7) použili n místo f, dostali bychom odhad parametru σ, který je sice konsistentní a vydatný, ale nebyl by nestranný, protože by - hlavně pro malé n - podhodnocoval parametr σ.

Jiným, ovšem méně vydatným a málo robustním odhadem parametru σ je rozpětí

$$R = x_{max} - x_{min} \tag{4.8}$$

kde x_{max}, x_{min} je nejmenší, resp. největší hodnota celého náhodného výběru; u seřazeného souboru je $x_{min} = x_1$ a $x_{max} = x_n$. Pro normálně rozdělené výsledky můžeme použít odhad směrodatné odchylky, založený na rozdělení rozpětí, totiž

$$s_R = k_n \, . \, R \tag{4.9}$$

kde hodnoty koeficientu k_n pro různá n najdeme v tab. 5. [1], kde je také uvedena vydatnost odhadu podle (4.9) oproti odhadu podle (4.7a). Je zřejmé, že pro $n \geq 8$ je vydatnost tohoto odhadu výrazně nižší, a proto se užívá jen pro menší soubory dat. Poměr $q = R/s$, tzv. „studentizované rozpětí, velmi závisí na rozdělení výsledků a proto není odhad podle (4.9) robustní. Je zřejmé, že $k_n = q^{-1}$. Při odhadech, založených na rozpětí, je $f = n$, protože R se počítá ze dvou nezávislých hodnot, které představují různé výběry z téhož základního souboru.

Směrodatná odchylka je metrologickou charakteristikou přesnosti výsledků, tj. charakteristikou náhodné chyby. Charakteristikou relativní hodnoty náhodné chyby je relativní směrodatná odchylka, vyjádřená jako

$$s_r = \frac{s}{\bar{x}} \tag{4.10a}$$

resp. v procentech

$$s_r(\%) = 100 \, \frac{s}{\bar{x}} \tag{4.10b}$$

Pak se označuje jako variační koeficient.

Pro odhad parametrů polohy a rozptýlení normálního rozdělení bylo zavedeno více různých „statistik“, z nichž každá má poněkud jiné vlastnosti. Otázka, který odhad je „nejlepší“, je však zcela nesmyslná: každý odhad je „nejlepší“ pro ty předpoklady, za kterých byl odvozen.

Řešený příklad:

Určete odhad parametrů μ a σ těchto výsledků:

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
8,35	0,022	0,000484
8,21	- 0,118	0,013924
8,32	- 0,008	0,000064
8,39	0,062	0,003844
8,37	0,042	0,001764
41,64	0,000	0,001764

Medián: 8,21 < 8,32 < 8,35 < 8,37 < 8,39

Rozpětí: R = 8,39 - 8,21 = 0,18

Průměr: $\bar{x} = 41,64 / 5 = 8,328$

Směrodatná odchylka: $s = \sqrt{0,02008 / 4} = 0,07085$

$$s_R = 0,18 \,.\, 0,4299 = 0,07738$$

Relativní směrodatná odchylka $s_r(\%) = (100 \,.\, 0,07085) / 8,328 = 0,85\%$

Shoda průměru a mediánu je dobrá, $s_r < 1\%$ *ukazuje velmi dobrou přesnost.*

4.1.2 Intervalové odhady

Všechny až dosud uvažované odhady byly vždy bodové, tj. vyjádřeny pomocí jediného čísla. Jakožto náhodné veličiny mají bodové odhady určitá pravděpodobnostní rozděleni: tak např. průměr má pro malé n Studentovo t-rozdělení, které v limitě přechází v normální, rozptyl s^2 má rozdělení χ^2, které v limitě také přechází v normální a s_R má rozdělení

rozpětí. Těmito rozděleními se v souvislosti s bodovými odhady nemusíme zabývat, ale použijeme jich při určování intervalových odhadů.

Nevýhodou bodového odhadu parametru μ z normálního rozdělení např. výpočtem aritmetického průměru $\bar{x}$ je, že tento odhad nevyjadřuje přesnost s jakou byl určen. Proto je často výhodnější určit interval spolehlivosti, ve kterém leží odhadovaná hodnota μ s vysokou, předem zvolenou pravděpodobností $(1 - \alpha)$. Interval spolehlivosti vychází tím širší, čím větší je pravděpodobnost, že v něm leží správná hodnota μ. Pro normální rozdělení je taková pravděpodobnost dána jako

$$F(z) = \{\xi \leq z(\alpha)\} = (1 - \alpha),\ z = \frac{x-\mu}{\sigma}$$

kde $F(z)$ je distribuční funkce rozdělení $N(0,1)$, viz odst. 2.3.3 a $z(\alpha)$ je taková hodnota normované náhodné veličiny z, pro kterou má odchylka $(x - \mu)$ pravděpodobnost α. Hodnotu α označujeme jako hladinu významnosti, $(1 - \alpha)$ je koeficient spolehlivosti a $z(\alpha)$ označujeme jako kritickou hodnotu normálního rozdělení na hladině významnosti α.

Vlastní výpočet intervalu spolehlivosti závisí na tom, zda známe hodnotu parametru σ nebo pouze jeho odhad s, a zda jsou výsledky správné, kdy střední chyba $\delta = 0$, nebo nejsou zcela správné a střední chyba je nenulová, $\delta \neq 0$.

A. Známe parametr σ nebo jeho odhad s, určený z velkého souboru dat, např. podle (4.7b) pro $(N - k) \geq 30$, je interval spolehlivosti dán jako

$$L_{1,2} = \bar{x} \pm \sigma \frac{z(\alpha)}{\sqrt{n}} \qquad (4.11)$$

kde n je počet paralelních stanovení, z nichž byl určen průměr $\bar{x}$ a L_1 je dolní a L_2 horní mez intervalu spolehlivosti, $L_2 - L_1 = 2\sigma\, z(\alpha) / \sqrt{n}$ je jeho šíře. O tomto intervalu spolehlivosti platí, že

$$P\{\bar{x} - \sigma \frac{z(\alpha)}{\sqrt{n}} \leq \mu \leq \bar{x} + \sigma \frac{z(\alpha)}{\sqrt{n}}\} = (1 - \alpha)$$

B. Známe-li pouze odhad s, určený z menšího souboru, musíme postupovat jinak. Víme, že podíl $(n - 1)\ (\sigma^2 / s^2)$ má χ^2 - rozdělení o $f = (n - 1)$ stupních volnosti. Náhodná veličina ξ je nezávislá na χ^2 a náhodná veličina

$$t = \frac{\xi}{\sqrt{\frac{\chi^2}{n}}} \qquad (4.12)$$

má Studentovo t-rozdělení. Proto můžeme určit interval spolehlivosti průměru jako

$$L_{1,2} = \bar{x} \pm s \frac{t(\alpha, f)}{\sqrt{n}} \qquad (4.13)$$

kde n je počet měření, z nichž byl určen průměr $\bar{x}$ a počet stupňů volnosti $f = n_S - 1$, kde n_S je počet paralelních stanovení, z nichž byl určen odhad s; zřejmě bude vždy $n_S \geq n$. Kritické hodnoty $t(\alpha,f)$ jsou uvedeny v tab. 6. [1]. O intervalu spolehlivosti podle (4.13) platí, že

$$P\{\bar{x} - s \frac{t(\alpha, f)}{\sqrt{n}} \leq \mu \leq \bar{x} + s \frac{t(\alpha, f)}{\sqrt{n}}\} = (1 - \alpha)$$

Pro výpočty, prováděné za použití počítače, není výhodné ukládat celé tabulky kritických hodnot do paměti: výhodnější je použití aproximace pro výpočet $t(\alpha,f)$. Jednu z takových aproximací presentuje tab. 7. [1]. Velmi jednoduchá, ovšem jen přibližná aproximace kritických hodnot $t(0,05;f)$ pro $4 \leq f \leq 60$ je

$$t(0,05;f) = 2 + \frac{2,5}{f} \qquad (4.14)$$

Interval spolehlivosti průměru lze určit i pomocí rozpětí jako

$$L_{1,2} = \bar{x} \pm R \,.\, K_n \qquad (4.15)$$

kde koeficienty K_n pro různé n a $(1 - \alpha)$ jsou uvedeny v tab. 8. [1]. Tento způsob je výhodný pro svou jednoduchost, je však použitelný pouze pro průměr, určený z malého souboru dat.

C. Jsou-li výsledky zatíženy soustavnou chybou δ, jsou dvě možnosti, jak vyjádřit interval spolehlivosti:

1. Podle GRABEHO rozšíříme interval spolehlivosti směrem nahoru i dolu o hodnotu střední chyby δ, tj. počítáme:

$$L_{1,2} = \bar{x} \pm \left(s \frac{t(\alpha, f)}{\sqrt{n}} + \delta\right)$$

Tento interval je symetrický okolo průměru a jeho šíře je o 2δ větší než intervalu podle (4.13); tento způsob má však nevýhodu v tom, že přesně neznáme pravděpodobnost, s níž leží hodnota μ v tomto intervalu, víme jen, že je $P \geq (1 - \alpha)$ pro zvolenou hodnotu α. Je znázorněn na obr. 8b.

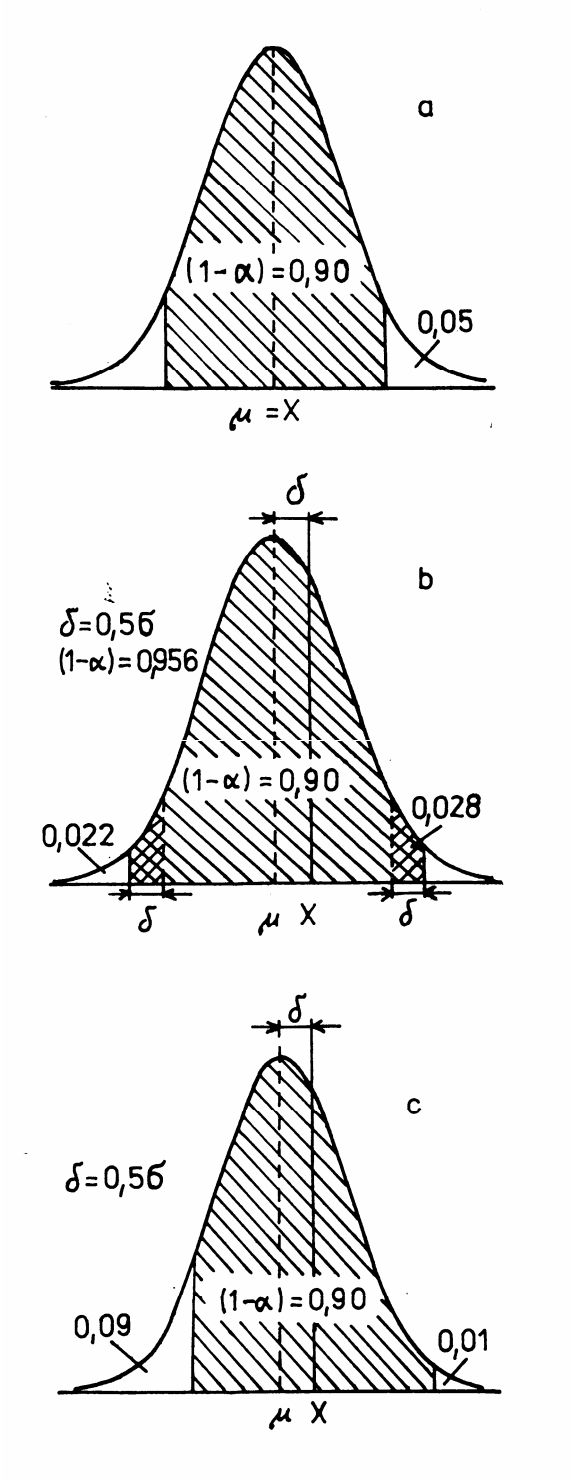
2. Určíme interval spolehlivosti za pomoci necentrálního t-rozdělení jako

$$L_{1,2} = \bar{x} \begin{cases} + s.t(\alpha,f : d) / \sqrt{n} \\ - s.t(\alpha,f : d) / \sqrt{n} - 2\delta \end{cases}$$

pro $\delta > 0$, resp. jako

$$L_{1,2} = \bar{x} \begin{cases} + s.t(\alpha,f : d) / \sqrt{n} + 2\delta \\ - s.t(\alpha,f : d) / \sqrt{n} \end{cases}$$

pro $\delta < 0$. Kritické hodnoty necentrálního t-rozdělení $t(\alpha,f:d)$ lze určit pomocí aproximace, uvedené i s příslušnými koeficienty v tab. 9. [1].



Obr. 8: Intervaly spolehlivosti průměru

Interval spolehlivosti průměru určený pomocí t-rozdělení podle (4.12), resp. rozšířený o 2δ podle GRABEHO návrhu jsou zobrazeny na obr. 8.a,b. Interval určený pomocí necentrálního t-rozdělení, který je nesymetricky rozložen okolo průměru, je uveden na obr. 8.c. V praxi se však vždy snažíme optimalizovat experimentální metodu tak, aby poskytovala pouze správné výsledky: pak přichází v úvahu pouze symetrický interval spolehlivosti podle (4.13) nebo (4.15) a určování nesymetrického intervalu snad pouze jako nouzové řešení.

Interval spolehlivosti rozptylu, o němž platí, že

$$P\{\frac{(n-1)s^2}{\chi^2_{n-1},\frac{\alpha}{2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{n-1},1-\frac{\alpha}{2}}\} = (1 - \alpha)$$

se častěji počítá jako jednostranný, tj. určuje se pouze jeho horní mez jako

$$\frac{(n-1)s^2}{\chi^2_{n-1},1-\frac{\alpha}{2}}$$

Dolní mez totiž nebývá v praktických případech zajímavá.

Řešený příklad:

Určete interval spolehlivosti průměru a horní mez intervalu spolehlivosti rozptylu těchto výsledků:

x_i	19,87	20,12	20,05	19,61	20,33	19,90

Průměr: $\bar{x} = 19,98$ Medián: $\tilde{x} = 19,975$ $n = n_S = 6$; $f = 5$

Směrodatná odchylka: s = 0,245927 *Rozpětí: R = 20,33 - 19,61 = 0,72*

Výpočet intervalu spolehlivosti ze směrodatné odchylky:

Kritická hodnota t(0,05;5) = 2,571 ..... převzato z tab. 6. [1]

Aproximace t(0,05;5) = ... podle tab. 7. [1]:

= 1,960 + 1/5(2,350 + 1/5(3,226 +1/5(0,621 + 1/5 . 4,549))) = 2,571

Přibližná aproximace: t(0,05;5) = 2 + 2,5/5 = 2,5

$L_{1,2} = 19,98 \pm 0,246 \frac{2,671}{\sqrt{6}} = 19,89 \pm 0,26$ *Dolní mez:* $L_1 = 19,72$

Horní mez: $L_2 = 20,24$

Výpočet intervalu spolehlivosti pomocí přibližné aproximace:

$L_{1,2} = 19,98 \pm 0,246 \frac{2,5}{\sqrt{6}} = 19,89 \pm 0,25$ *Dolní mez:* $L_1 = 19,73$

Horní mez: $L_2 = 20,23$

Výpočet intervalu spolehlivosti z rozpětí:

Pro n = 6 je $K_n = 0,399$ *... převzato z tab. 8. [1]*

$L_{1,2} = 19,98 \pm 0,72 \,.\, 0,399 = 19,89 \pm 0,29$ *Dolní mez:* $L_1 = 19,69$

Horní mez: $L_2 = 20,27$

Jednotlivé způsoby výpočtu vedou k poněkud odlišným hodnotám mezí $L_{1,2}$.

Výpočet horní meze intervalu spolehlivosti rozptylu:

Kritická hodnota $\chi^2_{5;0.95} = 1,15$ *... převzato z tab. 10. [1]*

Horní mez: $\sigma^2_{max} = \frac{5 \,.\, 0,2459^2}{1,15} = 0,263$ $\sigma_{max} = 0,513$

4.2 Statistické testování

Druhým významným úkolem statistické indukce je testování hypotéz o velikosti neznámého parametru pravděpodobnostního rozdělení nebo vztahu mezi parametry dvou a více výběrů. Postupuje se tak, že se nejprve vytvoří hypotéza a pak se hledají prostředky k jejímu ověření testem významnosti. Tak např. zjišťování správnosti výsledků měření nebo analýzy provádíme porovnáním výsledků se skutečnou hodnotou a usuzujeme, zda rozdíl mezi oběma hodnotami můžeme vysvětlit vlivem náhodných chyb a pokládat výsledky za správné nebo usoudíme, že rozdíl je způsoben soustavnou chybou a výsledky tedy správné nejsou. Vytvoříme si tzv. nulovou hypotézu H_0, že rozdíl mezi správnou a nalezenou hodnotou není statisticky významný a tuto hypotézu ověříme testováním. Přitom mohou nastat dvě možnosti: buď nulovou hypotézu přijmeme nebo ji zamítneme. V každém případě může být naše rozhodnutí buď správné nebo nesprávné. Zamítneme-li správnou hypotézu, mluvíme o chybě prvního druhu a přijmeme-li nesprávnou, mluvíme o chybě druhého druhu. Můžeme si to znázornit následujícím schématem:

H_0	Nulovou hypotézu	
	přijmeme	zamítneme
správná		chyba I. druhu
nesprávná	chyba II. druhu	

Testování se provádí porovnáním vypočtené hodnoty testovací charakteristiky (testovací „statistiky“) s kritickou hodnotou, kterou najdeme pro zvolenou hladinu významnosti v tabulkách nebo ji vypočítáme pomocí vhodné aproximace. Nulová hypotéza se zamítá, jestliže z experimentálních výsledků vypočtená hodnota „statistiky“ překročí její kritickou hodnotu, ale nemůžeme-li nulovou hypotézu zamítnout, znamená to jen, že nelze rozhodnout, zda je např. testovaný rozdíl dán jenom náhodnými chybami nebo zda je použitá metodika málo citlivá k tomu, aby rozdíl prokázala jako významný. Riziko chyby I. druhu je dáno hladinou významnosti $(1 - \alpha)$. Můžeme-li na základě logické úvahy určit, jaký rozdíl

testovaných parametrů je ještě přístupný, získáme otestováním specifikované alternativní hypotézy H_1 závěry o souhlasu testovaných veličin: zamítnutí H_1 pak znamená přijetí nulové hypotézy H_0. Vždy ovšem nebývá možné vytvořit specifikovanou alternativní hypotézu; nulovou lze vytvořit vždy.

Je popsáno mnoho statistických testů. Zde si ukážeme použití Studentova t-testu pro ověření správnosti a shodnosti výsledků měření a testování specifikované alternativní hypotézy o shodnosti, použití GRUBBSOVA a DEANOVA-DIXONOVA testu k určení odlehlých výsledků a testování shodnosti dvou rozptylů pomocí F-testu.

Testování rozdílu dvou směrodatných odchylek: Jsou dány dva odhady $s_1 \neq s_2$, které mají počty stupňů volnosti f_1 a f_2. Jako nulovou hypotézu testujeme předpoklad, že s_1 i s_2 jsou odhady téhož parametru σ. Můžeme si napsat schéma:

$$\left.\begin{array}{l} s_1 = \hat{\sigma}_1 \\ s_2 = \hat{\sigma}_2 \end{array}\right\rangle \; H_0: \sigma_1 = \sigma_2$$

Testovací kriterium (testovací statistika) je dána poměrem

$$F = \left(\frac{s_1}{s_2}\right)^2 \qquad (F \geq 1) \qquad (4.17)$$

Nulovou hypotézu zamítáme, když $F \geq F(\alpha, f_1, f_2)$. Tabulky kritických hodnot F - rozdělení jsou velmi obsáhlé; v tab. 11. [1] je uveden pouze jejich úsek pro $(1 - \alpha) = 0{,}95$.

Testování rozdílu průměrů od skutečné hodnoty: Jde o testování správnosti výsledku $\bar{x}$ porovnáním se skutečnou hodnotou μ (např. „ověřovací“ analýza standardního vzorku). Uvažujeme tedy následující nulovou hypotézu:

$$H_0: \qquad \bar{x} = \hat{\mu}$$

Testovací statistika je dána jako

$$t = \frac{|\mu - \bar{x}|}{s}\sqrt{n} \qquad (4.18)$$

Jestliže $t \geq t(\alpha,f)$, $f = n - 1$, zamítáme nulovou hypotézu.

Řešený příklad:

Titračním stanovením Mn ve standardním vzorku s obsahem 0,670% Mn bylo nalezeno: 0,69; 0,68; 0,70; 0,67; 0,67; 0,69; 0,66; 0,68; 0,67; 0,68; 0,68; 0,67 a 0,69% Mn. Je výsledek zatížen soustavnou chybou? Průměr výsledků $\bar{x} = 0,679\%$ Mn se zdánlivě dobře shoduje s $\mu = 0,670\%$ Mn. Provedeme však otestování t-testem. Směrodatná odchylka s = 0,0111516.

$$t = \frac{0,679 - 0,670}{0,0111516}\sqrt{13} = 2,910 > 2,179 = t(0,05;12)$$

Rozdíl je statisticky významný na hladině $(1 - \alpha) = 0,95$, zamítáme nulovou hypotézu a musíme předpokládat, že výsledky jsou zatíženy soustavnou chybou. Takovým konstatováním ovšem nesmíme zakončit ověřování titrační metody stanovení Mn a musíme se snažit soustavnou chybu odstranit. Skutečně bylo zjištěno, že použité chemikálie způsobovaly určitou malou hodnotu slepého pokusu; po jeho odečítání byly již výsledky správné.

Testování rozdílu dvou průměrů: Jde o testování shodnosti výsledků měření, provedených dvěma různými metodami na témže vzorku a zjišťujeme, zda obě metody dávají shodné výsledky, nebo jde o testování, kdy provádíme měření jednou metodou na dvou vzorcích a zjišťujeme, zda jde o dva vzorky téhož materiálu. Jako nulová hypotéza se testuje

$$\bar{x}_1 = \hat{\mu}_1$$

$$H_0: \mu_1 = \mu_2$$

$$\bar{x}_2 = \hat{\mu}_2$$

Experimentální data, která testujeme, mohou být dvojího druhu:

A. Jsou dány dva průměry $\bar{x}_1 \neq \bar{x}_2$, určené z n_1 a n_2 paralelních měření. Testovací charakteristika je

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{s_{1,2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$s_{1,2}^2 = \frac{\sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n_2} (x_{i2} - \bar{x}_2)^2}{n_1 + n_2 - 2} \qquad (4.19a)$$

a porovná se s kritickou hodnotou t-rozdělení pro $f = n_1 + n_2 - 2$. Nulovou hypotézu zamítáme, je-li $t \geq t(\alpha,f)$; nemůžeme ji zamítnout, je-li $t < t(\alpha,f)$, což ovšem neznamená shodnost obou průměrů. Má-li se prokázat souhlas mezi dvěma průměry, je třeba určit logicky zdůvodnitelný rozdíl mezi nimi, tj. $d = |\mu_1 - \mu_2|$, a jako specifikovanou alternativní hypotézu otestovat tento rozdíl pomocí testovacího kriteria

$$u = \frac{d}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \qquad (4.19b)$$

a specifikovaná alternativní hypotéza se zamítá, je-li $u \geq u(\alpha,\beta,f)$; kritické hodnoty jsou uvedeny v tab. 12. [1]. Kombinací otestování podle (4.19a) a (4.19b) lze s příslušným rizikem rozhodnout jak o přijetí, tak o zamítnutí hypotéz o shodnosti v rámci rozdílu d, tak o neshodnosti obou průměrů.

B. Je dáno $k \geq 2$ vzorků, z nichž každý je měřen (analyzován) pomocí dvou různých metod A, B; $n_A = n_B = 1$. Shodnost výsledků obou metod můžeme otestovat pomocí diferencí $x_{Ai} - x_{Bi}$, které vypočteme pomocí schématu:

			Výsledky	
Vzorek:	x_A	x_B	D_i	$(D_i - \overline{D})^2$
1	x_{A1}	x_{B1}	$x_{A1} - x_{B1}$	$(D_1 - \overline{D})^2$
2	x_{A2}	x_{B2}	$x_{A2} - x_{B2}$	$(D_2 - \overline{D})^2$
:				
k	x_{Ak}	x_{Bk}	$x_{Ak} - x_{Bk}$	$(D_k - \overline{D})^2$

$$\overline{D} = \frac{1}{k}\sum_{i=1}^{k} D_i$$

Určíme průměry: $\overline{x}_A = \frac{1}{k}\sum_{i=1}^{k} x_{Ai}$, $\overline{x}_B = \frac{1}{k}\sum_{i=1}^{k} x_{Bi}$ a vypočteme hodnotu testovací charakteristiky

$$t = \left|\overline{x}_A - \overline{x}_B\right| \sqrt{\frac{k(k-1)}{\sum (D_i - \overline{D})^2}} \qquad (4.20)$$

a porovnáváme s kritickou hodnotou t-rozdělení pro f = k - 1 a pro zvolenou hladinu významnosti; nejčastěji se používá (1 - α) = 0,95.

Dosud jsme se zabývali testováním parametrů a jejich odhadů; můžeme však testovat i jednotlivé výsledky, např. abychom zjistili, zda nejsou odlehlé od ostatních výsledků proto, že jsou zatíženy hrubou chybou. Testování odlehlých výsledků má v chemometrii dosti velký význam a proto si zde všimneme dvou testů odlehlosti. Pro testování odlehlosti je nulová hypotéza dána předpokladem, že testovaný výsledek je náhodným výběrem ze základního souboru. Při použití GRUBBSOVA testu testujeme rozdíl podezřelého výsledku od průměru

$$T = \frac{\left|x_i - \overline{x}\right|}{s};\ i = 1,n \qquad (4.21)$$

a porovnáváme s hodnotou $T(\alpha,n)\sqrt{\frac{n-1}{n}}$, kde $T(\alpha,n)$ jsou kritické hodnoty GRUBBSOVA rozdělení a s je odhad směrodatné odchylky ze všech výsledků, včetně podezřelého. Kritické hodnoty $T(\alpha,n)$ jsou uvedeny v tab. 13. [1]. Pro malé soubory dat je jednodušší použití testu podle DEANA a DIXONA, kdy za použití seřazených výsledků určíme

$$Q = \frac{|x_i - \bar{x}_{i-1}|}{R};\ i = 2,n \qquad (4.22)$$

kde $R = x_n - x_1$ je rozpětí. Porovnáváme s kritickými hodnotami $Q(\alpha,n)$, přičemž pro $n = 3$ musí být všechny výsledky vzájemně různé. Kritické hodnoty DEANOVA a DIXONOVA rozdělení jsou uvedeny v tab. 14. [1].

Řešený příklad:

V organické látce byl nalezen obsah uhlíku: 16,84%, 16,86%, 16,91%, 16,93% a 17,61% C. Může jít o látku, která obsahuje podle teorie 17,41% C?

Průměr $\bar{x}$ = 17,03% C, medián $\tilde{x}$ = 16,91% C; směrodatná odchylka: s = 0,32627. Interval spolehlivosti $L_{1,2} = 17,03 \pm 0,32627\ \frac{2,776}{\sqrt{5}} = 17,03 \pm 0,41$.

Hodnota 17,41 leží v intervalu spolehlivosti; mohlo by se tedy jednat o tuto látku. Nejvyšší výsledek 17,61% C je však podezřelý jako odlehlý: otestujeme testem podle DEANA a DIXONA:

$$Q = \frac{17,61 - 16,93}{17,61 - 16,84} = 0,883 > 0,642 = Q(0,05;5)$$

Nejvyšší výsledek je odlehlý, vyloučíme jej a dostaneme:

Průměr a medián $\bar{x} = \tilde{x}$ = 16,885% C, směrodatná odchylka: s = 0,04203. Interval spolehlivosti

$L_{1,2} = 16,885 \pm 0,04203\ \frac{3,182}{\sqrt{4}} = 16,885 \pm 0,067$; *hodnota 17,41% C leží mimo*

interval spolehlivosti; pokud jsou výsledky analýzy správné, nejde o tuto látku. Odlehlý výsledek, pokud by nebyl vyloučen, mohl způsobit vznik nesprávného výsledku a nesprávného rozhodnutí, provedeného na základě tohoto výsledku.

Odlehlý výsledek se někdy projeví tím, že rozdíl mezi průměrem a mediánem je větší než obvykle. Vyloučením odlehlého výsledku:

1. změní se průměr a zlepší souhlas průměru s mediánem;

2. zmenší se rozpětí a směrodatná odchylka;

3. obyčejně se zúží interval spolehlivosti přesto, že je třeba použít hodnotu $t(\alpha,f)$ pro menší f nebo hodnotu K_n pro menší n než před vyloučením.

Analogický efekt jako vyloučení odlehlé hodnoty má použití robustního odhadu hodnoty parametru μ. Ukážeme si to na příkladu:

Řešený příklad:

Z výsledků z předešlého příkladu vypočteme jako odhad střední hodnoty winsorizovaný průměr, přičemž provedeme jednostrannou winsorizaci výsledku, více odlehlého od mediánu.

Rozdíl nejnižšího výsledku od mediánu je 16,91 - 16,84 = 0,07%, rozdíl nejvyššího výsledku od mediánu je 17,61 - 16,91 = 0,70%. Vypouštíme tedy největší výsledek, nahrazujeme jej nejbližším sousedním a dostaneme soubor: 16,84 < 16,86 < 16,91 < 16,93 = 16,93%. Průměr $\bar{x}$ *= 16,894 se dobře shoduje s mediánem* $\tilde{x}$ *= 16,910%, směrodatná odchylka s = 0,04159, takže* $L_{1,2} = 16,89 \pm 0,052$ *;* $L_1 = 16,84\%$, $L_2 = 16,95\%$. *Nemůže jít o látku s obsahem 17,41% C. Všimněme si, že interval spolehlivosti winsorizovaného souboru je ještě užší než po vyloučení odlehlého výsledku v předchozím příkladu.*

Testování na předem zvolené hladině významnosti (1 - α) vede k jednoznačnému rozhodnutí, což je sice výhodné, ale představuje vytvoření diskontinuity ve spojité řadě hodnot kriterií např. pro správnost, shodnost, odlehlost atd. Proto někdy rozhodujeme takto:

$0,99 < (1 - \alpha)$	vysoce významné
$0,95 < (1 - \alpha) \leq 0,99$	významné
$0,90 < (1 - \alpha) \leq 0,95$	málo významné
$\leq 0,90$	nevýznamné

Pro vědecké účely je vhodné neporovnávat vypočtenou hodnotu testovací statistiky s kritickou hodnotou, ale určit, na jaké hladině významnosti, tj. s jakou pravděpodobností není nulová hypotéza splněna. To se dá provést dvojím způsobem:

a) Exaktně: zavést takovou aproximaci kritických hodnot testovacích statistik, která umožňuje výpočet příslušné hladiny významnosti pro jakoukoliv hodnotu testovací statistiky při daném počtu stupňů volnosti; tento způsob se ovšem dá realizovat pouze a použití počítače.

b) Přibližně: grafickou nebo početní interpolací mezi tabelovanými hodnotami; vyžaduje to však dosti obsáhlé tabulky kritických hodnot testovací statistiky. Dnes se prakticky nepoužívá.

Pak ovšem nevyjadřujeme výsledek testování výrokem, že nulová hypotéza je nebo není statisticky významná, ale že testovaná hypotéza může být přijata na hladině významnosti $(1 - \alpha)$.

Řešený příklad:

Látku s obsahem 7,10% N jsme analyzovali jednak spalovací metodou (A), jednak metodou KJELDAHLOVOU (B); máme určit, na jaké hladině $(1-\alpha)$ jsou výsledky jednotlivých metod, $\bar{x}_A$ a $\bar{x}_B$, shodné se skutečným obsahem.

x_A	x_B
7,10	7,04

$$t_A = \frac{7,11 - 7,10}{0,012}\sqrt{5} = 1,825$$

7,10		7,06
7,11		7,06
7,11		7,08
7,13		7,11
7,11	$\bar{x}$	*7,07*
0,0122	*s*	*0,0265*

$$t_B = \frac{7,10 - 7,07}{0,026}\sqrt{5} = 2,532$$

Tab. t(α;4):

(1 - α)	*t(α;4)*	
0,80	*1,533*	
0,85	*1,824*	
0,90	*2,132*	t_A=1,825
0,95	*2,776*	t_B=2,532

$t_A = 1,825 \approx 1,824 = t(0,85;4)$

$t(0,90;4) = 2,131 < 2,532 < 2,776 < t(0,95;4)$

Metoda A poskytuje výsledky, shodné se skutečnou hodnotou na hladině cca (1 - α) = 0,85.

Pro metodu B musíme provést výpočet lineární interpolací:

$$(1 - \alpha) = 0,90 \frac{0,95 - 0,90}{2,776 - 2,132} (2,532 - 2,132) = 0,93$$

4.3 Závislost dvou proměnných

Závislost mezi dvěma proměnnými může být v podstatě dvojího druhu:

1. Závislost funkční: Určité hodnotě argumentu x_i odpovídá jediná, určitá hodnota nezávisle proměnné y_i, tj. platí $y = f(x)$.

2. Závislost stochastická: Je to závislost, kdy buď závisle proměnná, event. obě proměnné jsou náhodné veličiny. V prvém případě, tj. je-li náhodná jen závisle proměnná, odpovídá určité hodnotě argumentu x_i celé pravděpodobnostní rozdělení hodnot náhodné veličiny y, která má očekávanou hodnotu μ_y a rozptyl σ_y^2, tj. platí, že $\mu_y = f(x_i)$. V druhém případě, tj. při vzájemné závislosti dvou náhodných veličin y a x s rozptyly σ_y^2 a σ_x^2 očekávanými hodnotami μ_y, μ_x, platí, že $\mu_y = f(\mu_x)$. Dvě náhodné proměnné však mohou být vzájemně závislé jen do jisté míry.

Při zpracování experimentálních dat můžeme snad jen vyjímečně předpokládat funkční závislost: zpravidla půjde o závislost stochastickou. Stochastická, přímá nebo nepřímá závislost, je řešitelná matematicko-statistickými metodami, které umožňují získat řadu užitečných informací o zpracovávaném souboru. Závislost náhodné veličiny y na pevných hodnotách argumentu x zpravidla zpracujeme formou regrese, kdy určujeme tvar závislosti; vzájemnou závislost dvou náhodných veličin y, x lze sice také zpracovat, ale toto zpracování vede v podstatě jen ke zjištěním jak silná je korelace mezi oběma proměnnými. Proto často i tehdy, jde-li o závislost dvou experimentálně zjišťovaných a tedy náhodných hodnot veličin y, x, zpracováváme soubor regresními metodami. Přitom buď veličinu, kterou zjišťujeme přesněji, pokládáme za argument, nebo provedeme „dvojí“ regresní zpracování, tj. jednou závislosti $x = f_1(y)$ a podruhé závislosti $y = f_2(x)$; např. kalibrační a analytická závislost v analytické chemii. Jinou možností zpracování závislosti dvou náhodných veličin je použití tzv. ortogonální regrese. Zatímco „obyčejná“ regrese má své oprávnění v případě, kdy $\sigma_y >> \sigma_x$, je ortogonální regrese vhodná zejména pro případ $\sigma_y \approx \sigma_x$. V případě, kdy se

rozptyly řádově neliší ani nejsou přibližně stejné, může přinést některé cenné informace o vzájemné závislosti, zejména o jeho tvaru, dvojí regrese.

Pro způsob zpracování regresivní závislosti je podstatné, zda známe nebo z teorie můžeme předpokládat tvar závislosti: pak jde v podstatě jen o hledání parametrů této závislosti, nebo tento tvar neznáme a hledáme vhodný model pro vyjádření tohoto tvaru. Velmi častým modelem je případ, kdy se jedná o lineární závislost na neznámých parametrech: tyto modely označujeme jako lineární vzhledem k parametrům. Velmi častá je lineární závislost dvou proměnných. Pak mluvíme o jednoduché lineární regresi, která je vyjádřená jako

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad (4.23)$$

kde ε_i je náhodná chyba, která má normální rozdělení $N(0,\sigma^2)$ a α, β jsou parametry lineární regresní rovnice. Odhady těchto parametrů, tj. koeficienty $a = \hat{\alpha}$, $b = \hat{\beta}$ určujeme takto: máme $n > 2$ bodů, tj. dvojic hodnot $[x_i, y_i]$, $i = 1, \ldots n$. To znamená, že pro určení dvou koeficientů a, b máme $n > 2$ rovnic $y_i = a + b.xi + \varepsilon_i$. Jde tedy o přeurčený systém rovnic, tj. máme více rovnic než neznámých. Takový systém se dá řešit jen pro určitou podmínku. Pro odhad koeficientů regresní rovnice zvolíme podmínku nejmenšího součtu čtverců, tj. podmínku

$$\sum_i (y_i - a - bx_i)^2 = \min, \qquad i = 1, \ldots n$$

K určení této podmínky derivujeme podle obou koeficientů a položíme rovno nule:

$$\frac{\partial}{\partial a}\sum_i (y_i - a - bx_i)^2 = -2 \sum_i (y_i - a - bx_i) = 0$$

$$\frac{\partial}{\partial b}\sum_i (y_i - a - bx_i)^2 = -2 \sum_i (y_i - a - bx_i) = 0$$

Úpravou těchto vztahů dostaneme tzv. normální rovnice

$$an + b\sum_i x_i - \sum_i y_i = 0$$

$$a\sum_i x_i + b\sum_i x_i^2 - \sum_i x_i y_i = 0$$

Jejich řešením pak dostaneme odhady koeficientů

$$b = \frac{(\sum_i x_i)(\sum_i y_i) - n\sum_i x_i y_i}{(\sum_i x_i)^2 - n\sum_i x_i^2} \qquad (4.24)$$

$$a = \frac{1}{n}(\sum_i y_i - b\sum_i x_i) \qquad (4.25)$$

Směrodatnou odchylku, která charakterizuje rozptýlení kolem regresní přímky, určíme jako

$$s_{y,x} = \sqrt{\frac{\sum_i (y_i - Y_i)^2}{n-2}} \qquad (4.26)$$

kde regresní hodnota $Y_i = a + bx_i$. Odhad směrodatné odchylky koeficientu $a = \hat{\alpha}$ počítáme jako

$$s_a = s_{y,x}\sqrt{\frac{1}{n} + \frac{\bar{x}}{\sum_i x_i^2 - \frac{1}{n}(\sum_i x_i)^2}} \qquad (4.27)$$

a odhad směrodatné odchylky koeficientu $b = \hat{\beta}$ počítáme podle vztahu

$$s_b = s_{y,x}\sqrt{\frac{s_{y,x}^2}{\sum_i x_i^2 - \frac{1}{n}(\sum_i x_i)^2}} \qquad (4.28)$$

kde $s_{y,x}$ počítáme podle (4.26).

Známe-li odhady koeficientů a, b a jejich směrodatné odchylky, můžeme určovat jejich intervaly spolehlivosti

$$L_{(a)1,2} = a \pm s_a t(\alpha,f) \qquad L_{(b)1,2} = b \pm s_b t(\alpha,f)$$

pro f = n - 2. Koeficienty a, b můžeme též pomocí t-testu testovat oproti předpokládaným hodnotám α, β. Nejčastější je testování hypotézy H_0: α = 0; testovací statistika je pak

$$t = \frac{|a|}{s_a}, \qquad f = n - 2$$

Tímto testováním zjišťujeme, zda regresní přímka prochází počátkem, tj. bodem x = 0, y = 0. Toto zjištění je často velmi důležité, např. při zjišťování hodnoty „slepého pokusu“ při kalibraci v analytické chemii. Obdobně testujeme hypotézu H_0: β = c, c = konst; v některých důležitých chemometrických aplikacích bývá c = 1.

Zjistíme-li otestováním, že rozdíl koeficientu α není od nuly statisticky významný, resp. že nula leží uvnitř intervalu $L_{(a)1,2}$, můžeme předpokládat regresní závislost

$$y = b \,.\, x \tag{4.29}$$

a koeficient b počítáme z jediné normální rovnice

$$b \sum_i x_i^2 - \sum_i x_i y_i = 0$$

jako

$$b = \frac{\sum_i x_i y_i}{\sum_i x_i^2} \tag{4.30}$$

Směrodatná odchylka

$$s_{y,x} = \sqrt{\frac{\sum_i (y_i - Y_i)^2}{n-1}} \qquad (4.31)$$

kde regresní hodnota $Y_i = b.x_i$.

Řešený příklad:

Máme zjistit koeficienty lineární regresní přímky pro následující data: Sestavíme tabulku

i	x_i	y_i	x_i^2	$x_i\, y_i$	Y_i	$(y_i - Y_i)^2$
1	0,00	0,02	0,00	0,000	0,01	0,0001
2	0,80	0,94	0,64	0,752	0,98	0,0016
3	1,60	2,01	2,56	3,216	1,96	0,0025
4	2,40	2,90	5,76	6,960	2,93	0,0009
5	3,20	3,91	10,24	12,512	3,91	0,0000
6	4,00	4,88	16,00	19,520	4,88	0,0000
	12,00	14,66	35,20	42,960		0,0051

$$b = \frac{12 \cdot 14{,}66 - 6 \cdot 42{,}96}{12^2 - 6 \cdot 35{,}2} = 1{,}21786 \quad a = \frac{1}{6}\,(14{,}66 - 1{,}21786 \,.\, 12{,}0) = 0{,}0076$$

$$s_{y,x} = \sqrt{\frac{0{,}0051}{4}} = 0{,}0357$$

Hodnota koeficientu a = 0,0076 je malá; proto otestujeme, zda je statisticky významně odlišná od $\alpha = 0$. Napřed vypočteme

$$s_a = 0{,}0357 \sqrt{\frac{1}{6} + \frac{2{,}2}{35{,}2 - 12{,}12 / 6}} = 0{,}0258$$

$$\text{Testování: } t = \frac{0{,}0076}{0{,}0258} = 0{,}3 < 2{,}776 = t(0{,}05;4)$$

Koeficient a je statisticky nevýznamně odlišný od nuly tj. můžeme závislost vyjádřit regresní rovnicí y = b . x. Sestavíme novou tabulku:

i	x_i	y_i	Y_i	$(y_i - Y_i)^2$
1	0,00	0,02	0,00	0,0004
2	0,80	0,94	0,98	0,0016
3	1,60	2,01	1,95	0,0036
4	2,40	2,90	2,93	0,0009
5	3,20	3,91	3,91	0,0000
6	4,00	4,88	4,89	0,0001

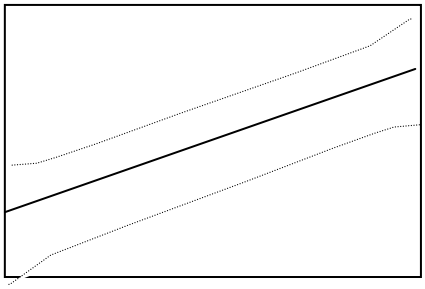
$$b = \frac{42{,}96}{35{,}20} = 1{,}22045 \qquad s_{y,x} = \sqrt{\frac{1}{5} 0{,}0066} = 0{,}0363$$

Nyní bychom měli otestovat F-testem, zda rozdíl mezi oběma hodnotami směrodatných odchylek $s_{y,x}$ pro model y = a + b . x a pro model y = b . x není statisticky významný: jejich shoda, tj. 0,0357 a 0,0363 je však tak dobrá, že testování není nutné.

Pro regresní hodnotu, tj. pro hodnotu Y vypočtenou z regresní rovnice můžeme určit interval spolehlivosti jako

$$L_{1,2} = (a + bx) \pm s_{y,x} \cdot t(\alpha,f) \sqrt{\frac{1}{n} + \frac{(X-\bar{x})^2}{\sum_i x_i^2 - \frac{1}{n}(\sum_i x_i)^2}} \qquad (4.32)$$

Výpočtem intervalu spolehlivosti pro libovolnou hodnotu X ze zpracované oblasti hodnot x dostaneme tzv. pás spolehlivosti pro regresní přímku. Má tvar, zobrazený na obr. 8.: je nejužší ve své střední části. Body, odlehlé od regresní přímky, můžeme vyloučit po otestování GRUBBSOVÝM T-testem.



Obr. 9: Pás spolehlivosti regresní přímky

Lineární závislost je v chemometrii velmi důležitá, protože mnohé chemické, fyzikálně-chemické a fyzikální zákonitosti jsou v podstatě lineární: k odchylkám od lineárního průběhu dochází často tehdy, kdy se buď uplatňují nové vlivy, s nimiž teorie nepočítá nebo tehdy, kdy se neuplatní některé teorií předpokládané.

Postup zpracování nelineární regresní závislosti bude záviset především na tom, zda známe racionální vztah pro závislost sledovaných proměnných, příp. zda jej můžeme odvodit na základě nějaké teorie. Není-li tomu tak, pokoušíme se vystihnout závislost mezi proměnnými empirickou funkcí, která ve sledovaném intervalu vhodně aproximuje sledovanou závislost. Musíme tedy probírat jeden typ empirické závislosti za druhým, až najdeme nejvýhodnější; často se osvědčuje vyjádření polynomem. Hledání vhodné závislosti může usnadnit schéma 1., kde jsou uvedeny průběhy těchto závislostí. Jindy lze vhodnou transformací převést nelineární závislost na lineární a řešit pak metodami lineární regrese.

Tento způsob, je-li možný, bývá snadnější a je velmi často používán zejména ve fyzikální chemii. Některé linearizační transformace jsou uvedeny ve schématu 1.

Použití výpočetní techniky bylo velkým přínosem právě pro zpracování regresních závislostí a je možno zcela oprávněně tvrdit, že metody nelineární regrese při zpracování rozsáhlých souborů dat by bez počítačů nebyly vůbec prakticky realizovatelné.

Příkladem regrese, která je lineární v parametrech, ale popisuje nelineární závislost mezi proměnnými, je tzv. polynomická regrese, při níž vyjadřujeme vztah mezi oběma proměnnými polynomem n-tého stupně

$$y = k_0 + k_1 x + k_2 x^2 + \ldots + k_n x^n = \sum_{i=0}^{n} k_i x^i \qquad (4.33)$$

kde celočíselná hodnota n vyjadřuje stupeň polynomu a k_0, k_1, ... , k_n jsou neznámé koeficienty. Tyto koeficienty určíme řešením normálních rovnic

$$k_0 n + k_1 \sum_i x_i + \ldots + k_n \sum_i x_i^n - \sum_i x_i y_i = 0$$

$$k_0 \sum_i x_i + k_1 \sum_i x_i^2 + \ldots + k_n \sum_i x_i^{n+1} - \sum_i x_i^2 y_i = 0$$

:

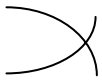
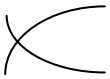
$$k_0 \sum_i x_i^n + k_1 \sum_i x_i^{n+1} + \ldots + k_n \sum_i x_i^{2n} - \sum_i x_i^n y_i = 0$$

Specielním případem je přímková regrese, tj. polynom prvního stupně.

Schéma 1: Některé nelineární závislosti a jejich linearizace

Vztah:	Tvar závislosti:	Linearizační transformace
		$Y = A + B \,.\, X$

$y = \sum_{i=0}^{n} k_i \cdot x^i$	Velmi flexibilní polynomická závislost
$y = \sum_{i=0}^{n} k_i \left(\frac{1}{x}\right)^i$	Velmi flexibilní polynomická závislost

$y = a \,.\, b^x$		$b > 0$	$Y = \ln y$	$X = x$
		$b < 0$	$A = \ln a$	$B = \ln b$
$y = a \,.\, e^x$		$b > 0$	$Y = \ln y$	$X = x$
		$b < 0$	$A = \ln a$	$B = b$
$y = a + b/x$		$b > 0$	$Y = y$	$X = 1/x$
		$b < 0$	$A = a$	$B = b$
$y = a + b \,.\, \log x$		$b > 0$	$Y = y$	$X = \log x$
		$b < 0$	$A = a$	$B = b$
$\log y = a + b \,.\, x$		$b > 0$	$Y = \log y$	$X = x$
		$b < 0$	$A = a$	$B = b$

Otevřená zůstává otázka volby vhodného stupně polynomu: Přitom musíme rozlišit, o který ze tří případů se jedná:

1. Regrese polynomem známého stupně;

2. Určení stupně polynomu;

3. Redukce členů polynomu.

Jestliže na základě zkušenosti nebo požadované přesnosti známe stupeň polynomu n a pro zpracování máme k dispozici m > n dvojic hodnot, řešíme soustavu normálních rovnic, což ovšem vyžaduje počítač a vhodný program. Jindy musíme napřed určit stupeň polynomu. Volba stupně polynomu je velmi závažná, protože zvolíme-li nízký stupeň, rovnice nemusí dosti dobře vystihnout skutečnou závislost a zvolíme-li příliš vysoký stupeň, „proloží se" křivka popsaná regresní rovnicí všemi body a nedosáhneme vyrovnání chyb a měření. Proto pro účely „vyrovnání" používáme spíše nižší stupeň polynomu a tam, kde jde o aproximaci např. kritických hodnot, apod., volíme spíše vyšší stupeň. Velmi „nebezpečné" jsou odlehlé výsledky, které mohou vést k nevhodné volbě stupně polynomu. Proto postupujeme tak, že provedeme výpočet pro nízký stupeň a pomocí GRUBBSOVA testu zjišťujeme odlehlost těch bodů, které se více liší od regresní hodnoty. Někdy provádíme redukci koeficientů polynomu, tj. vynechání těch koeficientů, které při otestování t-testem proti nule shledáme statisticky nevýznamné.

Správné provedení nelineární regrese, ať už za použití polynomu nebo jiného vztahu, případně po linearizaci, je záležitostí nejen matematickou, ale především vhodné logické úvahy a analýzy celého problému. Zde platí více než kde jinde, že chemometrie je chemický obor a tak volba vhodného modelu (tj. vztahu, kterým daný jev, fenomén, popisujeme) by měla být spíše otázkou racionálního nalezení takového vztahu než problému co nejlepší přiléhavosti regresní rovnice k experimentálním datům. Zde je do jisté míry třeba respektovat skutečnost, na kterou upozornil C. F. VON WEIZSACKER, že totiž přírodní zákony jsou v podstatě vyjádřitelné jednoduchými matematickými vztahy, ale komplikují je četné náhodné vlivy. Cílem regrese je právě vystižení určité „přírodní zákonitosti" při současné eliminaci vlivu náhodných okolností.

Je-li mezi dvěma náhodnými experimentálně určovanými proměnnými nějaká závislost, musíme především určit, jak je těsná. K tomu slouží tzv. korelační koeficient ρ; ten charakterizuje těsnost závislosti dvou náhodných veličin pro případ, že jde o lineární závislost, přičemž nabývá hodnot od -1 do +1. Kladných nabývá pro přímou a záporných pro nepřímou závislost, při nezávislosti obou proměnných je $\rho = 0$, při funkční závislosti je $\rho = 1$

a čím více se jeho hodnota blíží jedné, tím je závislost obou proměnných těsnější. Odhad korelačního koeficientu

$$\hat{\rho} = r = \frac{n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{\sqrt{[n\sum_i x_i^2 - (\sum_i x_i)^2][n\sum_i y_i^2 - (\sum_i y_i)^2]}} = \frac{cov_{xy}}{s_x \cdot s_y} \qquad (4.34)$$

kde kovariance

$$cov_{xy} = \frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Je zřejmé, že pro x = y přechází kovariance v rozptyl.

K testování nulové hypotézy H_0: $r = \hat{\rho}$, $\rho = 0$ používáme kritických hodnot $r(\alpha,n)$, uvedených v tabulce 15. Je-li $|r| \geq r(\alpha,n)$, hypotézu H_0 zamítneme. Při určování korelačního koeficientu má velký vliv rozsah souboru, ze kterého vycházíme: odhad r je totiž málo vydatným odhadem ρ. Tvrzení, že nemá praktický smysl určovat koeficient korelace pro malý počet dvojic dat ($n \leq 10$) se dá ověřit pomocí tzv. z - transformace, tj. veličiny

$$z = 1/2 \ln \frac{1+r}{1-r}$$

která má normální rozdělení se směrodatnou odchylkou $\sigma = (n - 3)^{-2} = \frac{1}{\sqrt{n-3}}$. Ukážeme to na příkladu.

Příklad:

Pro r = 0, určený z n = 7 je šíře intervalu spolehlivosti, určená pomocí z - transformace pro $(1 - \alpha) = 0{,}95$ taková, že $-0{,}96 \leq \rho \leq +0{,}96$, tj. nelze rozhodnout, jde-li o poměrně silnou nepřímou, žádnou nebo dosti silnou přímou závislost. Určíme-li však r = 0 z n = 70, je pro $(1 - \alpha) = 0{,}95$ interval spolehlivost

$-0{,}23 \leq \rho \leq +0{,}23$, takže podle tab. 15. [1] nelze zamítnout nulovou hypotézu; budeme bez nejmenších rozpaků mluvit o naprosté nezávislosti obou veličin.

Korelační koeficient nám poskytuje pouze informaci o tom, jak těsná je vzájemná závislost obou proměnných a o tom, jde-li o přímou nebo nepřímou závislost; tvar této závislosti pokládejme pro jednoduchost za lineární. Proto dosti často při analýze závislosti dvou náhodných veličin postupujeme buď pomocí dvojí lineární regrese nebo zavádíme ortogonální regresi.

Při dvojí lineární regresi určíme regresní rovnici pro závislost y na x a regresní rovnici pro závislost x na y, tedy rovnice:

$$y = a_{y,x} + b_{y,x} \,.\, x \qquad (4.35a)$$

$$y = a_{x,y} + b_{x,y} \,.\, y \qquad (4.35a)$$

Určení vzájemné závislosti obou proměnných je v případě dvojí regrese usnadněno tím, že platí:

$$r = \sqrt{b_{y,x} \cdot b_{x,y}} \qquad (4.35b)$$

kde $b_{y,x}$ a $b_{x,y}$ jsou směrnice rovnic (4.35a). Z toho vyplývá, že pro $r = 1$ musí $b_{y,x} = b_{x,y}^{-1}$, ale zároveň i to, že pro $0 < |r| < 1$ se regresní přímky budou od sebe lišit.

Jindy, zejména při $s_x \approx s_y$, zavádíme ortogonální regresi, která je založena na tom, že minimalizujeme kolmou vzdálenost bodu od regresní přímky. Nabývá tvaru

$$y = a + b.x$$

$$x = \frac{y-a}{b} \qquad (4.36)$$

kde

$$b = \frac{A}{2} \pm \sqrt{\left(\frac{A}{2}\right)^2 + 1}$$

$$a = \frac{1}{n}\left(\sum_i y_i - b\sum_i x_i\right)$$

a pomocná veličina

$$A = \frac{\left(\sum_i x_i\right)^2 - n\sum_i x_i^2 - \left(\sum_i y_i\right)^2 + n\sum_i y_i^2}{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}$$

Ze vztahu (4.36) vyplývá, že při ortogonální regresi nabývají obě směrnice pro jakékoliv hodnoty r vzájemně reciproké hodnoty. V případě, že použijeme ortogonální regrese, musíme koeficient korelace odhadnout pomocí vztahu (4.34).

Někdy, hlavně při posuzování kalibračních přímek v analytické chemii, se vedle koeficientu korelace zavádí tzv. koeficient determinace

$$D = 100\ r^2 \qquad (4.37)$$

který je vždy kladný a D = 0 charakterizuje nezávislost, D = 100 funkční závislost obou proměnných. Neumožňuje ovšem rozlišit, jde-li o závislost přímou nebo nepřímou.

U proměnných, mezi nimiž můžeme předpokládat jenom nepříliš silnou korelaci, určujeme někdy tzv. obrysovou elipsu (elipsu spolehlivosti). Je to elipsa, jejíž delší osa leží na přímce, která půlí úhel mezi oběma regresními přímkami podle (4.35a) a která má podobný význam, jako pás spolehlivosti regresní závislosti. Pro posouzení variability bodů má význam plocha této elipsy, která úzce souvisí s veličinou $Q = s_x s_y (1 - r^2)$, kde směrodatné odchylky

$$s_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}} \qquad \text{resp.} \qquad s_x = \sqrt{\frac{\sum_i (y_i - \bar{y})^2}{n-1}}$$

charakterizují přesnost určení obou proměnných, x, y a r je odhad korelačního koeficientu podle (4.34). Je zřejmé, že plocha obrysové elipsy je tím větší, čím větší jsou směrodatné odchylky a čím víc se |r| liší od jedné. Veličina

$$(1 - r^2) = \frac{\sum_i (y_i - Y_i)^2}{\sum_i (y_i - \bar{y})^2} \qquad (4.38)$$

je užitečná tím, že ukazuje jaká část variability hodnot y je dána jejich závislostí na hodnotách x.

Řešený příklad:

Máme vyhodnotit závislost dvou náhodných veličin, x a y, jejichž hodnoty jsou sestaveny podle velikosti x_i do schématu:

i	x_i	y_i	x_i^2	y_i^2	$x_i . y_i$
1	1,0	3,0	1,00	9,00	3,00
2	2,1	1,1	4,41	1,21	2,31
3	2,5	6,0	6,25	36,00	15,00
4	3,0	4,2	9,00	17,64	12,60
5	3,5	3,5	12,25	12,25	12,25
6	4,1	1,5	16,81	2,25	6,15
7	5,0	8,4	25,00	70,56	42,00
8	5,3	6,6	28,09	43,56	34,98
9	5,4	4,5	29,16	20,25	24,30
10	6,2	2,8	38,44	7,84	17,36
11	6,2	5,6	38,44	31,36	34,72
12	7,2	7,1	51,84	50,41	51,12
13	7,8	4,6	60,84	21,16	35,88
14	8,1	9,0	65,61	81,00	72,90
15	8,9	6,8	79,21	46,24	60,52
16	9,1	9,1	82,81	82,81	82,81
Součty:	85,4	83,8	549,16	533,54	507,90

Dvojí regrese metodou nejmenších čtverců: $y = 1{,}771 + 0{,}649\,x$

$x = 1{,}983 + 0{,}641\,y$

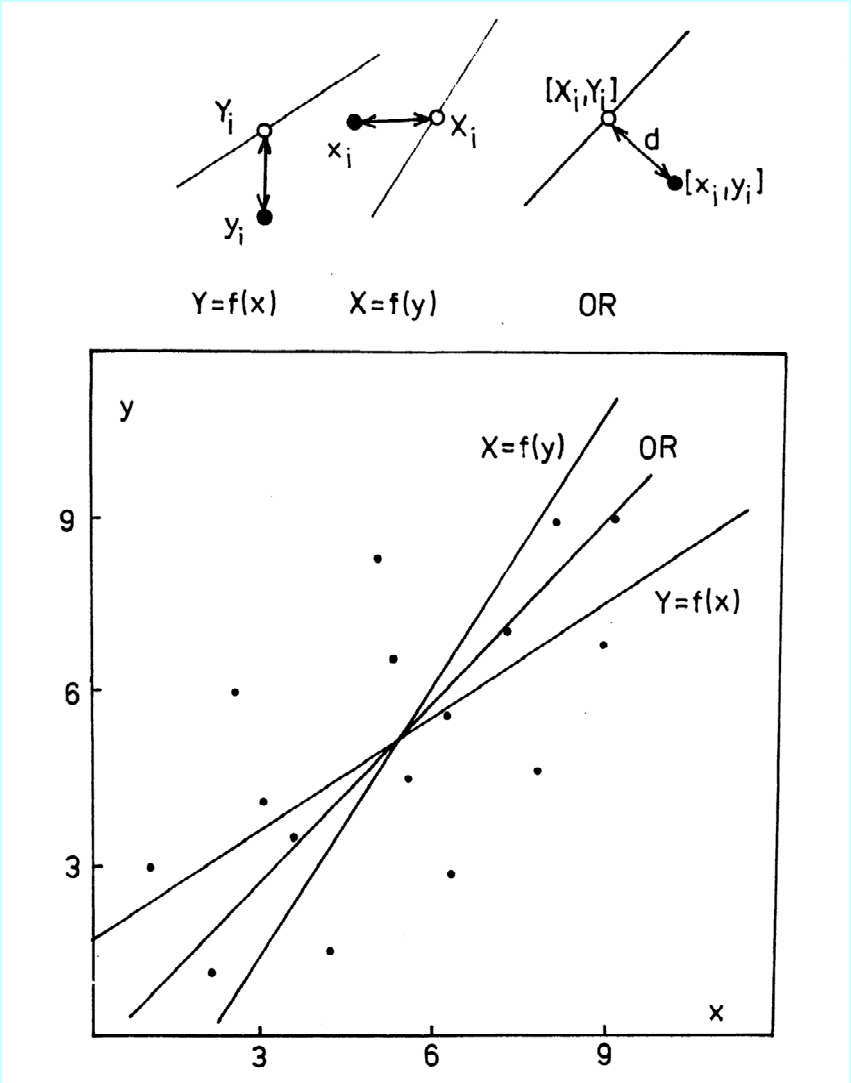
Ortogonální regrese: $y = -0{,}182 + 1{,}016\,x$

$$x = + 0{,}182 + 0{,}984\, y$$

Korelační koeficient $r = \sqrt{0{,}649 \cdot 0{,}641} = 0{,}645$ *je pro* $n = 16$ *statisticky významný na hladině* $(1 - \alpha) = 0{,}95$*; je tedy mezi oběma proměnnými přímá závislost.*

Pro ortogonální regresi je $1{,}016 \,.\, 0{,}984 = 0{,}9997 \cong 0$.

Grafické znázornění této závislosti je znázorněno na obr.10.



Obr. 10: Grafické zpracování dat z předešlého příkladu dvojí regresí (X =

f(y), Y = f(x) - metoda nejmenších čtverců) a ortogonální regresí.

4.4 Vícenásobná lineární regrese

Při statistické analýze často zjišťujeme, že měřené hodnoty y závisejí na více faktorech. Není vždy možné provést takovou sérii pokusů, při nichž by se měnila pouze hodnota jediného faktoru a ostatní byly konstantní: to by totiž umožňovalo sledovat vliv jednotlivých faktorů jednoduchou regresí. Pak je nutné zpracovat experimentální výsledky ve vztahu k současné změně všech faktorů. Jsou-li očekávané hodnoty náhodných veličin y_i lineárně závislé na několika vzájemně nezávislých proměnných $x_1, x_2, \ldots x_k$, tj. platí-li, že

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i \qquad (4.39)$$

můžeme pro výpočet odhadů koeficientů α, β_i odvodit soustavu normálních rovnic a jejich řešením určit odhady a, b_i. Tak např. pro k = 3, tj. pro tři nezávislé proměnné, jsou normální rovnice:

$$\begin{aligned}
an &+ b_1\sum x_{i1} &+ b_2\sum x_{i2} &+ b_3\sum x_{i3} &= \sum y_i \\
a\sum x_{i1} &+ b_1\sum x_{i1}^2 &+ b_2\sum x_{i1}x_{i2} &+ b_3\sum x_{i1}x_{i3} &= \sum x_{i1}y_i \\
a\sum x_{i2} &+ b_1\sum x_{i1}x_{i2} &+ b_2\sum x_{i2}^2 &+ b_3\sum x_{i2}x_{i3} &= \sum x_{i2}y_i \\
a\sum x_{i3} &+ b_1\sum x_{i1}x_{i3} &+ b_2\sum x_{i2}x_{i3} &+ b_3\sum x_{i3}^2 &= \sum x_{i3}y_i
\end{aligned}$$

Tuto soustavu rovnic pak řešíme.

Řešený příklad:

Máme určit rovnici pro vícenásobnou lineární regresi dat, sestavených podle velikosti y do následující tabulky:

y	x_1	X_2	x_3

- 8,20	1,9	0,2	15,0
- 7,26	1,0	0,6	12,8
- 0,36	2,0	0,9	6,5
+ 0,15	3,0	0,3	7,2
+ 4,87	8,3	0,1	9,0
+ 6,68	4,5	0,8	2,0
+ 7,17	7,6	0,9	5,6
+ 7,39	5,0	0,4	2,1
+ 7,85	10,0	0,5	0,8
+ 9,46	6,0	0,7	0,9
+10,93	6,6	1,0	0,1
+14,91	9,9	1,2	0,1
53,59	65,8	7,6	62,1

Grafické zpracování závislosti $y = f(x_1,x_2,x_3)$ není možné; zpracováním dílčích lineárních regresí tvaru $y = f(x_i)$, $i = 1,2,3$ metodou nejmenších čtverců dostaneme závislosti, které nejsou výstižné. Teprve trojnásobná lineární regrese $y = 3,0 + 1,2x_1 + 0,1x_2 - 0,9x_3$ popisuje velmi dobře tuto závislost.

Metody jednoduché a vícenásobné lineární regrese a polynomické regrese nacházejí velké uplatnění v chemometrii a jsou základem celé řady specielních metod. Pokud přijmeme na zkušenosti založenou představu, že se v praxi nejčastěji uplatňuje buď lineární jednoduchá i vícenásobná nebo polynomická regrese, je výhodné výraz pro závislou proměnnou y zobecnit vztahem

$$y = a_0 f_0 + a_1 f_1 + .. + a_k f_k = \sum_{i=0}^{k} a_i f_i \qquad (4.40)$$

kde a_i jsou regresní parametry a f_i jsou tzv. regresory, i = 0,1, .. , k. Regresory mohou být mocniny nezávisle proměnné $f_i = x^i$ (při lineární regresi je i = 0,1) nebo mocniny její určité funkce (např. může být $f_i = (1/x^i$ apod.). Při linearizaci může být f_i funkcí x (např. $f_1 = \log x$ apod.); při vícenásobné regresi $f_i = x_i$ apod. Také y může být funkcí nezávisle proměnné. Sestavení a řešení normálních rovnic umožní výpočet regresních parametrů a_i tak, jak bylo uvedeno dříve; při použití počítače se však lépe uplatní výpočet za použití maticových operací.

5 LITERATURA

[1] ECKSCHLAGER, K. *Chemometrie* [Skripta]. UK, Praha 1991, 156 s.