

# Úvod do zpracování měření

## Teorie chyb

Opakujeme-li měření téže fyzikální veličiny za stejných podmínek několikrát za sebou, dostáváme zpravidla různé hodnoty. Měřené veličině přísluší však jediná správná hodnota. Každou odchylku naměřené hodnoty od správné hodnoty nazýváme obecně chybou. Chybou měření  $\Delta X$  budeme rozumět rozdíl mezi hodnotou správnou  $X$  a hodnotou  $x$  získanou měřením, tedy

$$\Delta X = X - x. \quad (1)$$

Chyba může být jak kladná, tak i záporná. Je-li chyba kladná, musíme ji k naměřené hodnotě přičíst, abychom dostali hodnotu správnou, a naopak ji odečítáme, jde-li o chybu zápornou. Udáváme-li chybu rozdílem správné hodnoty a naměřené hodnoty dané veličiny, tj. absolutně, mluvíme o absolutní chybě měřené veličiny. Rovnice (1) je pak rovnicí pro absolutní chybu.

Jestliže vyjádříme chybu relativně vůči měřené hodnotě, docházíme k pojmu relativní chyby měřené veličiny. Relativní chybou  $\delta$  měřené veličiny rozumíme poměr absolutní chyby  $\Delta X$  této veličiny a správné hodnoty veličiny  $X$ . Pro relativní chybu tedy platí

$$\delta = \frac{\Delta X}{X}. \quad (2)$$

Relativní chybu lze také vyjádřit poměrem naměřené a správné hodnoty dané veličiny:

$$\delta = 1 - \frac{x}{X}. \quad (3)$$

Relativní chyba se velmi často udává v procentech. Z obou uvedených výrazů (2 i 3) je patrné, že také relativní chyba může nabývat kladných i záporných hodnot.

Podle jejich původu dělíme chyby do tří skupin:

**Chyby hrubé** vznikají při měření prováděném nedbale nebo nepozorně, s nedokonalými či vadnými přístroji, při užití nevhodné metody. Naměřená hodnota se při opakovaném měření značně liší od ostatních, a proto je nutné ji nahradit novým měřením nebo ji při konečném zpracování výsledků neuvažovat.

**Chyby systematické** (soustavné) jsou způsobeny stále stejnými a pravidelnými vlivy, tedy výsledek měření je soustavně větší nebo menší než správná hodnota. Podle toho můžeme systematické chyby přisoudit určité znaménko. Původ systematických chyb je obvykle buď v měřicí metodě (založené na určitých zjednodušujících předpokladech), v měřicích přístrojích (např. posunutí počátku (nuly) na stupnici, závislost výchylky na měřené veličině neodpovídá dělení stupnice apod.), nebo ve způsobu činnosti pozorovatele (např. odhad a zaokrouhlování zlomků dílků na stupnici, pozorování stupnice a ukazatele z nevhodného směru – chyba úkosu, paralaxa). V řadě případů je možno systematické chyby vyloučit vhodnými korekcemi. Systematické chyby nelze vyloučit statistickými metodami.

**Chyby náhodné** vznikají zcela náhodně vzájemným působením pozorovatele, přístroje a prostředí. Jejich původ nemůžeme odhalit. Každou náhodnou chybu můžeme považovat za složenou z velkého počtu velmi malých náhodně vzniklých a ojedinele nepozorovatelných elementárních chyb. O těchto elementárních chybách můžeme předpokládat, že jejich znaménka i velikosti jsou nepravidelně rozděleny a aby vznikla pozorovatelná chyba, musí se jich složit větší počet. Elementární chyby jsou kladné i záporné a jejich složením dojde pravděpodobně stejně často k chybám kladným i záporným. Nejčastěji se sejde přibližně stejný počet elementárních chyb kladných i záporných, čímž vzniknou malé náhodné chyby. Méně často se vyskytuje případ, že převažují elementárních chyby stejného znaménka, a pak

vznikne náhodná chyba větší. Takové případy jsou málo pravděpodobné, takže počet náhodných chyb bude s velikostí chyby ztlačně klesat.

Chyby systematické nás svým způsobem informují o správnosti měření, chyby náhodné o přesnosti měření.

### **Normální rozdělení**

Obecně lze říci, že toto rozdělení je použitelné všude tam, kde na kolísání náhodné veličiny působí velký počet nepatrných a vzájemně nezávislých jevů.

Pro nahodilé rozdělení měřených hodnot při počtu měření, které se blíží nekonečnu, platí vztah, odvozený Gaussem – tzv. normální statistické rozdělení, jemuž odpovídá i analogické vyjádření pro rozdělení četnosti náhodných chyb. Ze statistického rozboru tohoto problému plyne několik důležitých závěrů, které umožňují určit nejpravděpodobnější hodnotu měřené veličiny a interval, v němž se dá očekávat skutečná hodnota s předem zvolenou pravděpodobností:

Kdybychom mohli vykonat nekonečný počet měření, pak by z přesné platnosti zákona četnosti plynulo, že počet kladných chyb je rovný počtu záporných chyb a že se tedy součet všech chyb rovná nule. Aritmetický průměr všech měření by pak udával správnou hodnotu měřené veličiny. Při skutečných měřeních můžeme najít pouze nejpravděpodobnější hodnotu měřené veličiny.

Předpokládejme pro veličinu  $x$  měřením získané hodnoty  $x_1, x_2, \dots, x_n$ . Předpokládejme dále, že chyby v jednom směru (kladné odchylky) jsou právě tak pravděpodobné jako chyby ve směru druhém (záporné odchylky), takže součet všech chyb je roven nule. Označíme-li pravděpodobnou hodnotu měřené veličiny  $\bar{x}$ , pak platí

$$(\bar{x} - x_1) + (\bar{x} - x_2) + \dots + (\bar{x} - x_n) = 0 \quad (4)$$

a odtud plyne pro pravděpodobnou hodnotu  $\bar{x}$  měřené veličiny výraz

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i . \quad (5)$$

Pravděpodobnou hodnotou je aritmetický průměr naměřených hodnot. To ovšem neznamená, že aritmetický průměr je přesně rovný správné hodnotě. Jeho smysl je ten, že kdybychom měli velký počet řad o konečném počtu měření, vedl by aritmetický průměr častěji ke správné hodnotě, než kdybychom hodnotu měřené veličiny počítali jakýmkoli jiným způsobem.

Každá hodnota  $\Delta_k = \bar{x} - x_k$  udává odchylku měření od aritmetického průměru. Abychom určili střední chybu jednotlivého měření, nemůžeme odchylky sečíst a dělit počtem měření, protože součet odchylek od aritmetického průměru je rovný nule. Proto odchylky umocníme a sečteme; součet označíme  $\sum \Delta^2$ . Dělíme-li tento součet počtem měření, dostaneme průměr ze čtverců chyb, který se ve statistice nazývá rozptyl nebo také variance a značí se  $\sigma_n^2$ .

$$\sigma_n^2 = \frac{\sum \Delta^2}{n} \quad (6)$$

Odmocnina z tohoto průměru je směrodatná odchylka  $\sigma_n$

$$\sigma_n = \sqrt{\frac{\sum \Delta^2}{n}} . \quad (7)$$

Tuto hodnotu bychom mohli považovat za střední chybu jednoho měření, kdyby aritmetický průměr byl správnou hodnotou. Musíme však uvážit, že pro určení směrodatné odchylky máme k dispozici jen výběr ze souboru všech možných měření. Jedno měření potřebujeme k naměření hodnoty, zbývajících  $n-1$  měření ke kontrole výpočtu chyby. Proto

pro výpočet střední chyby jednoho měření bereme  $n-1$  místo  $n$ . Výběrová směrodatná odchylka  $\sigma_{n-1}$  nazývaná též střední kvadratická chyba jednoho měření je

$$\sigma_{n-1} = \sqrt{\frac{\sum \Delta^2}{n-1}}. \quad (8)$$

Nás však bude především zajímat, jakou chybou je zatížen výsledek měření - aritmetický průměr. Tento průměr je stanoven z většího počtu naměřených hodnot, máme tedy větší jistotu, že se skutečné hodnotě blíží aritmetický průměr, než pouze jediná hodnota měření. Projeví se to i v chybách: aritmetickému průměru přísluší menší chyby, než jednotlivým měřením. Teorie chyb vede k výsledku, že chyba aritmetického průměru je  $\sqrt{n}$  krát menší než chyba jednoho měření, přičemž  $n$  je počet měření.

Směrodatná odchylka aritmetického průměru (střední kvadratická chyba) je dána vztahem

$$\bar{\sigma} = \sqrt{\frac{\sum \Delta^2}{n(n-1)}}. \quad (9)$$

Vztah (9) není příliš vhodný pro praktický výpočet, protože pro výpočet odchylek od průměru je třeba mít průměr předem vypočítaný. Můžeme vyjádřit:

$$\sum \Delta^2 = \sum (x_i - \bar{x})^2 = \sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum x_i^2 - \frac{1}{n}(\sum x_i)^2, \quad (10)$$

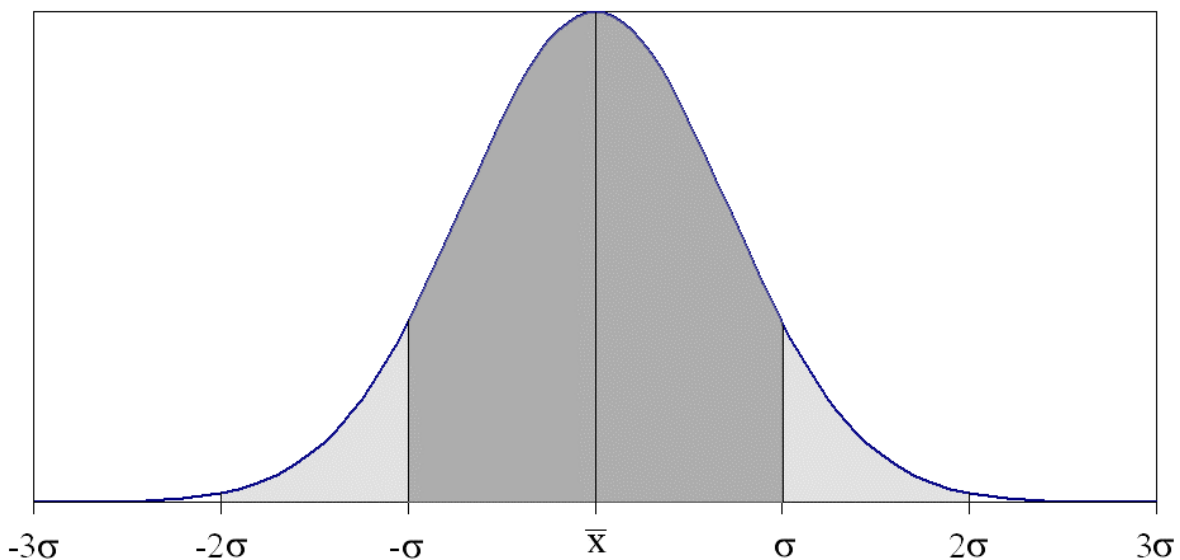
využili jsme při tom skutečnosti, že

$$2\bar{x} \sum x_i = 2 \frac{\sum x_i}{n} \sum x_i = \frac{2}{n}(\sum x_i)^2 \text{ a } \sum \bar{x}^2 = n\bar{x}^2 = n \frac{(\sum x_i)^2}{n^2} = \frac{1}{n}(\sum x_i)^2. \quad (11)$$

Do (1.9) dosadíme (1.10) a (1.11) a dostaneme

$$\bar{\sigma} = \sqrt{\frac{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}{n(n-1)}}. \quad (12)$$

Na obrázku 1 je nakreslena funkce hustoty pravděpodobnosti pro normální rozdělení.



Obr. 1: Funkce hustoty pravděpodobnosti pro normální rozdělení.

Plocha pod křivkou (integrál funkce) je úměrná pravděpodobnosti, se kterou správná hodnota může nabývat hodnot vynesných na ose  $x$ . V intervalu  $(-\sigma, \sigma)$  (na obr. 1 vybarveno tmavě) je tato pravděpodobnost 0,683, to znamená, že v tomto intervalu by mělo být 68% hodnot. V intervalu  $(-2\sigma, 2\sigma)$  (na obr. 1 vybarveno světle i tmavě) je pravděpodobnost 0,955. V intervalu  $(-3\sigma, 3\sigma)$  pak je to 0,997. Kromě střední chyby uvádíme někdy také pravděpodobnou chybu, která je rovna  $2/3$  střední chyby. Její význam je tento: je stejně pravděpodobné, že chyba jednoho měření (libovolně vybraného) je menší než pravděpodobná chyba, jako že tato chyba je větší než pravděpodobná chyba. Při velkém počtu měření je tedy polovina skutečných chyb menší, druhá polovina větší než pravděpodobná chyba. Pravděpodobná chyba jednoho měření je

$$\vartheta = \frac{2}{3} \sigma_{n-1}. \quad (13)$$

V některých případech používáme ještě krajní chybu  $\chi$ , která je rovna trojnásobku střední chyby:  $\chi = 3\sigma$ . U krajní chyby máme pravděpodobnost 99,73 %, že se nám v měření nevyskytne hodnota s chybou větší než je krajní chyba. Jednu chybu větší než  $\chi$  můžeme tedy očekávat průměrně v 370 měřeních.

Vzájemné vztahy mezi uvedenými chybami jsou následující:

$$\vartheta : \sigma : \chi = 0,67 : 1 : 3. \quad (14)$$

Stejně vztahy jako mezi chybami jednoho měření jsou i mezi odpovídajícími chybami aritmetického průměru.

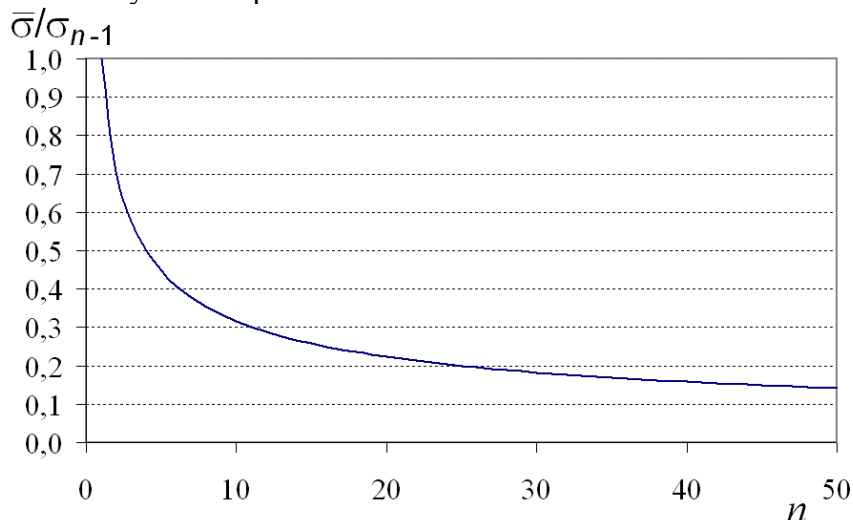
Na Gaussově křivce (obr. 1) odpovídá pravděpodobné chybě hodnota, jejíž pořadnice dělí plochu Gaussovy křivky na části, z nichž prostřední zaujímá polovinu celkové plochy, obě krajní také polovinu. Geometrický význam střední kvadratické chyby je ten, že v místě  $\sigma$  má Gaussova křivka inflexní bod.

Aby bylo zřejmé, do jaké míry je zaručen výsledek měření, připisujeme k němu jeho střední kvadratickou chybu. Pišeme tedy výsledek ve tvaru:

$$x = \bar{x} \pm \bar{\sigma}. \quad (15)$$

Číselně uvádíme chybu zpravidla pouze na jedno platné místo a počet číslic ve výsledku omezíme tak, aby chyba zasahovala pouze do posledního místa. Například:  $m=(27,32 \pm 0,04)$  g. V případě, že je mantisa chyby 1, uvádíme chybu zpravidla na dvě místa (např.  $m=(27,32 \pm 0,12)$  g).

Pokud z nějakých důvodů uvádíme jinou chybu než střední kvadratickou, je třeba na tento fakt v textu výslovně upozornit!



Obr. 2: Závislost chyby průměru na počtu měření. Chyba průměru je vynášena jako násobek výběrové chyby jednoho měření.

Je zřejmé, že čím větší počet měření vykonáme, tím máme větší jistotu při stanovení výsledné hodnoty a tím menší bude chyba výsledku. Závislost střední chyby aritmetického průměru na počtu měření je graficky znázorněna na obr.2. Vidíme, že se vzrůstajícím počtem měření klesá chyba aritmetického průměru zpočátku prudce, pak mírně. Z této křivky můžeme odhadnout, kolik musíme vykonat měření, abychom dosáhli požadované přesnosti. Obvykle stačí měřit desetkrát; při dalším zvyšování počtu měření vzrůstá přesnost výsledku jen velmi zvolna.

### **Výpočet aritmetického průměru a chyby (příklad)**

#### **Ruční zpracování**

Posuvným měřítkem byla stanovena desetkrát tloušťka  $x$ , přičemž byly odhadovány ještě desetiny dílků stupnice. Výsledky jsou uvedeny v tabulce:

i	$x_i$ (cm)	$x_i^2$ (cm <sup>2</sup> )
1	0,256	0,0655
2	0,258	0,0666
3	0,255	0,0650
4	0,255	0,0650
5	0,254	0,0645
6	0,256	0,0655
7	0,257	0,0660
8	0,255	0,0650
9	0,259	0,0671
10	0,254	0,0645
$\Sigma$	2,559	0,654873

$$\bar{x} = \frac{\sum x_i}{n} = \frac{2,559}{10} = 0,2559 \text{ cm}$$

Aritmetický průměr tloušťky je 0,2559 cm.

$$\bar{\sigma} = \sqrt{\frac{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}{n(n-1)}} = \sqrt{\frac{0,654873 - 2,559^2/10}{90}} = \sqrt{\frac{0,654873 - 0,6548481}{90}} = 0,00053 \text{ cm.}$$

Směrodatná odchylka aritmetického průměru je 0,00053 cm.

Zaokrouhlíme na jednu platnou číslici:  $\bar{\sigma} = \pm 0,0005$  cm.

Výsledek měření napíšeme tak, že k aritmetickému průměru připišeme střední kvadratickou chybu zaokrouhlenou na jedno platné místo:  $x = (0,2559 \pm 0,0005)$  cm. Pravděpodobná chyba aritmetického průměru je rovna dvěma třetinám směrodatné odchylky:

$$\bar{\vartheta} = \frac{2}{3} \bar{\sigma} = \pm \frac{2}{3} \cdot 0,00053 = \pm 0,00035 \text{ cm.}$$

#### **Zpracování v Excelu**

Výpočet průměru s směrodatné odchylky průměru je v Excelu velmi jednoduchý. Pro výpočet aritmetického průměru obsahuje funkci PRŮMĚR(). Pro směrodatnou odchylku průměru není k dispozici přímá funkce a je třeba použít funkce SMODCH(), která vrací směrodatnou odchylku jednoho měření vypočítanou podle vztahu (1.7). Abychom získali směrodatnou odchylku průměru, je třeba tuto hodnotu vydělit, v souladu se vztahem (1.9), odmocninou z počtu měření zmenšeného o jednu.

	A	B	C
1	<i>i</i>	<i>x</i>	vzorec
2	1	0,256	
3	2	0,258	
4	3	0,255	
5	4	0,255	
6	5	0,254	
7	6	0,256	
8	7	0,257	
9	8	0,255	
10	9	0,259	
11	10	0,254	
12	$\bar{x}$	0,2559	=PRŮMĚR(B2:B11)
13	$\bar{\sigma}$	0,00053	=SMODCH(B2:B11)/ODMOCNINA(A11-1)

Obr. 2: Výřez listu Excelu s výpočtem průměru a jeho směrodatné odchylky.

### Výpočet chyby hodnoty funkce z chyb nezávisle proměnných

Než přejdeme k určení chyby aritmetického průměru, předpokládejme, že máme z výsledků měření několika vzájemně nezávislých veličin  $x, y, z, \dots$ , určit hodnotu veličiny

$$V = f(x, y, z, \dots) \quad (16)$$

(Veličina  $V$  je tedy výsledkem nepřímých měření).

Jsou-li chyby jednotlivých měřených veličin  $\sigma(x), \sigma(y), \sigma(z), \dots$ , (nemusí to ovšem být právě směrodatné odchylky, mohou to být chyby odpovídající jiné pravděpodobnosti výskytu, avšak pro všechny veličiny  $x, y, z, \dots$ , stejného druhu), pak při počítání chyby veličiny  $V$  s nimi pracujeme podobně jako s diferenciály nezávisle proměnných. Z teorie pravděpodobnosti pro chybu veličiny  $V$  dostáváme

$$\sigma(V) = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 [\sigma(x)]^2 + \left(\frac{\partial f}{\partial y}\right)^2 [\sigma(y)]^2 + \dots} \quad (17)$$

Je-li  $V = f(x)$  (funkce jedné nezávisle proměnné), pak

$$\sigma(V) = \left| \frac{df}{dx} \right| \sigma(x) = |f'(x)| \sigma(x). \quad (18)$$

Např. je-li  $V = ax$ , je  $\sigma(V) = a\sigma(x)$ .

Zavedeme-li relativní chybu  $\delta(x) = \frac{\sigma(x)}{x}$  pak pro  $V = x^k$  je

$$\delta(V) = kx^{k-1} \sigma(x) \quad (19)$$

a odtud

$$\delta(V) = \frac{\sigma(V)}{V} = k \cdot \delta(x). \quad (20)$$

Pro  $V = x \pm y$  je  $\frac{\partial f}{\partial x} = 1$ ,  $\frac{\partial f}{\partial y} = \pm 1$

a

$$\sigma(V) = \sqrt{\sigma^2(x) + \sigma^2(y)}. \quad (21)$$

Geometricky to znamená, že chybu součtu nebo rozdílu dvou veličin určíme jako délku přepony v pravouhlém trojúhelníku, o odvěsnách rovných velikostem chyb jednotlivých sčítanců. Toto pravidlo snadno rozšíříme i na větší počet sčítanců.

Pro součin  $V = x \cdot y$  dostaneme

$$\sigma(V) = \sqrt{x^2 \cdot \sigma^2(y) + y^2 \cdot \sigma^2(x)} \quad (22)$$

nebo relativní chybu součinu

$$\delta(V) = \sqrt{\delta^2(x) + \delta^2(y)} \quad (23)$$

Vidíme, že relativní chyba součinu je vyjádřena podobným vztahem, jako absolutní chyba součtu. Snadno se odvodí podobné vztahy i pro chybu součinu  $V = x \cdot y \cdot z$  a podílu

$V = \frac{x}{y}$ . (Odvod'te sami, obojí i pro relativní chyby).

**Příklad 1:** Vypočteme objem  $V$  válečku a jeho střední kvadratickou chybu  $\sigma(V)$  užitím vzorce  $V = \pi \cdot r^2 \cdot h$ , kde  $r$  je poloměr válečku a  $h$  jeho výška.

Mikrometrem byl změřen průměr  $d$  válečku:  $d = (2,442 \pm 0,004)$  cm, posuvným měřítkem výška  $h$  válečku:  $h = (4,56 \pm 0,01)$  cm. Vypočteme nejdříve poloměr válečku.

Poloměr  $r = \frac{d}{2} = 1,221$  cm. Střední chyba poloměru je rovna polovině střední chyby

průměru:  $\sigma(\bar{r}) = \frac{\sigma(\bar{d})}{2} = 0,002$  cm. Poloměr válečku je tedy  $r = (1,221 \pm 0,002)$  cm.

Dosadíme do vzorce  $V = \pi \cdot r^2 \cdot h$ :

$$V = 3,142 \cdot (1,221)^2 \cdot 4,56 = 21,25.$$

Protože průměr je měřen na čtyři místa, výška na tři, počítáme objem zkráceně na čtyři místa. Objem  $V = 21,25$  cm<sup>3</sup>.

Střední chybu tohoto výsledku vypočteme dosazením do vzorce

$$\sigma(V) = \pm \sqrt{\left[ \frac{\partial(V)}{\partial r} \sigma(\bar{r}) \right]^2 + \left[ \frac{\partial(V)}{\partial h} \sigma(\bar{h}) \right]^2}.$$

Protože parciální derivace  $\frac{\partial V}{\partial r}$  a  $\frac{\partial V}{\partial h}$  jsou

$$\frac{\partial V}{\partial r} = 2\pi r h, \quad \frac{\partial V}{\partial h} = \pi r^2,$$

je  $\sigma(\bar{V}) = \pm \sqrt{[2\pi r h \sigma(\bar{r})]^2 + [\pi r^2 \sigma(\bar{h})]^2}$

a po úpravě

$$\sigma(\bar{V}) = \pm \sqrt{(\pi r^2 h)^2 \left[ \frac{2\sigma(\bar{r})}{r} \right]^2 + (\pi r^2 h)^2 \left[ \frac{\sigma(\bar{h})}{h} \right]^2}.$$

Absolutní střední chyba výsledku je dána vzorcem

$$\sigma(\bar{V}) = \pm V \sqrt{\left[ \frac{2\sigma(\bar{r})}{r} \right]^2 + \left[ \frac{\sigma(\bar{h})}{h} \right]^2}$$

a relativní chyba

$$\delta(\bar{V}) = \frac{\sigma(\bar{V})}{V} = \pm \sqrt{\left[ \frac{2\sigma(\bar{r})}{r} \right]^2 + \left[ \frac{\sigma(\bar{h})}{h} \right]^2}.$$

Numericky počítáme absolutní střední chybu objemu na jedno místo, tj pod odmocninou na dvě místa různá od nuly:

$$\sigma(\bar{V}) = \pm 21,25 \cdot \sqrt{\left( \frac{0,004}{1,221} \right)^2 + \left( \frac{0,01}{4,56} \right)^2} = \pm 21,25 \cdot \sqrt{0,0033^2 + 0,0022^2} =$$

$$= \pm 21,25 \cdot \sqrt{0,000011 + 0,000005} = 21,25 \cdot \sqrt{0,000016} = \\ = \pm 21,25 \cdot 0,004 \doteq 0,08.$$

Střední kvadratická chyba objemu válečku je  $0,08 \text{ cm}^3$ . Výsledek píšeme ve tvaru:  
Objem  $V = (21,25 \pm 0,08) \text{ cm}^3$ .

**Poznámka 1:** Při výpočtu jsme viděli, že relativní chyba poloměru, který je ve vzorci pro výpočet objemu ve druhé mocnině, se uplatnila dvojnásobně, bylo by proto vhodné měřit poloměr s větší přesností!

Vypočteme ještě relativní chybu objemu:

$$\frac{\sigma(\bar{V})}{V} = \frac{0,08}{21,25} = 0,0038.$$

Relativní chyba objemu je  $0,0038$ , tj. přibližně  $0,4\%$ .

**Poznámka 2:** U funkcí typu  $u = x^k y^m z^n$  vychází pro relativní chybu vztah

$$\frac{\sigma(\bar{u})}{u} = \pm \sqrt{\left[ \frac{k \cdot \sigma(\bar{x})}{x} \right]^2 + \left[ \frac{m \cdot \sigma(\bar{y})}{y} \right]^2 + \left[ \frac{n \cdot \sigma(\bar{z})}{z} \right]^2}.$$

**Příklad 2:** Určete objem koule z naměřeného průměru  $d$ :

Protože platí vztah  $V = \frac{\pi}{6} d^3$ ,

určíme  $\sigma(\bar{V}) = \frac{\pi}{2} d^2 \sigma(\bar{d})$ ,

takže  $\frac{\sigma(\bar{V})}{V} = 3 \frac{\sigma(\bar{d})}{d}$ .

Sami zvažte, co plyne z provedeného rozboru chyby.

## Regresní analýza

V praxi se často setkáme s úkolem, kdy nějaká proměnná  $y$  je funkcí nezávisle proměnné  $x$ , tedy  $y=f(x)$ . Z hodnot  $\{x_i, y_i\}$  pak máme odhadnout parametry funkční závislosti. Zpravidla předpokládáme, že hodnoty  $x_i$  jsou dány pevně a hodnoty  $y_i$  byly získány měřením. Kdyby měření hodnot  $y_i$  nebylo zatíženo chybami, platilo by  $y_i=f(x_i)$ . Ve skutečnosti však platí  $y_i=f(x_i)+\Delta_i$ , kde  $\Delta_i$  je chyba  $i$ -tého měření. Body  $[x_i, y_i]$  jsou pak vlivem chyb rozptýleny kolem křivky  $y=f(x)$ . Obecně funkce  $y=f(x)$  obsahuje  $p$  neznámých konstant - parametrů, které označíme  $b_0, \dots, b_{p-1}$ . Máme-li soustavou bodů  $[x_i, y_i]$  proložit křivku  $y=f(x; b_0, \dots, b_{p-1})$ , musíme určit (statisticky odhadnout) neznámé parametry  $b_0, \dots, b_{p-1}$ , které se vyskytují v rovnici křivky. Při tom vyžadujeme, aby se křivka co nejvíc přiblížila blížíla bodům  $[x_i, y_i]$ . Statistický odhad parametru  $b_i$  označme  $\beta_i$ . Způsob odhadu  $\beta_i$  závisí na tom, jak definujeme "přiblížení". Mohli bychom například požadovat, aby součet absolutních hodnot odchylek bodů od křivky byl minimální. V praxi se však nejčastěji za kritérium přiblížení považuje suma čtverců hodnot  $y_i=f(x_i; \beta_0, \dots, \beta_{p-1})$  a odhadem parametrů  $\beta_0, \dots, \beta_{p-1}$  jsou pak hodnoty, které tento součet čtverců minimalizují.

Označíme-li



$$S = \sum_{i=1}^n (y_i - f(x_i; \beta_0, \dots, \beta_{p-1}))^2, \quad (24)$$

budou odhady  $\beta_i$  určeny z podmínky

$$S = \min. \quad (25)$$

Touto podmínkou je vyjádřen princip metody nejmenších čtverců. O křivce  $y=f(x; \beta_0, \dots, \beta_{p-1})$  říkáme, že byla body  $[x_i, y_i]$  proložena metodou nejmenších čtverců.

Nejčastěji se setkáme s případem, kdy je očekávaná závislost lineární

$$y=b_0+b_1x. \quad (26)$$

Chceme tedy nalézt parametry  $\beta_0$  a  $\beta_1$  tak, aby co nejlépe odpovídaly zadaným bodům.

Podle (24 a 25) můžeme odhady  $\beta_0$  a  $\beta_1$  určit z podmínky

$$S = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = \min. \quad (27)$$

Hodnoty parametrů  $\beta_0$  a  $\beta_1$ , které minimalizují sumu čtverců odchylek  $S$ ,

$$\frac{\partial S}{\partial \beta_0} = 0 \quad \text{a} \quad \frac{\partial S}{\partial \beta_1} = 0 \quad (28)$$

dostaneme soustavu dvou rovnic

$$\frac{\partial S}{\partial \beta_0} = n\beta_0 + \beta_1 \sum_{i=1}^n x_i - \sum_{i=1}^n y_i = 0 \quad (29)$$

a

$$\frac{\partial S}{\partial \beta_1} = \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i = 0 \quad (30)$$

Její řešení získáme odhady  $\beta_0$  a  $\beta_1$ , parametrů  $b_0$  a  $b_1$ :

$$\beta_0 = \frac{1}{n} \left( \sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i \right) \quad (31)$$

$$\beta_1 = \frac{n \sum_{i=1}^n x_i y_i - \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \quad (32)$$

Podobně lze nalézt odhady parametrů i pro jiné (složitější) regresní funkce.

Bez odvození napíšeme odhady směrodatných odchylek  $\sigma_{\beta_0}$  a  $\sigma_{\beta_1}$  parametrů  $\beta_0$  a  $\beta_1$ .

Označme:

$$S_0 = \sum_{i=1}^n y_i^2 - \frac{\left( \sum_{i=1}^n y_i \right)^2}{n} - \beta_1 \left( \sum_{i=1}^n x_i y_i - \frac{\left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{n} \right), \quad (33)$$

$$s = \sqrt{\frac{S_0}{n-2}}, \quad (34)$$

směrodatné odchylky parametrů  $\beta_0$  a  $\beta_1$  pak vypočítáme ze vztahů:

$$\sigma_{\beta_0} = s \sqrt{\frac{\frac{1}{n} \sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2}} \quad (35)$$

$$\sigma_{\beta_1} = \frac{s}{\sqrt{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2}} \quad (36)$$

### Výpočet lineární regrese pomocí Excelu

K výpočtu lineární regrese metodou nejmenších čtverců slouží v Excelu funkce =LINREGRESE(), která vrací matici parametrů regresní funkce. Protože funkce vrací matici, je třeba s ní pracovat jako s maticovým vzorcem:

1) označíme v listu Excelu prázdnou oblast o pěti řádcích a dvou sloupcích, do které se umístí výsledky lineární regrese.

2) zadáme vzorec =LINREGRESE(y;x;b;stat), kde  $y$  je pole závisle proměnných (sloupec hodnot  $y$ ),  $x$  je pole nezávisle proměnných (sloupec hodnot  $x$ ),  $b$  je logická hodnota udávající, zda má být konstanta  $\beta_0$  rovna 0 (je-li  $b$  PRAVDA nebo 1, hodnota  $\beta_0$  se počítá, je-li  $b$  NEPRAVDA nebo 0, je pevně dáno  $\beta_0 = 0$ ).

3) po napsání vzorce zmáčkneme současně klávesy Ctrl+Shift+Enter (tím říkáme, že se má vzorec rozepsat do všech prvků matice); nebude-li vám výpočet regresní přímky fungovat, s vysokou pravděpodobností jste místo Ctrl+Shift+Enter odklepli jen Enter

Výsledná matice pak obsahuje hodnoty:

$\beta_1$	$\beta_0$
$\sigma_{\beta_1}$	$\sigma_{\beta_0}$
$r^2$	$\sigma$
$F$	počet stupňů volnosti
$SS_{reg}$	$SS_{resid}$

kde  $\beta_0$  a  $\beta_1$  jsou odhady parametrů  $b_0$  a  $b_1$  z rovnice (26),  $\sigma_{\beta_0}$  a  $\sigma_{\beta_1}$  jsou jejich směrodatné odchylky,  $r^2$  je koeficient determinace,  $\sigma$  směrodatná odchylka odhadu  $y$ ,  $F$  je  $F$ -statistika (používá se při statistickém testování), počet stupňů volnosti (v případě regresní rovnice (26) je to počet hodnot zmenšená o 2),  $ss_{reg}$  je regresní součet čtverců a  $ss_{resid}$  reziduální součet čtverců.

### Korelační koeficient

Mějme dvě řady proměnných  $x_i$  a  $y_i$ . V předchozích kapitolách jsme se pokoušeli nalézt parametry optimálně charakterizující vztah mezi těmito proměnnými. Míru závislosti mezi proměnnými je možné částečně odhadnout ze směrodatných odchylek parametrů charakterizujících tento vztah, kdy můžeme předpokládat, že čím větší jsou relativní chyby těchto parametrů, tím slabší bude závislost. My však potřebujeme kvantitativní veličinu, která nám popíše, jak se změní veličina  $y$  při nějaké změně veličiny  $x$ . Při tom veličiny  $x$  a  $y$  mohou být zcela nesouměřitelné. Abychom mohli veličiny  $x$  a  $y$  srovnat, musíme je standardizovat a to tak, že od každé veličiny odečteme průměr a rozdíl vydělíme směrodatnou odchylkou. Standardizované veličiny  $x_i^+$  a  $y_i^+$  jsou definovány vztahy:

$$x_i^+ = (x_i - \bar{x}) / \sigma_x, \quad (37)$$

$$y_i^+ = (y_i - \bar{y}) / \sigma_y. \quad (38)$$

Tím jsme zajistili, že  $x_i^+$  i  $y_i^+$  mají nulovou střední hodnotu a jednotkovou směrodatnou odchylku. V tomto okamžiku už můžeme diskutovat o tom, jak se změní  $y_i^+$  při nějaké změně  $x_i^+$ . Veličinou, která popisuje tento vztah, je korelační koeficient  $r$ . Korelační koeficient je možné vypočítat ze vztahu:

$$r = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{\left[ n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 \right] \left[ n \sum_{i=1}^n y_i^2 - \left( \sum_{i=1}^n y_i \right)^2 \right]}} \quad (39)$$

Zaměříme se na otázku, jakých hodnot může korelační koeficient nabývat. Existuje-li mezi veličinami  $x$  a  $y$  pozitivní lineární závislost, pak vzroste-li  $x$  o jednu směrodatnou odchylku, vzroste i  $y$  o jednu směrodatnou odchylku a  $r=1$ . Existuje-li mezi veličinami  $x$  a  $y$  negativní lineární závislost, pak vzroste-li  $x$  o jednu směrodatnou odchylku, klesne  $y$  o jednu směrodatnou odchylku a  $r = -1$ . Není-li mezi proměnnými žádná závislost, nedojde při jakékoliv změně proměnné  $x$  k žádné změně proměnné  $y$  a korelační koeficient  $r = 0$ .

Už rozumíme, jaký význam mají extrémní hodnoty korelačního koeficientu. Pokusme se teď interpretovat, jaký význam má korelační koeficient 0,43 nebo  $-0,16$ . Hodnota 0,43 indikuje, že s rostoucím  $x$  roste  $y$ , hodnota  $-0,16$  pak znamená, že s rostoucím  $x$  klesá  $y$ .

Pomocí korelačního koeficientu můžeme testovat nulovou hypotézu  $r=0,0$  (mezi proměnnými  $x$  a  $y$  není závislost). Testovací veličinou je

$$t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}}. \quad (40)$$

Je-li testovací veličina  $t$  větší než hodnota Studentova rozdělení na dané hladině významnosti a s příslušným počtem stupňů volnosti, můžeme zamítnout nulovou hypotézu  $r=0,0$ .

Druhá mocnina korelačního koeficientu se nazývá koeficient determinace a určuje, jak velká část rozptylu veličiny  $y$  je vysvětlitelná veličinou  $x$ .

## Výpočet regresní přímky (příklad)

### Ruční výpočet

Mějme deset experimentálně zjištěných dvojic  $x_i$  a  $y_i$  zadaných prvními třemi sloupci následující tabulky. Dopočítejme hodnoty  $x^2$ ,  $y^2$  a  $xy$ , a doplňme je do dalších tří sloupců. Spočítejme v každém sloupci součet hodnot a zapišme ho do posledního řádku tabulky:

$i$	$x$	$y$	$x^2$	$y^2$	$xy$
1	10	338	100	114308	3381
2	20	343	400	117391	6852
3	30	348	900	121266	10447
4	40	354	1600	125646	14179
5	50	361	2500	130323	18050
6	60	368	3600	135382	22077
7	70	372	4900	138620	26062
8	80	379	6400	143365	30291
9	90	385	8100	147910	34613
10	100	390	10000	152217	39015
<b><math>\Sigma</math></b>	<b>550</b>	<b>3638</b>	<b>38500</b>	<b>1326428</b>	<b>204967</b>

V posledním řádku tabulky máme všechny veličiny potřebné pro vyčíslení vztahů (31-36).

Dosažením do vztahu (32) vypočítáme:

$$\beta_1 = \frac{10 \cdot 204967 - 550 \cdot 3638}{10 \cdot 38500 - 550^2} = \frac{48770}{82500} = 0,591$$

a dosazením do vztahu (31):

$$\beta_0 = \frac{1}{10}(3638 - 0,591 \cdot 550) = 331,3.$$

Tím jsme vypočítali odhady parametrů  $\beta_0$  a  $\beta_1$ . Nyní odhadneme jejich směrodatné odchylky. Ze vztahu (33) vypočteme  $S_0 = 3,5112$  a pak ze vztahu (34)  $s = 0,6625$ . Veličinu  $s$  dosadíme do vztahů (35) a (36) a dostaneme  $\sigma_{\beta_0} = 0,45$  a  $\sigma_{\beta_1} = 0,0073$ . Vypočítali jsme tedy odhady parametrů regrese  $\beta_0 = (331,3 \pm 0,5)$  a  $\beta_1 = 0,591 \pm 0,007$ .

### Výpočet v Excelu

Vstupní hodnoty jsou uloženy v listu Excelu:

	A	B	C
1	i	x	y
2	1	10	338
3	2	20	343
4	3	30	348
5	4	40	354
6	5	50	361
7	6	60	368
8	7	70	372
9	8	80	379
10	9	90	385
11	10	100	390
12	suma	550	3638

Vybereme v listu Excelu oblast o dvou sloupcích a pěti řádcích a do příkazového řádku vepíšeme vzorec =LINREGRESE(C2:C11;B2:B11;1;1):

The screenshot shows the Excel interface. The formula bar contains the formula =LINREGRESE(C2:C11;B2:B11;1;1). A selection of cells is highlighted in the worksheet, covering the range A15:B19. The formula bar also shows 'CORREL' and a dropdown arrow.

Vložíme vzorec maticově do vybraných buněk současným stiskem kláves Ctrl+Shift+Enter:

The screenshot shows the Excel interface after entering the array formula. The formula bar shows the array formula = {=LINREGRESE(C2:C11;B2:B11;1;1)}. The result is displayed in the selected cells A15:B19:

	A	B	C	D	E	F
14						
15	0,591	331,313				
16	0,007	0,453				
17	0,999	0,662				
18	6560,338	8,000				
19	2879,355	3,511				
20						

Ze zvolené oblasti nás nejvíc zajímají první dva řádky. V prvním řádku jsou odhady regresních parametrů, ve druhém pak jejich směrodatné odchylky.