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 ΔX budeme rozumět rozdíl mezi hodnotou správnou X a hodnotou x získanou měřením, tedy

$$\Delta X = X - x . \tag{1}$$

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 δ měřené veličiny rozumíme poměr absolutní chyby ΔX této veličiny a správné hodnoty veličiny X. Pro relativní chybu tedy platí

$$\delta = \frac{\Delta X}{X} . \tag{2}$$

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

$$\delta = 1 - \frac{x}{X} . \tag{3}$$

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 chybě přisoudit určité znaménko. Původ systematických chyb je obvykle buď v měřící metodě (založené na určitých zjednodušujících předpokladech), v měřících 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 ojediněle 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 znatelně 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, ….. , 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

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

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

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

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 σ_n^2.

$$\sigma_n^2 = \frac{\sum \Delta^2}{n} \tag{6}$$

Odmocnina z tohoto průměru je směrodatná odchylka σ_n

$$\sigma_n = \sqrt{\frac{\sum \Delta^2}{n}} \ . \tag{7}$$

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 σ_{n-1} nazývaná též střední kvadratická chyba jednoho měření je

$$\sigma_{n-1} = \sqrt{\frac{\sum \Delta^2}{n-1}}\,. \tag{8}$$

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

$$\overline{\sigma} = \sqrt{\frac{\sum \Delta^2}{n(n-1)}}\,. \tag{9}$$

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

$$\sum \Delta^2 = \sum \left(x_i - \overline{x}\right)^2 = \sum \left(x_i^2 - 2x_i\overline{x} + \overline{x}^2\right) = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2, \tag{10}$$

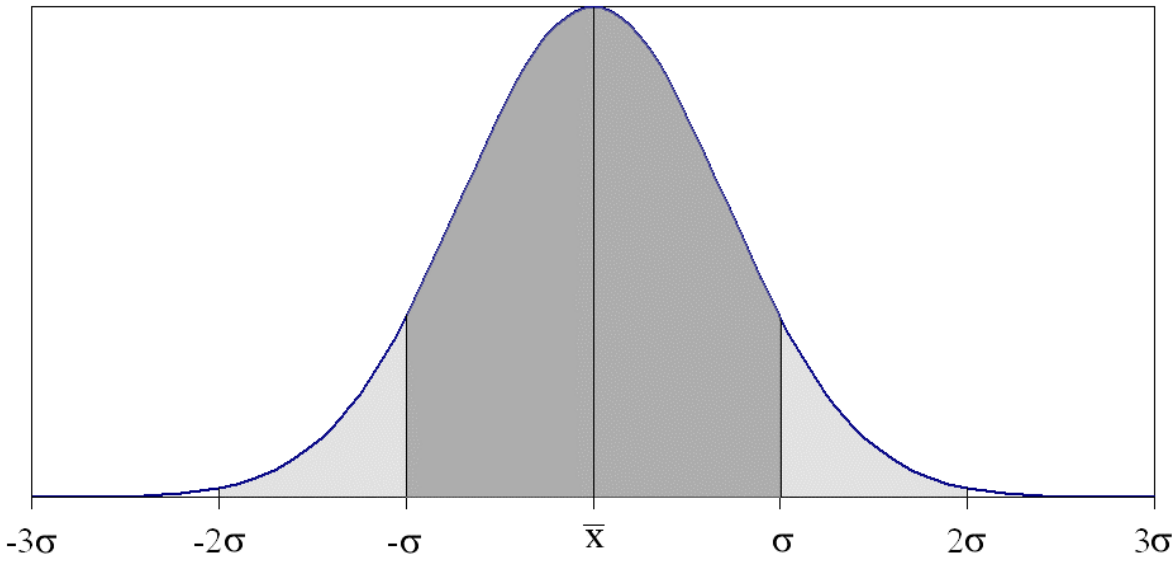
využili jsme při tom skutečnosti, že

$$2\overline{x}\sum x_i = 2\frac{\sum x_i}{n}\sum x_i = \frac{2}{n}\left(\sum x_i\right)^2 \text{ a } \sum \overline{x}^2 = n\overline{x}^2 = n\frac{\left(\sum x_i\right)^2}{n^2} = \frac{1}{n}\left(\sum x_i\right)^2. \tag{11}$$

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

$$\overline{\sigma} = \sqrt{\frac{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}{n(n-1)}}\,. \tag{12}$$

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 vynesených na ose *x*. V intervalu (-σ, σ) (na obr. 1 vybarveno tmavě) je tato pravděpodobnost 0,683, to znamená. že v tomto intervalu by mělo být 68% hodnot. V intervalu (-2σ, 2σ) (na obr. 1 vybarveno světle i tmavě) je pravděpodobnost 0,955. V intervalu (-3σ, 3σ) 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

$$\vartheta = \frac{2}{3}\sigma_{n-1}. \tag{13}$$

V některých případech používáme ještě krajní chybu χ, 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 χ můžeme tedy očekávat průměrně v 370 měřeních.

Vzájemné vztahy mezi uvedenými chybami jsou následující:

$$\vartheta : \sigma : \chi = 0{,}67 : 1 : 3. \tag{14}$$

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

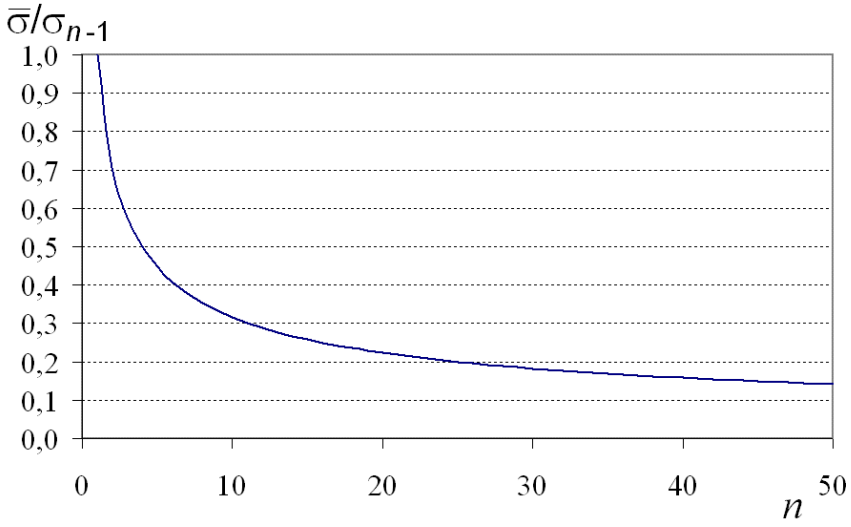
Na Gaussově křivce (obr. 1) odpovídá pravděpodobné chybě hodnota, jejíž pořadnice dělí plochu Gaussovy křivky na části, z nichž prostřední zaujímá polovinu celkové plochy, obě krajní také polovinu. Geometrický význam střední kvadratické chyby je ten, že v místě σ má Gaussova křivka inflexní bod.

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

$$x = \bar{x} \pm \bar{\sigma}. \tag{15}$$

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 ± 0,04) g. V případě, že je mantisa chyby 1, uvádíme chybu zpravidla na dvě místa (např. *m*=(27,32 ± 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^2)
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
Σ	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.

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

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 připíšeme střední kvadratickou chybu zaokrouhlenou na jedno platné místo: x=(0,2559±0,0005) cm. Pravděpodobná chyba aritmetického průměru je rovna dvěma třetinám směrodatné odchylky:

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

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	x	vzorce
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, ……, určit hodnotu veličiny

$$V = f(x,y,z,\ldots) \tag{16}$$

(Veličina V je tedy výsledkem nepřímých měření).

Jsou-li chyby jednotlivých měřených veličin $\sigma(x)$, $\sigma(y)$, $\sigma(z)$, ……, (nemusí to ovšem být právě směrodatné odchylky, mohou to být chyby odpovídající jiné pravděpodobnosti výskytu, avšak pro všechny veličiny x, y, z, …., 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

$$\sigma(V) = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 [\sigma(x)]^2 + \left(\frac{\partial f}{\partial y}\right)^2 [\sigma(y)]^2 + \ldots..} \,. \tag{17}$$

Je-li $V = f(x)$ (funkce jedné nezávisle proměnn0), pak

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

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

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

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

a odtud

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

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

a

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

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

Pro součin $V = x.y$ dostaneme

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

nebo relativní chybu součinu

$$\delta(V) = \sqrt{\delta^2(x) + \delta^2(x)}\,. \tag{23}$$

Vidíme, že relativní chyba součinu je vyjádřena podobným vztahem, jako absolutní chyba součtu. Snadno se odvodí podobné vztahy i pro chybu součinu $V = x.y.z$ a podílu $V = \frac{x}{y}$. (Odvoďte sami, obojí i pro relativní chyby).

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

Mikrometrem byl změřen průměr d válečku: d = (2,442 ± 0,004) cm, posuvným měřítkem výška h válečku: h = (4,56 ± 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± 0,002) cm.

Dosadíme do vzorce $V = \pi.r^2.h$:

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

Protože průměr je měřen na čtyři místa, výška na tři, počítáme objem zkráceně na čtyři místa. Objem V = 21,25 cm^3.

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

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

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

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:

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

$$= \pm 21{,}25.\sqrt{0{,}000011 + 0{,}000005} = 21{,}25.\sqrt{0{,}000016} =$$
$$= \pm 21{,}25.0{,}004 \doteq 0{,}08\,.$$

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

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

Příklad 2: Určete chybu objemu koule vypočítaného 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 Δ_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 $b_0,...b_{p-1}$. Máme-li soustavou bodů $[x_i,y_i]$ proložit křivku $y=f(x;b_0,..., b_{p-1})$, musíme určit (statisticky odhadnout) neznámé parametry $b_0,..., b_{p-1}$, které se vyskytují v rovnici křivky. Při tom vyžadujeme, aby se křivka co nejvíc přiblížila blížila bodům $[x_i,y_i]$. Statistický odhad parametru b_i označme β_i. Způsob odhadu β_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 kritétium přiblížení považuje suma čtverců hodnot $y_i=f(x_i; \beta_0, ..., \beta_{p-1})$ a odhadem parametrů $\beta_0, ..., \beta_{p-1}$ jsou pak hodnoty, které tento součet čtverců minimalizují.
Označíme-li

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

budou odhady β_i určeny z podmínky

$$S = \min. \tag{25}$$

Touto podmínkou je vyjádřen princip metody nejmenších čtverců. O křivce $y=f(x; \beta_0,..., \beta_{p-1})$ říkáme, že byla body $[x_i, y_i]$ proložena metodou nejmenších čtverců.
Nejčastěji se setkáme s případem, kdy je očekávaná závislost lineární

$$y=b_0+b_1x. \tag{26}$$

Chceme tedy nalézt parametry β_0 a β_1 tak, aby co nejlépe odpovídaly zadaným bodům.
Podle (24 a 25) můžeme odhady β_0 a β_1 určit z podmínky

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

Hodnoty parametrů β_0 a β_1, které minimalizují sumu čtverců odchylek S,

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

dostaneme soustavu dvou rovnic

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

a

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

Jejím řešením získáme odhady β_0 a β_1, parametrů b_0 a b_1:

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

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

Podobně lze nalézt odhady parametrů i pro jiné (složitější) regresní funkce.
Bez odvození napíšeme odhady směrodatných odchylek σ_{β_0} a σ_{β_1} parametrů β_0 a β_1.
Označme:

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

$$s = \sqrt{\frac{S_0}{n-2}}, \tag{34}$$

směrodatné odchylky parametrů β_0 a β_1 pak vypočítáme ze vztahů:

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

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

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 β_0 rovna 0 (je-li *b* PRAVDA nebo 1, hodnota β_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:

β_1	β_0
σ_{β_1}	σ_{β_0}
r^2	σ
F	počet stupňů volnosti
ss_{reg}	ss_{resid}

kde β_0 a β_1 jsou odhady parametrů parametrů b_0 a b_1 z rovnice (26), σ_{β_0} a σ_{β_1} jsou jejich směrodatné odchylky, r^2 je koeficient determinace, σ 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:

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

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

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:

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

Zaměřme 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. Kladná hodnota 0,43 indikuje, že s rostoucím x roste y, záporná 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

$$t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}\,. \tag{40}$$

Je-li testovací veličina t větší než hodnota Studentova rozdělení na dané hladině významnosti α, s příslušným počtem stupňů volnosti $t_{1-\alpha/2}(n\text{-}2)$, můžeme zamítnout nulovou hypotézu r=0,00.

Tabulka: Kvantily Studentova rozdělení pro k=n-2 stupňů volnosti

k	1	2	3	4	5	6	7	8	9	10	12	15	20	30	∞
α=0,05	12,71	4,30	3,18	2,78	2,57	2,45	2,37	2,31	2,26	2,23	2,18	2,13	2,09	2,04	1,96
α=0,01	63,66	9,93	5,84	4,60	4,03	3,71	3,50	3,36	3,25	3,17	3,06	2,95	2,85	2,75	2,58

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	194	100	37636	1940
2	20	389	400	151321	7780
3	30	332	900	110224	9960
4	40	466	1600	217156	18640
5	50	483	2500	233289	24150
6	60	618	3600	381924	37080
7	70	591	4900	349281	41370
8	80	674	6400	454276	53920
9	90	742	8100	550564	66780
10	100	900	10000	810000	90000
Σ	**550**	**5389**	**38500**	**3295671**	**351620**

V posledním řádku tabulky máme všechny veličiny potřebné pro vyčíslení vztahů (31-36). Dosazením do vztahu (32) vypočítáme:

$$\beta_1 = \frac{10.351620 - 550.5389}{10.38500 - 550^2} = \frac{48770}{82500} = 6{,}694$$

a dosazením do vztahu (31):

$$\beta_0 = \frac{1}{10}\left(5389 - 6{,}694.550\right) = 170{,}7\,.$$

Tím jsme vypočítali odhady parametrů β_0 a β_1. Nyní odhadneme jejich směrodatné odchylky. Ze vztahu (33) vypočteme $S_0 = 21866$ a pak ze vztahu (34) $s = 52{,}28$. Veličinu s dosadíme do vztahů (35) a (36) a dostaneme $\sigma_{\beta_0} = 35{,}71$ a $\sigma_{\beta_1} = 0{,}58$. Vypočítali jsme tedy odhady parametrů regresní rovnice $\beta_0 = \left(170 \pm 40\right)$ a $\beta_1 = 6{,}74 \pm 0{,}6$.

Ze vztahu (36) vypočítáme hodnotu korelačního koeficientu

$$r = \frac{10.351620 - 550.5389}{\sqrt{\left[10.38500 - 550^2\right]\left[10.3295671 - 5389^2\right]}} = \frac{552250}{568348} = 0{,}972$$

Na závěr otestujme pomocí vztahu (40) hodnotu r. Vypočítáme parametr t:

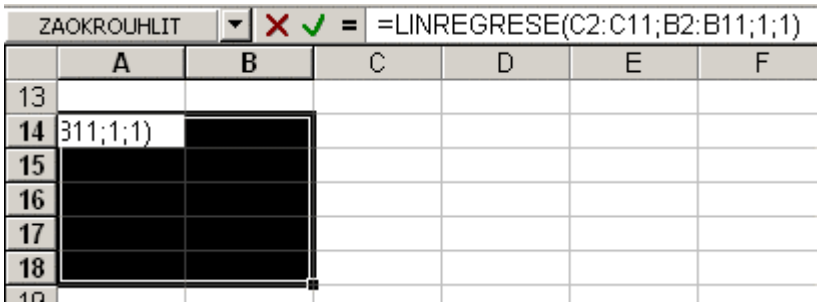
$$t = \sqrt{8}\,\frac{0{,}972}{\sqrt{1 - 0{,}972^2}} = 11{,}63$$

Srovnáním s tabulkou kvantilů Studentova rozdělení pro k=8 a α=0,05 resp. α=0,01, které jsou 2,31 resp. 3,36 vidíme, že tato hodnota značně převyšuje kritickou hodnotu Studentova rozdělení. Můžeme tedy zamítnout nulovou hypotézu a pokládat vliv proměnné x na proměnnou y za prokázaný. Zdůrazněme, že tento test nám pouze potvrdil korelaci mezi proměnnými x a y. Podobnou informaci můžeme získat i z dříve vypočítané hodnoty $\beta_1 = 6{,}74 \pm 0{,}6$. Protože je hodnota směrnice přímky větší než trojnásobek její chyby, můžeme opět tvrdit, že směrnice β_1>0 a tedy existuje korelace mezi proměnnými x a y.

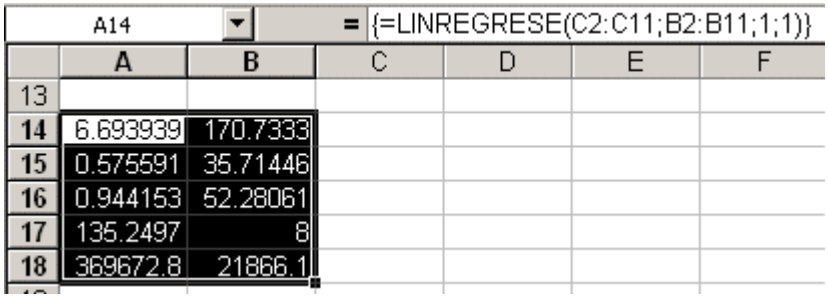
Výpočet v Excelu
Vstupní hodnoty jsou uloženy v listu Excelu:

	A	B	C
1	i	x	y
2	1	10	194
3	2	20	389
4	3	30	332
5	4	40	466
6	5	50	483
7	6	60	618
8	7	70	591
9	8	80	674
10	9	90	742
11	10	100	900
12	Σ	550	5389

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



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



Ze zvolené oblasti nás nejvíc zajímají první tři řádky. V prvním řádku jsou odhady regresních parametrů, ve druhém pak jejich směrodatné odchylky. V prvním sloupci třetího řádku je pak druhá mocnina korelačního koeficientu r^2. Získali jsme tedy všechny důležité parametry regresní přímky.
Pokud se zobrazilo číslo jen v jedné buňce vybrané oblasti nebo je v některých buňkách chybové hlášení #HODNOTA! případně ########, nevložili jste pravděpodobně vzorec klávesami Ctrl+Shift+Enter, ale jen Enter nebo rozkopírováním jedné buňky.