12. Statistické zpracování experimentálních výsledků

Přesnou hodnotu měřené veličiny lze získat pouze dokonalým a přesným měřením. To však není možné, protože každé měření podléhá chybám. Obvykle předpokládáme, že přesná hodnota zvolené veličiny se rovná tzv. **nejlepšímu odhadu**, který získáme statistickým zpracováním dat. Těsnost shody výsledku měření s přesnou hodnotou je **experimentální přesnost** měření. Číselným vyjádřením přesnosti měření je její **chyba**, která se může obecně skládat ze dvou složek:

- **Náhodná chyba** je výsledkem náhodných jevů působících v době měření, které nelze předvídat, opakovat ani eliminovat. Pozorujeme tedy rozptyl výsledků.
- **Soustavná chyba** je způsobena stálým nebo neustále se měnícím příspěvkem k hodnotě měřené veličiny, který lze zcela nebo částečně korigovat (např. novou metodou měření, kalibrací, použitím standardů, apod.).

Odlehlý výsledek (hrubá chyba) je extrémním případem chyby způsobené selháním člověka nebo přístroje. Odlehlý výsledek je třeba vyloučit z experimentálních dat buď přímo, nebo po testu na odlehlé hodnoty.

Experimentální přesnost měření je charakterizována **průměrem** a **nejistotou**. Statistická teorie předpokládá, že tyto hodnoty lze zlepšit opakováním měření. Datový soubor od A_1 do A_n získáme, když n – krát zopakujeme měření veličiny A .

Aritmetický[1] **průměr**

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n} A_i \qquad (1.)$$

je obvykle považován za nejlepší odhad měřené veličiny A . Čím větší n, tím je nejlepší odhad přesné hodnoty - tedy průměrná hodnota $\bar{A}$ (při absenci systematické a hrubé chyby) blíže hodnotě A .

Výhodné je zavést odchylku každého měření od průměru

$$\varepsilon_i = A_i - \bar{A} \qquad (2.)$$

pak nejlepší odhad **nejistoty** je funkcí **směrodatné odchylky** definované vztahem:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(\varepsilon_i)^2}{n-1}} \qquad (3.)$$

kde výraz ve jmenovateli $n - 1 = \nu$ je tzv. počet stupňů volnosti.

Platí, že aritmetický průměr $\bar{A}$ je tzv. bodovým odhadem přesné hodnoty A , zatímco směrodatná odchylka s je tzv. intervalový odhad nejistoty A . Provedeme-li $n = \infty$ měření (za této situace nazýváme směrodatnou odchylku často **standardní nejistotou**) získáme dle vztahu (3.) interval $\bar{A} \pm s$ do kterého nové měření A_{n+1}, pokud se jedná o **normální (Gaussovo) rozdělení chyb,** spadne s pravděpodobností 68.27%. Rozšíříme-li interval na $\bar{A} \pm 2s$ (resp. $\bar{A} \pm 3s$) bude nové měření A_{n+1} v tomto intervalu ležet s pravděpodobností blízkou 95%, kdy **hladina spolehlivosti** je $\alpha = 0{,}05$ (resp. 99.9%, kdy $\alpha = 0{,}001$). Zde srovnej s hodnotami $1{,}960$ (resp. 3,090) pro $n = \infty$ (viz **TABULKA I**). Pokud bychom opakovali celé měření A_1 až A_n opět pro $n = \infty$, budou nově vyhodnocené hodnoty $\bar{A}$ i s stejné a obě hodnoty $\bar{A}$ rovny nejlepšímu odhadu správné hodnoty.

[1] V některých případech se k získání nejlepšího odhadu přesné hodnoty používají jiné způsoby. Např. geometrický, kvadratický nebo harmonický průměr. V sociálních vědách se používá často místo průměru medián a místo směrodatné odchylky percentily.

Pokud je počet provedených měření n malý, se používá nejčastěji **Studentovo rozdělení**[2]**.** K dosažení stejné pravděpodobnosti na dané hladině spolehlivosti α musíme interval rozšířit násobením směrodatné odchylky s koeficientem t, který uvádí **TABULKA I** pro $\nu = n - 1$. Tímto zavádíme **rozšířenou nejistotu:**

$$\bar{A} \pm t \,.s \tag{4.}$$

a **interval spolehlivosti** (*eng.* **confidence interval**):

$$\bar{A} \pm t \,. \frac{s}{\sqrt{n}} \tag{5.}$$

Pro malý počet měření nejlepším odhadem správné hodnoty není $\bar{A}$, ale správná hodnota leží v intervalu spolehlivosti (5.) s pravděpodobností dle zvolené hladiny spolehlivosti α.

Cílem experimentu může být také získat hodnotu y, která je funkcí provedeného experimentálního měření. Například v případě, kdy $y = f(x_1, x_2)$ je funkcí dvou nezávislých experimentálních vstupů $x_1 \pm s_{x_1}$ a $x_2 \pm s_{x_2}$ Tehdy je třeba věnovat pozornost **šíření chyb.** Zjednodušený konzervativní odhad standardní nejistoty výsledku s_f zjistíme v případě, že se chyby obou měření nemohou kompenzovat, ze standardních nejistot s_{x_1} a s_{x_2} dle vztahu:

$$s_f = \sqrt{\left(\frac{\partial f}{\partial x_1} s_{x_1}\right)^2 + \left(\frac{\partial f}{\partial x_2} s_{x_2}\right)^2} \tag{6.}$$

kde $\frac{\partial f}{\partial x_1}$ a $\frac{\partial f}{\partial x_2}\frac{\partial f}{\partial y}$ jsou parciální derivace funkce $F = f(x_1, x_2)$. Tuto standardní nejistotou výsledku s_f pak můžeme rozšířit dle vztahu (4.) a (5.). Obvykle na $\alpha = 0{,}05$.

Vzhledem k možnostem dnešní výpočetní techniky se obvykle získávají výsledky s nadbytečně vysokým počtem číslic. **Počet platných číslic v zápisu výsledku** však musí být v souladu s odhadem nejistoty. Toho docílíme zaokrouhlením výsledku s ohledem na chybu měření.

TABULKA I: Kritické hodnoty t-rozdělení, kde ν je počet stupňů volnosti a α je hladina spolehlivosti (s údajem o intervalové spolehlivosti měření v procentech).

ν	$\alpha = 0.05$ (95%)	$\alpha = 0.001$ (99,9%)	ν	$\alpha = 0.05$ (95%)	$\alpha = 0.001$ (99,9%)
2	4,303	22,326	12	2,179	3,930
3	3,182	10,213	13	2,160	3,852
4	2,776	7,173	14	2,145	3,787
5	2,571	5,893	15	2,131	3,733
6	2,447	5,208	16	2,120	3,686
7	2,369	4,785	17	2,110	3,646
8	2,306	4,501	18	2,101	3,610
9	2,262	4,297	19	2,093	3,579
10	2,228	4,144	20	2,086	3,552
11	2,201	4,025	∞	1,960	3,090

[2] Pro velmi malé počty měření $n < 10$ je alternativou rozdělení Dean-Dixon. (Q-test, analytická chemie).

Výsledky experimentálního měření zapisujeme ve vědeckých výstupech s uvedením intervalové spolehlivosti vyjádřené procenty nebo hodnoty α. Například pokud máme 7 experimentálních měření a zvolíme $\alpha = 0{,}05$ můžeme v textu použít formulace:

- „latentní teplo je s 95% spolehlivostí v intervalu $\Delta H = (31{,}52 \pm 0{,}65)\ kJ/mol$“. Interval spolehlivosti zde byl vypočítán pomocí Studentova rozdělení ($s = 0{,}70, \nu = 6,\ t = 2.447$).
- „průměrné latentní teplo ze 7 měření je $\Delta H = 31{,}52\ kJ/mol$ se směrodatnou odchylkou $s = 0{,}70$ g“.

Mimo odbornou komunitu obvykle postačuje uvést výsledek experimentálního měření s použitím relativní chyby měření v procentech:

- „latentní teplo je $\Delta H = 31{,}52\ kJ/mol$ s relativní chybou měření $0{,}49\%$“. Tím, že jsme použili rozšířené Studentovo rozdělení ($s = 0{,}70,\ \nu = 6,\ t = 2.447$) čtenáře nezatěžujeme.

Lineární a nelineární regrese

V praxi se setkáváme s případy, kdy závislost měřené veličiny $y = f(x)$ obsahuje jeden nebo více parametrů. Častým případem je opakované naměření, kdy získáváme dvojice experimentálních hodnot x_i a y_i. Hodnoty mohou být svázány lineární závislostí, pak platí že:

$$y_i = a + b \cdot x_i + \delta_i \qquad (7.)$$

nebo nelineární funkcí:

$$y_i = f(x_i, a_1, \ldots a_m) + \delta_i \qquad (8.)$$

kde a a b (respektive $a_1, \ldots a_m$) jsou parametry lineární (resp. nelineární) závislosti a δ_i jsou tzv. rezidua.

Parametry získáváme regresí experimentální závislosti. K regresi lze s výhodou použít vhodné software. Jednou z možností je použití SW nástroje „Řešitel“, která je volitelnou součástí softwaru MS EXCEL.

Do úvodních buněk SW Excel vložíme odhad parametrů. Vytvoříme si tabulku se sloupci hodnot x_i a y_i. Do buněk dalšího sloupce zadáme instrukce k výpočtu hodnot $Y_i = a + b \cdot x_i$ (respektive $Y_i = f(x_i, a_1, \ldots a_m)$) s použitím odhadů výchozích parametrů a experimentálních hodnot x_i. V dalších sloupcích vypočítáme rezidua δ_i mezi experimentálními hodnotami y_i a hodnotami vypočtenými z odhadů parametrů Y_i. Do závěrečného sloupce doplníme hodnoty $(\delta_i)^2$ a nakonec vypočteme hodnotu účelové funkce $MF = \sum_{i=1}^{n} (\delta_i)^2$. Nástroj „řešitel“, pak použijeme pro minimalizaci účelové funkce MF a získání optimalizovaných hodnot parametrů lineární (resp. nelineární) závislosti. Tuto metodu nazýváme s ohledem na tvar funkce MF metodou nejmenších čtverců.

Vzhledem k tomu, že měření podléhají chybám, je nutné vyhodnocovat větší počet experimentálních dvojic x_i a y_i, minimálně tři na každý optimalizovaný parametr.

Výpočet parametrů lineární závislosti a test jejich významnosti

Alternativou k numerické lineární regresi metodou nejmenších čtverců je analytický výpočet regresních parametrů a a b. Metoda je vhodná i pro posouzení statistické významnosti parametrů regresní přímky.

1. Pro hodnoty x_i a y_i vypočítáme aritmetické průměry $\bar{x}$ a $\bar{y}$ (nezaokrouhlujeme!).

2. Hodnoty x_i a y_i centrujeme, tj. získáme odchylky: $x_{ic} = x_i - \bar{x}$, $y_{ic} = y_i - \bar{y}$.

3. Vypočítáme hodnot sum $\sum(x_{ic})^2$, $\sum(y_{ic})^2$, $\sum(x_{ic} \cdot y_{ic})$. Parametry a a b pak spočítáme dle následných vztahů:

$$b = \frac{\sum(x_{ic} \cdot y_{ic})}{\sum(x_{ic})^2} \quad (9.)$$

$$a = \bar{y} - b \cdot \bar{x} \quad (10.)$$

5. Ze získaných parametrů a a b a z nezávisle proměnných hodnot x_i vypočítáme vyrovnané hodnoty Y_i a konečné rezidua δ_i jako rozdíly hodnot naměřenýchy_i a vyrovnaných Y_i:

$$Y_i = a + b \cdot x_i \quad (11.)$$

$$\delta_i = y_i - Y_i \quad (12.)$$

6. Úspěšnost regresního modelu se testuje pomocí **standardní odchylky regrese** s_R a pomocí **korelačního koeficientu** r.

$$s_R = \sqrt{\frac{\sum(y_i - Y_i)^2}{n-2}} \quad (13.)$$

$$r = \frac{\sum(x_{ic} \cdot y_{ic})}{\sqrt{\sum(x_{ic})^2 \cdot \sum(y_{ic})^2}} \quad (14.)$$

$$t^r = r \cdot \sqrt{\frac{n-2}{1-r^2}} \quad (15.)$$

7. Regrese je tím lepší, čím je r^2 bližší číslu *1* pro exaktnost se však musí testovat ve vztahu k počtu experimentálních dat. Obvykle očekáváme, že bude významný alespoň na *0.1%*-ní hladině spolehlivosti, t.j. $\alpha = 0.001$. Je-li hodnota t^r vypočítaná z hodnoty r větší než kritická hodnota *t*-rozdělení pro odpovídající počet stupňů volnosti tj. $\nu = n - 2$, lze z *99.9 %*-ní pravděpodobností usoudit, že odpovídající lineární závislost není náhodná.

8. Ze směrodatné odchylky regrese s_R vypočítáme směrodatné odchylky parametrů a a b

$$s_a = s_R \cdot \sqrt{\frac{1}{n} + \frac{{s_R}^2}{\sum(x_{ic})^2}} \quad (16.)$$

$$s_b = s_R \cdot \sqrt{\frac{1}{\sum(x_{ic})^2}} \quad (17.)$$

9. Otestujeme významnost parametrů a a b t-testem dle Studentova rozdělení. Cílem může být rozhodnout, zda lze lineární závislost zjednodušit na prostou konstantní funkci $y = a$ (Tj. ověřit hypotézu I: „správná hodnota $B = 0$“[3]) nebo na prostou lineární funkci $y = b \cdot x$ (Tj. hypotéza II: „správná hodnota $A = 0$“[4]).

Pokud platí:

$|b - B| = |b| < s_b . t_{95\%}$

pak platí hypotéza I a jedná se s pravděpodobností 95% o konstantní funkci $y = a$. Je-li platné:

$$|a - A| = |a| < s_a . t_{95\%} \quad (18.)$$

platí hypotéza II a funkce je s pravděpodobností 95% prostá lineární $y = b \cdot x$.

---

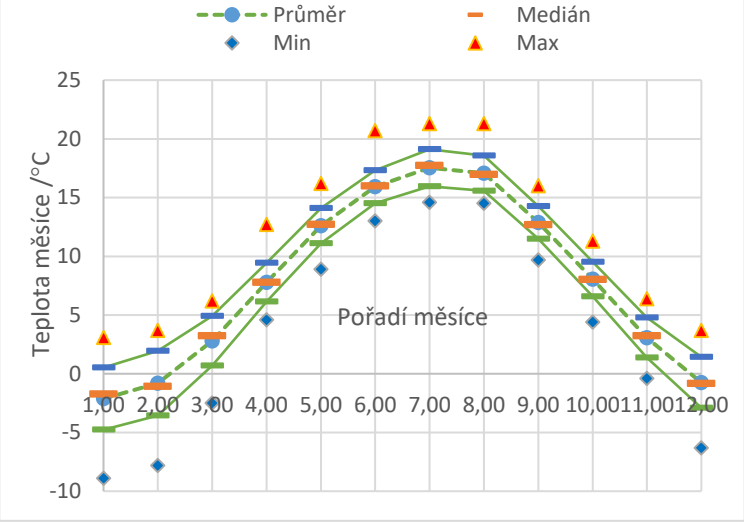
[3] Častý případ u adsorpčních kalibračních křivek.

[4] Viz případ posouzení klimatické změny.

Grafické znázornění experimentální závislosti

Hodnoty, které v experimentu měníme, vynášíme na osu x. Naměřené hodnoty na osu y. Obvykle provádíme experiment tak, že nejistoty x_i jsou zanedbatelné.

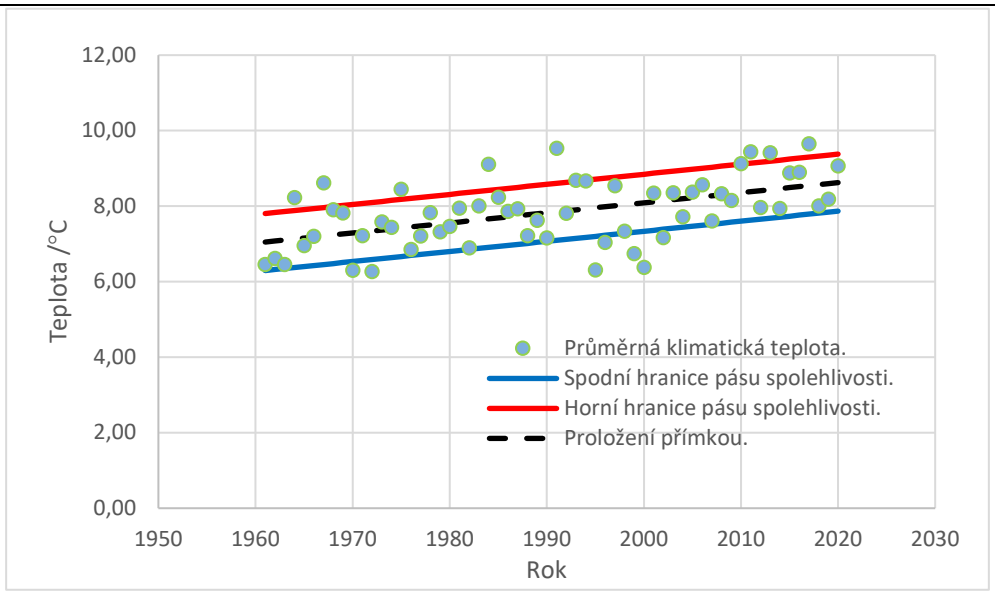
Nejistotu hodnot y_i vypočteme dle vztahů v úvodní části kapitoly 12, kde A_i nahradíme experimentálními hodnotami y_i. Nejistoty naznačíme úsečkami ve směru osy y nebo pásem kolem odpovídajících průměrných hodnot (viz **Obr. 1**).



Obr. 1: Průměrná teplota měsíce bez uvážení klimatické změny, doplněna o další statistické údaje. Pás spolehlivosti je pro $t_{95\%} = 1{,}960$ ($\alpha = 0{,}05$).

V případě lineární regrese experimentální závislosti y na x vynášíme experimentální body $[y_i,\ x_i]$, které můžeme doplnit o pás spolehlivosti (viz **Obr. 2**). Pás získáme tak, že pro každou hodnotu x_i vypočítáme s_R dle vztahů (13.). Po vynásobení s_R odpovídající kritickou hodnotou t –rozdělení vyneseme získaný výsledek (viz **Obr. 2**).

Výsledný graf musí být lehce srozumitelný. Při tvorbě grafu nesmíme zapomenout na vhodnou velikost symbolů a os, správné značení hodnot a jednotek na osách, rozlišení experimentálních a teoretických bodů za pomoci legendy a titulek grafu s jasným popisem.



Obr. 2: Klimatická změna průměrné roční teploty proložená rovnicí (18.) s pásem spolehlivosti pro 95% ($\alpha = 0{,}05$) ($t_{95\%} = 2{,}002$).

12.a. Statistické vyhodnocení změn klimatu

Klimatická data podléhají přirozeným fluktuacím, avšak od počátku průmyslové revoluce jsou negativně ovlivněna člověkem. Jedním z nejdůležitějších parametrů změn klimatu je průměrná teplota či množství srážek. O tom, že ke změnám klimatu dochází, se můžeme přesvědčit statistickým zpracováním měření klimatických dat v čase t.

V kratším období (nízké desítky roků) je možné pro změnu klimatu, např. pro průměrnou roční teplotu T_a, předpokládat lineární funkci:

$$T_a = a + b \cdot t \qquad (19.)$$

Avšak pro údaje sledované zpětně k počátku průmyslové revoluce (parní stroj: 1765) vykazují některé parametry klimatu nelineární růst (pro obsah CO_2 platí exponenciála!).

ÚKOL: Statisticky vyhodnoťte data klimatického pozorování v místě svého bydliště. Použijte například data, které uvádí Portál ČHMÚ : Historická data : Počasí : Základní informace (chmi.cz) a zpracujte jeho data v období od roku 1961. Rozhodněte, zda klimatická změna je statisticky významná či nikoliv na hladině 95%.

POTŘEBY: tabulkový procesor (například MS EXCEL s SW nástroji „Řešitel“ a „Analytické nástroje“ (*eng.: Solver, Analysis Toolpack*).

Seznamte se s jednoduchým statistickým zpracováním dat v kapitole 12. Vytvořte si soubor klimatických dat měsíčních a průměrných ročních teplot ve zvoleném regionu v období nejméně 60 roků. Předpokládejte lineární trend změny klimatu ve tvaru (19.).

1. **VÝPOČET PRŮMĚRNÉ TEPLOTY VE ZVOLENÉM MĚSÍCI BEZ UVÁŽENÍ KLIMATICKÉ ZMĚNY A BEZ POUŽITÍ STATISTICKÝCH FUNKCÍ.**
   - Vyberte například leden a spočtěte jeho průměrnou teplotu jako aritmetický průměr za celé klimatické období.
   - Spočtěte pro vybraný měsíc: maximum, minimum, medián, standardní odchylku, interval spolehlivosti pro $\alpha = 0.05$ a $\alpha = 0.001$, percentil 25 % a 75%.
   - Vyhodnoťte ostatní měsíce
2. **LINEÁRNÍ REGRESE KLIMATICKÉ ZMĚNY BEZ POUŽITÍ STATISTICKÝCH FUNKCÍ.**
   - Vytvořte graf závislosti průměrné roční teploty T_a na roku měření.
   - Vypočítejte parametry a a b rovnice klimatické změny (19.) a její statistiku dle vztahů v kapitole 12.
3. **PŘEPOČÍTEJTE VŠECHNA ZÍSKANÁ STATISTICKÁ DATA** za pomoci statistických nástrojů software MS EXCEL.

VYHODNOCENÍ. K vyhodnocení statistických dat (medián, percentily,…) zvoleného měsíce v roce bez uvážení klimatické změny a bez použití statistických funkcí je vhodné si nejprve seřadit data od nejnižší pozorované teploty v sledovaném měsíci k hodnotě nejvyšší.

PROTOKOL: Zdroj klimatických dat. Klimatické období, region. Použitá verze MS EXCEL. **Tabulka 1:** Dlouhodobá data klimatického pozorování: na řádcích: rok pozorování, ve sloupcích kalendářní měsíc, průměrná roční teplota, následovaná sloupci pomocných dat (např. odchylky δ_i, atd.). Pod tabulkou aritmetický průměr teploty měsíce, min. a max. hodnota, medián, percentily 0,25 a 0,75. **Tabulka 2:** Pro zvolený měsíc (například leden): ve sloupcích: název počítané hodnoty, data získaná bez použití statistických funkcí, data získaná za pomoci statistických funkcí SW EXCEL, název statistické funkce vč. argumentů. Na řádcích pak: hodnoty: n, ν, průměrná teplota měsíce, hodnota $\sum_{i=1}^{n}(\varepsilon_i)^2$, směrodatná odchylka s, hodnoty

pravděpodobnosti t oboustranného Studentova rozdělení pro $\nu = n - 1$, $\alpha = 0{,}05$ a $\alpha = 0{,}001$, interval spolehlivosti pro $\alpha = 0{,}05$ a $\alpha = 0{,}001$, medián, max. a min. hodnota, percentily (25%, 75%). **Společný graf 1:** závislost průměrné teploty jednotlivých měsíců v roce bez uvážení klimatické změny včetně symbolů hodnot mediánu, maxima, minima a percentilů. **Graf 2:** závislost průměrné roční teploty na roku pozorování s vyznačením lineárního trendu. **Dále:** parametry a a b rovnice klimatické změny (19.) **Tabulka 3:** pro klimatickou změnu ve sloupcích: název počítané hodnoty, data získaná bez použití statistických funkcí, data získaná za pomoci statistických funkcí SW EXCEL, název statistické funkce vč. argumentů. Na řádcích pak: hodnoty n, ν, $\bar{x}$, $\bar{y}$, suma $\sum(x_{ic})^2$, $\sum(y_{ic})^2$, $\sum(x_{ic} \cdot y_{ic})$, parametr a a b , pak s_R, r^2, s_a, s_b, $s_b.t_{95\%}$, součin $s_b.t_{95\%}$ **Dále:** rozbor testu významnosti parametru b v rovnici klimatické změny (19.). Zvýšení klimatické teploty od r. 1961.