**ENV006 Statistické zpracování environmentálních dat**

**Vize**

**Za půl roku budou kurz opouštět lidé, kteří jsou nadšeni ze zpracování dat. Budou vědět, že zpracování dat je logické a v principu jednoduché, přestože dat může být hodně a jejich zpracovávání časově i matematicky náročné. Budou mít jasnou představu o základních metodách a jejich principech a budou vědět, kde najít řešení metod složitějších. Budou přesvědčeni o nezbytnosti správného zpracování dat a jeho veliké užitečnosti. Budou schopni pochopit výsledky statisticky zpracovaných dat v běžném životě a základní zpracování vědecké.**

**Budou rozumět základním metodám a budou vědět, kde najdou jejich vzorečkyádu a popis pro matematické vypočítání. Nadto, budou schopni potřebné zpracování naklikat v Excelu i bez opužití naprogramovaných funkcí, Originu nebo Statistice a budou vědět, co tyto programy za ně neudělají.**

Osnova

1. Úvod, zpracování dat a jeho užitečnost, seznámení s obsahem kurzu

Charakteristiky měření a výsledků

* + Typy dat
  + Fyzikální velična a její zápis
  + Správnost, přesnost, neurčitost
  + Správný zápis výsledků, počet desetinných míst, operace s platnými číslicemi
  + Frekvenční rozložení a jeho znázornění – histogram, distribuční funkce

1. Základní popisná statistika
   * Charakteristiky středu – průměr a medián
   * Charakteristiky rozptylu – rozptyl (variance), standardní odchylka, (spooled)
   * Limity spolehlivosti
   * Test na chybu
2. Šíření chyb – odvození a příklady
3. Test 1
4. Normální rozložení
   * Populace × vzorek
   * Pravděpodobnost výskytu, kvantily
   * Interval spolehlivosti – jednostranný, dvoustranný, limity spolehlivosti
   * Statistické testování – chyby prvního a druhého řádu
   * Z – statistika
5. Jednovzorkové hypotézy
   * *t*-test
   * Jak publikovat rozptyl okolo středu
   * Odhad nových limitů intervalu spolehlivosti, velikost vzorku za požadovaných parametrů
   * Výpočet síly testu
   * Hypotézy o rozptylu – χ2
6. Test 2
7. Dvouvzorkové hypotézy

* S pooled, rozptyl rozdílů mezi průměry, *t*-test
* Limity spolehlivosti pro Ho je zamítuta, interval spolehlivosti rozdílu
* Ho není zamítnuta – vážený průměr a jeho limit spolehlivosti
* Nestejné rozptyly –Welchovo přibližné t = Behrens-Fisher t’ test
* Neparametrické testy - Wilkoxon-Mann-Whitney pořadový test
* Párové test

1. Vícevzorkové hypotézy a analýza rozptylu - ANOVA, násobné srovnání
   * Jednocestná ANOVA
   * Welchův test při nestejných rozptylech
   * Neparametrická ANOVA (Kruskalův-Wallisův test)
   * Násobé srovnání - Tukeyův test, Tukey-Kramer, Scheffého test násobných kontrastů (*post hoc* test)
2. Test 3
3. Regrese a korelace. Lineární regrese, kalibrační křivka, kalibrační křivka s vnitřním standardem
4. SVD, Target factoral analysis

Prostředky

Hlava, papír, tužka, Microsoft Excel, Origine, Statistica, Matlab, Maple.

**Literatura**

David Harvey: Modern Analytical Chemistry, ISBN 0–07–116953–9 (kapitola 4)

Skoog, Leary: Principles of Instrumental Analysis, Appendix 1

Jerrold H. Zar: Biostatistical Analsis

Paul Mac Berthouex, Linfield C. Brown: Statistics for Environmental Engineers°

<http://www.stat.cmu.edu/~cshalizi/350/>

Zuund Richard, Meier Peter: Statistical Methods in Analytical Chemistry

Jiří G. K. Ševčík: Metodologie měření v analytické chemii

# Hodina - Charakteristiky měření a výsledků

* 1. ***Typy dat***

Data na **racionální** stupnici Kolikrát?

(Ratio scale) Spojitá data

Data na **intervalové** stupnici O kolik?

Data na **ordinální** stupnici Větší, menší?

Diskrétní data

Data na **nominální** stupnici Rovná se?

Racionální stupnice – musí mít konstantní velikost intervalu a skutečnou nulu. Př. délka, počet částic, váha, objem, rychlost, čas, teplota na absolutní stupnici (v Kelvinech) ...

Racionální číslo (Dušek nazývá podílové číslo)– lze vyjádřit jako zlomek – podíl dvou celých čísel.

Kolikrát je špejle delší než zápalka? O kolik je špejle delší než zápalka?

Intervalová stupnice – má konstantní velikost intervalu, ale nemá skutečnou nulu. Př. dvě běžné stupnice teploty – Celsiova a Fahrenheitova.

Jaký je rozdíl mezi 5 °C a 10 °C? Jaký je rozdíl mezi 20 °C a 25 °C? 20 °C není 2 × více než 10 °C.

Absolutní teplota v Kelvinech má nulu s fyzikálním významem.

Ordinální stupnice (*ordinal* – určující pořadí)

Objekt je kratší nebo delší, zabarvení světlejší nebo tmavší. Buňky nebo lidi podle velikosti roztřídím do pěti kategorii – 1, 2, 3, 4, 5. Podobně úspěšnost řešení písemky.

Data by často mohla být zaznamenána na racionální stupnici – tam se mohu ptát o kolik či kolikrát. Ne však na ordinální stupnici. Kdo má na vysvědčení 3 není třikrát horší než ten, kdo má 1. Jen ten kdo má 3 je horší než ten, kdo má 2. Nemohu počítat průměr známek. Př. C2170–známky.

Nominální stupnice (kategorická)

Lidé jsou muži nebo ženy, praváci nebo leváci, barva vlasů lidí je černá, hnědá, zrzavá nebo světlá. Vzorky jsou zmraženy nebo nezmraženy, s nebo bez přidání katalyzátoru. Na louce rostou jahody, maliny a ostružiny. Statické zpracování používané na racionální, intervalové či ordinální stupnici není použitelné na nominální.

Spojitá a diskrétní data

Spojitá – kontinuální proměnná – délka, váha ...

Diskrétní – diskontinulání – počet knedlíků na talíři, počet žiraf v lese, počet zpracovaných vzorků.

Racioální, intervalová a ordinální data mohou být spojitá nebo diskrétní, nominální ze vsé podataty jsou diskrétní.

**Příklad C2170-1-Známky**

* 1. ***Správnost, přesnost a neurčitost***

**Správnost**

* blízkost měření opravdové hodnotě, kterou měříme (accuracy).
* vyjadřuje se jako absolutní chyba průměru, xt je akceptovaná hodnota



* nebo jako relativní chyba průměru.



%, ‰, ppm, ppb, ppt, ppq

**Chyba** – rozdíl mezi jedním měřením a skutečnou hodnotou (error). Je mírou správnosti.

**Chyby (Harvey)**

* hrubé **–** vzniklé nepozorností, poškozením přístroje
* systematické chyby (determinate)



* + chyba vzorkování
  + chyba metody
  + chyba měření – maximální systematická chyba měření – tolerance (maximální chyba měření uváděná výrobcem)
  + chyba pozorovatele
  + konstantní systematická chyba × proporcionální chyba
* náhodné chyby (random = indeterminate = stochastic) -> normální rozložení chyb



**Přesnost**

* vzájemná blízkost opakovaného měření téhož (precision).
* úplně přesné může být jen pozorování, které zahrnuje počítání objektů. Všechna ostatní měření obsahují chyby a dávají jen přiblížení pravdě.

Zar Fig 1.1

* může být vyjádřena jako rozsah (range), standardní odchylka (standard deviation), rozptyl (variance), keoficient variance (coefficient of variantion).
* vyjadřuje odhad náhodné (indeterminate) chyby, která ovlivňuje měření.

Opakovatelnost – přesnost analýz jednoho vzorku jednou metodou (repeatability).

Reprodukovatelnost – přesnost analýz více vzorků, provedené více lidmi, různými metodami (reproducibility).

Reprodukovatelnost nemůže být lepší než opakovatelnost.

**Neurčitost měření**

* vyjadřuje rozsah možných hodnot, které mohou u měření nebo výsledku být očekávány.
* zahrnuje systematické (systematic = determinate) i náhodné (random = indeterminate = stochastic) chyby

**Příklad 4.8 –pipety**

**Příklady tolerancí – Harvey 76**

**Nezávislost měření**

* podmínka pro výpočet průměru, standardní odchylky, ... je opravdová nezávislost měření

Př.1 UV-vis spektroskopie – chyba měření je zanedbatelná oproti chybě vážení a ředění. Pro stanovení molárního absorpčního koeficientu je potřeba tří několika miligramových navážek na lodičce, rozpuštění v odměrné baňce a následné ředění. Dobré měření by mělo vyjít v rozmezí ± 3 %.

Př. 2. V HPLC je separace látek a následná integrace píků zatížena stejnou nebo větší chybou jako vážení a ředění. Proto je běžně akceptovanou praxí dvojí vážení a dvojí nastříknutí jednoho vzorku.

* 1. ***Správný zápis výsledků***

# platných míst (significant figures) – je počet všech jistých míst a první nejisté.

8 znamená 7.50000... až 8.49999... 1 platná číslice

8.3 znamená 8.25000... až 8.34999... 2 platné číslice

8.0 znamená 7.95000...až 8.04999... 2 platné číslice

Vědecký zápis

7.2 × 104, 7.20 × 104

Zaokrouhlování

Operace s platnými číslicemi – hrubé pravidlo

+, - na *n* desetinných míst

\*, / na *n* platných číslic

Doporuční Zar (str. 73) – Uvádět standardní chybu (standardní odchylku průměru) na 2 platné číslice. Pak standardní odchylka a průměr budou uvedeny na stejný počet desetinných míst, rozptyl může být uveden na dvojnásobný počet desetinných míst.

Zuunda (str. 25) standardní odchylka z malého počtu pozorování má význam na jedno či dvě platné číslice.

* 1. ***Frekvenční rozložení***

Frekvenční tabulka – vhodná pro zobrazení velkého množství dat.

Frekvenční rozložení – mezi jednotlivé kategorie, znázorněno pomocí sloupcového grafu (bar graph).

Stejná šířka sloupců, sloupce ukazovat od nuly, o množství vypovídá výška a plocha. Pro nominální data nechávat mezi sloupci mezery.

Histogram – sloupcový graf pro spojitá data.

Frekvenční polygon.

n(x) absolutní četnost x (freqency)

p(x) relativní četnost x; p(x) = n(x) / n (relative frequency)

N(x) kumulativní četnost nepřevyšujících x

F(x) kumulativní relativní četnost hodnot nepřevyšujcících x = distribuční funkce

Kvantily.