Kemometrija v analizni kemiji

- Študijska literatura
- Kemometrijske teme sklopa predavanj
- Uvod
- Osnove statistike z eno spremenljivko

Prof. dr. Jure Zupan

Študijska literatura

1. D.L Massart, B.G.M. Vandengiste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers Verbeke, Handbook of Chemometrics and Qualimetrics: Part A, Elsevier, Amsterdam, 1997.
2. B.G.M. Vandengiste, D.L Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, Amsterdam, 1998.
3. J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, Wiley-VCH, Weinheim, 1999
4. J. Zupan, Računalniške metode za kemike, DZS, Ljubljana, 1992
5. Teach/Me software, Hans Lohninger, Springer Verlag Berlin 1999

Kemometrijske teme razdeljene na sklope

		Sklop
M. Pompe		
1.	Uvod, vsebina, kaj je kemometrija	(0.5)
2.	Opisna statistika (centr. limitni teorem, mediana, mode, μ, σ, kontrolne karte)	(1)
3.	Napovedna statistika (primerjave, t-test, F-test, α in β napaka)	(1)
4.	Kalibracjska premica (vsi parametri kalibracijske premice in napake)	(1)
5.	Meja detekcije, območja potrditve analita, kvantitativna določitev	(0.5)
6.	Analiza variance – ANalysis Of VAriance ANOVA(F-test)	(1)
7.	MLR (polinomski modeli, faktorji, transponirana matrika , inverzna matrika)	(1)
8.	Metrika v prostoru, razdalje, grupiranje (clustering), dendrogrami	(1)
M. Novič		
9.	Eksperimentalni načrti (dvonivojski, večnivojski)	(1)
10.	Metoda glavnih osi - Principal Component Analysis (PCA)	(1)
11.	Validacija modelov (razdelitev na tri nize, leave-one-out, bootstrap)	(1)
12.	Nevronske mreže (elementi nevronskih mrež, EBP, Kohonen, CPR)	(2)
13.	Optimizacijska metoda SIMPLEX	(1)
14.	Optimizacija z genetskim algoritmom (GA)	(1)
15.	Predstavitve kemijskih struktur (deskriptorji, spektru-podobne predstavitve)	(1)
	Skupaj	(15)

Besedo *kemometrija*, ki izhaja iz švedske besede *kimometrie* in njene angleške različice *chemometrics*, je leta 1971 na univerzi v Umei skoval in prvič uporabil švedski kemik in statistik prof. Svante Wold. Že tri leta kasneje, 10. junija 1974 v Seattlu, ZDA, so ob ustanovitvi Kemometrične družbe (*Chemometric Society*) v njenem statutu opredelili področje in namen *kemometrije* takole: *kemometrija* je področje znanosti, ki proučuje *"razvoj in uporabo matematičnih in statističnih metod za ugotavljanje pomembnih kemijskih informacij"*.

Posebnost *kemometerije* ni toliko v raziskavah in pomenu samih metod, kot v tem, da preučuje te metode hkrati s pogledom na kemijske strukture. V kemiji so vsi eksperimentalni podatki in lastnosti spojin vedno povezani s kemijskimi strukturami. Kemijska struktura spojine je težko določljiva lastnost, ker je sestavljena iz cele vrste posamičnih podatkov, ki so že v svojem temeljnem bistvu nedoločljivi. Zato se natančni določitvi kemijske strukture lahko samo približamo z vedno boljšimi eksperimentalnimi in računskimi postopki, povsem eksaktno pa je ne moremo določiti. Tu tiči tudi težava pri iskanju povezav med kemijskimi strukturami in lastnostmi spojin. Za večino navadnih poskusov in raziskav lastnosti spojin zadostujejo sorazmerno enostaven načini opisov struktur.

Kljub temu so tudi enostavni zapisi kemijskih struktur (npr. oznake atomov s simboli elementov in število kemijskih vezi s črticami) težko razumljivi računalnikom in še posebej neprikladni za splošno računanje. Strukturne zapise spojin je vedno težko obdelovati z navadnimi postopki, s katerimi matematiki in statistiki obdelujejo števila in znake. Zato se *kemometrija* ukvarja z razvojem novih metod in njihovim prilagajanjem za vnos kemijskih struktur kot vstopnih podatkov. Hkrati uvaja kemometrija ideje s področij numerične matematike, računalništva, umetne inteligence in informatike. Iskanje in razvoj novih metod za delo s kemijskimi strukturami ter prenašanje in prilagajanje starih že uveljavljenih metod, se odraža tudi v občasnem popravljanju definicije *kemometrije*. Prvotni opredelitvi se namreč dodajajo vedno nova področja iz katerih črpa nove ideje, kot npr.: *računalniške metode*, *umetna inteligenca*, *nevronske mreže* idr.

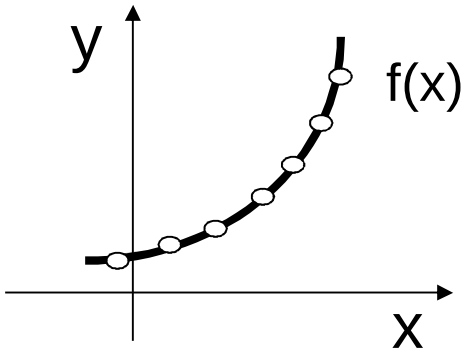
V kemometriji raziskujemo različna, večinoma matematična orodja za obdelavo kemijskih informacij. Za vsako od orodij, ki jih uporabljamo moramo vedeti:

- kako deluje,
- kdaj ga lahko uporabljamo in
- kakšne so njegove prednosti, slabosti, območje delovanja in omejitveni pogoji

Poleg tega, da kemometrija proučuje različne matematične metode, se ukvarja tudi z obdelavo zelo različnih vrst in oblik podatkov, ki so večinoma kemijske informacije in njihovi zapisi. Kot "kemijske" podatke lahko omenimo posamezne fizikalno-kemijske meritve, skupke meritev o neki spojini, različne spektre in kromatograme, zapise kemijskih struktur, procesne vektorje, časovne vrste, alfanumerične podatke, binarne (diskretne) informacije o preiskovanih stanjih itd.

Pri odločitvah o tem, katero metodo bomo uporabili, je oblika podatkov lahko zelo pomembna. Najpomembnejšo obliko podatkov predstavlja njihova kompleksnosti. Glede na kompleksnost podatkov, ki jih različne metode lahko obdelujejo ločimo na take, ki lahko obdelujejo le posamezne spremenljivke ali skalarje (univariate methods) in tiste, ki obdelujejo objekte predstavljene z vektorji, t.j., z več spremenljivkami hkrati, npr. $X = (x_1, x_2, .. x_i ... x_n)$ (multi-variate methods). Seveda so tudi primeri, ko neka metoda deluje na obeh vrstah podatkov. Primer za tako "univerzalno" metodo je linearna regresija, ki jo lahko uporabimo za izračun navadne kalibracijske premice $y = b_0 + b_1 x$ ali pa v bolj kompleksni obliki za izračun regresijskega modela z veliko spremenljivkami hkrati $\boldsymbol{Y = ||M||X}$ (multiple-linear-regression model ali MLR) .

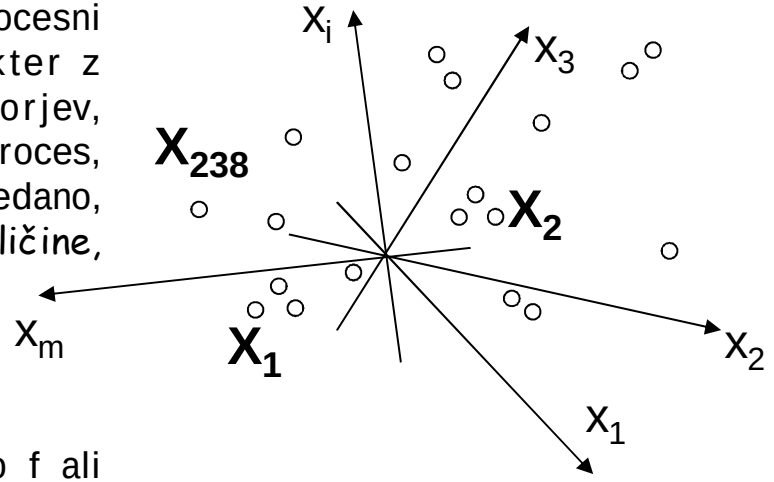
Skalarji ali uni-variatni podatki so posamezne merljive ali **izračunljive količine**. Npr, posamezne koncentracije, temperatura, pH, **povprečje** meritev, standardni odmik ali napaka meritve, vrednost *F* pri ANOVA testu ipd. Metodološko so obdelave odvisnosti ene lastnosti *y* od ene spremenljivke npr. *x*, razmeroma enostavne:



$$y = f(x,p_1,p_2,...p_n)$$

Kjer so p_1, p_2,... p_n parameteri funkcije f

Bolj zapletene so obdelave podatkov, ki so predstavljeni v obliki vektorjev. Vektorji združujejo **več** merljivih **veličin** (spremenljivk) v enotnem zapisu objekta $X = (x_1, x_2, ,,,, x_m)$. V tem primeru je vsak objekt obdelave, kot je npr. spekter, struktura, kompleksna analiza vsebnosti **več** komponent hkrati, procesni vektor, **časovna** vrsta itd., opisan z nizom podatkov. Spekter z nizom intenzitet, struktura z nizom strukturnih deskriptorjev, procesni vektor z nizom vseh spremenljivk pomembnih za proces, ki so izmerjene ob **določenem času** itd. Metodološko gledano, opisujemo lastnosti sistemov, ki so zopet lahko vekorske **količine**, s funkcijami, matrikami ali drugimi kompleksnimi operatorji:



$$\boldsymbol{Y = M(X,P)}$$

P je niz parametrov $(p_1, p_2, ...p_s)$, ki opredeljujejo funkcijo f ali model M. **Določitev** niza parametrov, v **različnih** metodoloških pristopih, s katerimi opisujemo opazovane kemijske sisteme, je ena glavnih nalog kemometrije.

Cilj pouka kemometrije je:

- **izpopolniti védenje, kako najprimerneje obdelovati kemijske podatke (meritve, strukture, spektre, procese, lastnost itd.) da dobimo čim več koristnih in željenih informacij.**
- **podrobno pozanvanje in proučevanje kemometrijskih metod za obdelavo kemijskih informacij.**

Poznavanje	Sposobnost opraviti ali narediti
statistike:	Opisati in/ali napovedi lastnosti vzorcev (meritev) in populacij
vizualizacije:	Podati jasen pregled in predstavitev kompleksnih več-dimenzijskih podatkov v 2d-prostoru
razdelitev:	Izbrati in določiti primerne skupine podatkov za različne namene (učni, testni in kontrolni nizi, referenčni nizi, določitev ubežnikov itd.)
klasifikacije:	Izdelati ali izbrati referenčne skupine objektov in napovedovati predpisane kategorije, ki jim pripadajo neznani vzorci
modeliranja:	Izdelati primerne matematične modele za kvantitativno napovedovanje lastnsti objektov naših raziskav (spojine, spektri, strukture, procesi,...)
optimizacije:	Izbrati in določiti najugodnejših pogoje, spremenljivke, metode in lastnosti pri danih robnih zahtevah in pogojih
računalniške obdelave struktur:	Obvladati najrazličnejše načine zapisov (predvsem uniformnih) kemijskih struktur.

Ena od pomembnih nalog uporabe kemometrijskih metod je zagotavljanje in kontrola kvalitete

Osnove statistike z eno spremenljivko

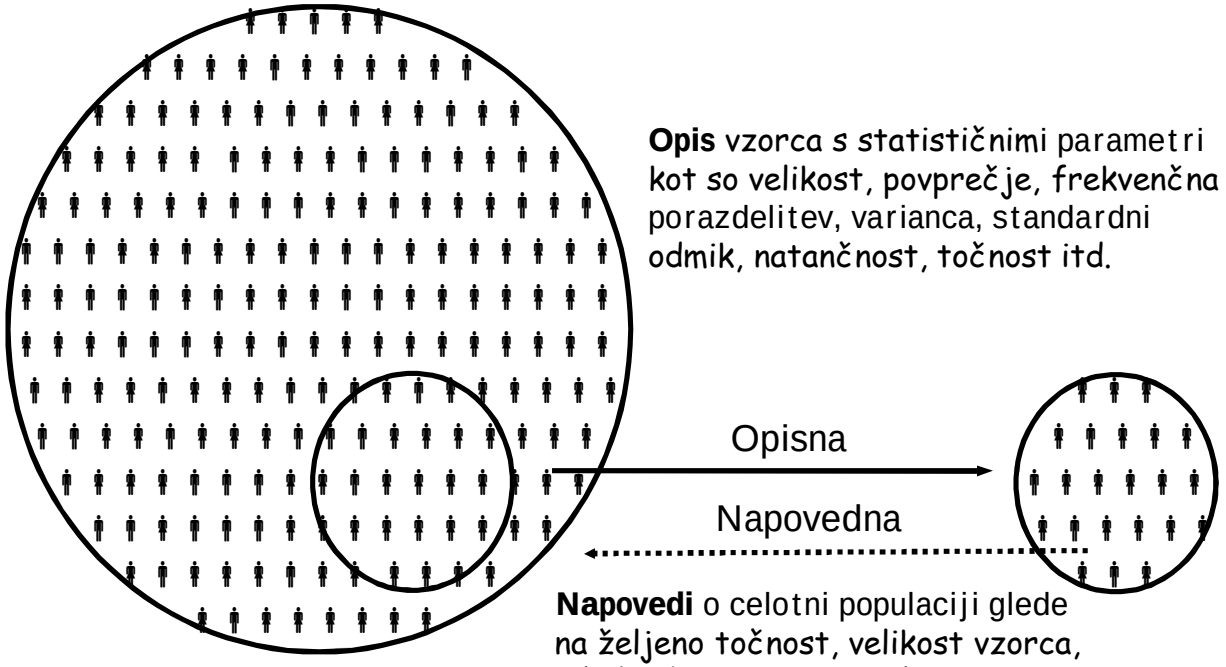
Razlika med opisno in napovedno statistiko

Populacija in vzorec

Statistika je del matematičnih ved, ki preučuje množice objektov, njihove opise in značilnosti ter ugotavlja zveze, ki veljajo med različno velikimi deli skupin istovrstnih objektov. Skupino istovrstnih objektov (kemijske analize izdelkov, živil, zdravil, procesnih postopkov, kemijskih struktur, receptur izdelkov itd.,) imenujemo populacijo. Vsem objektom v populaciji lahko izmerimo eno ali več lastnosti. Numerične vrednosti posameznih lastnosti objektov v populaciji so različno porazdeljene. ***Zelo malo*** lastnosti je porazdeljenih v normalni ali Gaussovi porazdelitvi.

1. primer. Poznamo opis in parametre velike množice. Z merjenjem lastnosti objektov v manjšem vzorcu, želimo ugotoviti ali mali vzorec pripada veliki množici ali ne. (Ugotavljanje kompatibilnosti s standardi).

2. primer. Želimo opisati zelo veliko množico, ki nas zanima, a ne poznamo njenih statističnih parametrov. Zato izberemo manjši vzorec in na njem izmerimo in izračunamo ustrezne vrednosti. (Napoved lastnosti velike množice).



Opis vzorca s statističnimi parametri kot so velikost, povprečje, frekvenčna porazdelitev, varianca, standardni odmik, natančnost, točnost itd.

Napovedi o celotni populaciji glede na željeno točnost, velikost vzorca, ničelno hipotezo in izračunane podatke iz vzorca.

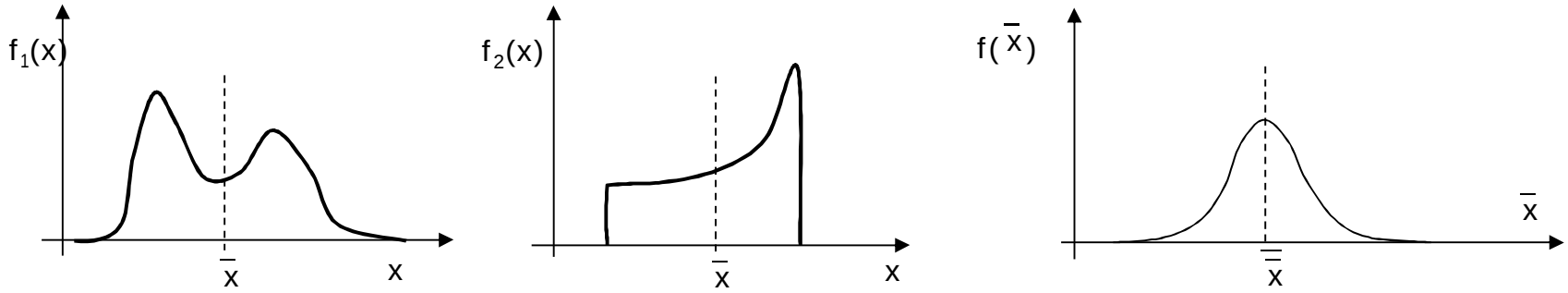
Povprečje, varianca, standardni odmik in napaka povprečja

Osnovni parametri, ki opisujejo vzorce in populacije so:

- povprečje - $$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$
- modus - vrednost, ki jo ima največ objektov (meritev) v vzorcu, $$modus\{x_1, x_2, \ldots, x_i, \ldots, x_m\} = x_i \leftarrow max\{f(x_i)\}$$
- mediana - vrednost, ki razdeli vzorec z *N* objekti (meritvami) na dva številčno enaka dela, ki imata po delitvi ali *N/2* (sodi vzorci) ali *(N-1)/2* objektov (lihi vzorci). $$x_{mediana} = \frac{1}{2}\left[x_{\frac{N}{2}} + x_{\frac{(N+1)}{2}} \right], \quad ali$$ $$x_{mediana} = x_{\frac{N+1}{2}}$$
- varianca - povprečni kvadrat odmikov posameznih vrednosti od središča vzorca. $$v = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{1}{N}(\sum_{i=1}^{N} x_i)^2}{N-1}$$
- standardni odmik - kvadratni koren variance. $$s = \sqrt[2]{v} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}} = \sqrt{\frac{\sum_{i=1}^{N} x_i^2 - \frac{1}{N}(\sum_{i=1}^{N} x_i)^2}{N-1}}$$
- razpon - razlika med najmanjšo in največjo vrednostjo iste lastnosti objektov v populaciji $$Razpon = x_{max} - x_{min}$$
- napaka povprečja – je odvisna od velikosti vzorca (N), medtem ko standardni odmik ni! $$s_{\bar{x}} = \frac{s}{\sqrt{N}}$$

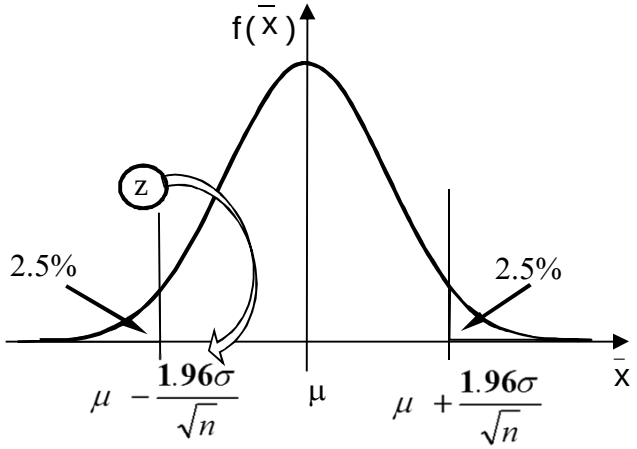
Središčni limitni teorem (central limit theorem)

Porazdelitev povprečnih vrednosti dovolj velikih vzorcev vzetih iz poljubne populacije se približuje normalni (Gaussovi) porazdelitvi. Če je osnovna populacija normalna, je porazdelitev povprečij vzorcev vedno normalna, ne glede na njihovo velikost.



Ne glede na to, kakšno porazdelitev (f_1 ali f_2) ima vrednost *x* v neki populaciji, bo ***porazdelitev povprečij*** (desno) vedno dober približek normalni (Gaussovi) porazdelitvi.

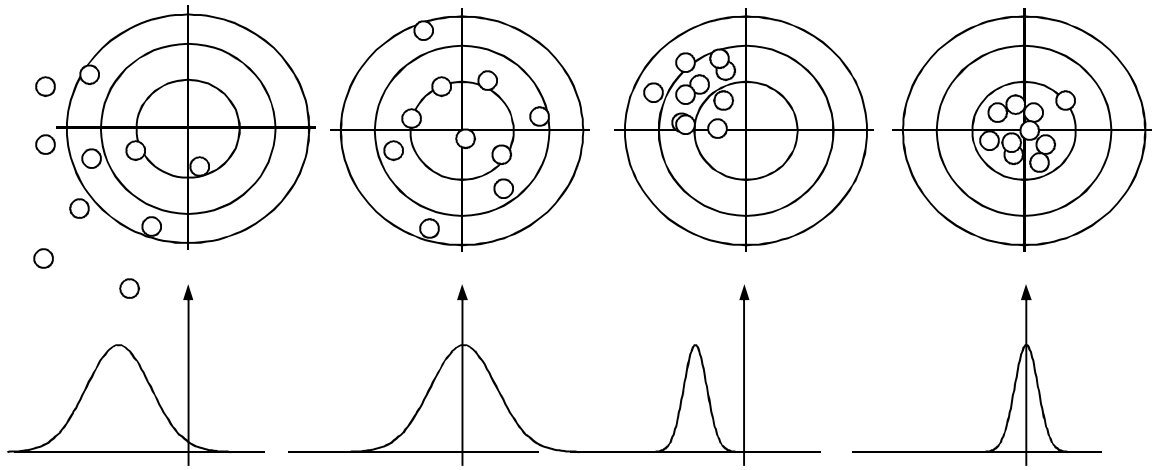
Interval zaupanja za povprečne vrednosti (confidence interval for the mean)



95 % vseh vrednosti, ki so normalno porazdeljene, pade v znotraj intervala $\pm 1.96\sigma$. Mejne vrednosti intervala "z" (n.pr. z=1.96) so tabelirane glede na odstotke površine α, (n.pr.: α=0.05), ki jo vrednost *"z"* omejuje. Vrednost *"z"* pove, koliko standardnih odmikov širok interval potrebujemo, da okrog povprečne vrednosti zajamemo (1-α)% populacije.

Tak interval imenujemo ***interval zaupanja*** in je odvisen od števila meritev s katerim ga določamo. Ker gledamo populacijo povprečnih vrednosti, standardni odmik populacije meritev σ pa je lastnost merske metode, moramo standardni odmik merske metode σ deliti s korenom števila meritev *n*. Pri določitvi širine intervala zaupanja je pomebno, da lahko z večjimi vzorci (večji *n*) zožamo interval zaupanja na sprejemljivo velikost.

Točnost in natančnost (accuracy and precision)



Točnost (accuracy) in natnačnost (precision) meritev. Meritev na skrajni levi a) je nenatančna in netočna, naslednja b) je točna ni pa natančna, meritev c) je natančna, a ni točna in zadnja meritev d) je hkrati točna in natančna.

Točnost (accuracy) meritve je odvisna od razlike med izmerjenim povprečjem vzorca in dejansko povprečno vrednostjo populacije (tarče). Čim večja je ta razlika, tem manjša je točnost. Točnost je pogosto povezana z neznanimi napakami pri meritvah (bias), s slabimi standardi in s pojavi, ki brez naše vednosti vplivajo na meritve.

Natančnost (precision) meritve je neposredno povezana z velikostjo standardnega odmika. Čim večji je standardni odmik, tem slabša je natančnost. Natančnost je povezana z naravo meritve in jo navadno zelo dobro poznamo in tudi določimo. Natančnost povprečja vzorcev lahko popravimo z večanjem števila meritev. Natančnosti same metode pa ne moremo izboljšati.