Testy dobré shody

= Ověření shody rozdělení s pravděpodobnostním modelem.

Umožňují srovnání **empirického** (výběrového) rozdělení s jistým rozdělením **teoretickým (základním)**.

- **χ^2 – test dobré shody**
- **Kolmogorovův - Smirnovův test**
- **Davidův test normality** (SPSS: test normality → Shapiro-Wilkův test)

χ^2 – test dobré shody

- používají se nejčastěji pro ověřování těchto hypotéz:
  - výběr pochází z populace, v níž jsou relativní četnosti jednotlivých variant **roztříděny podle nějakého znaku do k skupin**
  - výběr **pochází z rozdělení určitého typu** (např. normální), jehož parametry jsou dány
- Testuje, zda skutečná četnost odpovídá nějakému předpokladu četnosti
- Používá se pro ověřování hypotéz v kontingenční tabulce
- H0: výběr pochází z populace s normálním rozdělením
- H1: non H0

Nevyhovují-li některé četnosti této podmínce, lze dosáhnout jejího splnění sloučením několika sousedních tříd. Tím ovšem dojde také ke snížení počtu stupňů volnosti, neboť k (počet tříd intervalů) je rovno počtu tříd po sloučení.

Hlavní podmínka použití testu:

- Teoretické četnosti větší než 5 (n) = expected (80 % hodnot)

Kolmogorovův - Smirnovův test

- Používá se k ověření hypotézy, že **pořízený výběr pochází z rozdělení se spojitou distribuční funkcí F(x), která je plně specifikovaná.**
- Známe typ i příslušné parametry rozdělení
- Vychází z původních **napozorovaných hodnot, a nikoliv z údajů setříděných do skupin.**
- Nedochází ke ztrátě informace obsažené ve výběru.
- Test je založen na **porovnání kumulativních četností**

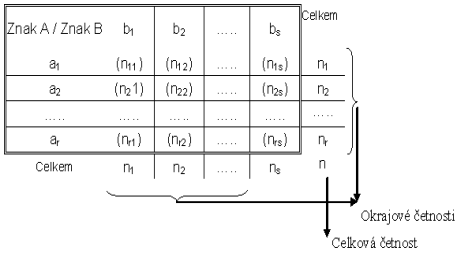
Test je založen na porovnání kumulativních četností.
H0: výběr pochází z populace s … rozdělením
H1: non H0
Nej…kumulativní četnosti empirické
Noj…kumulativní četnosti teoretické
n…rozsah sledovaného souboru
max…je největší rozdíl mezi četnostmi Nej a Noj

Kritický obor je vymezen nerovností.
$D > D_{n(1-\alpha)}$ → H0 zamítáme nulovou hypotézu o shodě mezi empirickým a teoretickým rozdělením na hladině významnosti alfa.

Tabulka kritických hodnot Dα je sestavena pouze pro n <= 40. Pro výběry větších rozsahů se musí kritické hodnoty určit podle vztahů:

Analýza kvalitativních znaků

- Cílem analýzy je:
  - **otestovat závislosti mezi proměnnými**
    - x2 test nezávislosti
    - Exaktní testy: Fisherův faktoriálový test nebo test Monte Carlo
  - **změřit sílu závislosti**
    - koeficienty kontingence
    - koeficienty asociace



Kontingence

- je vztahem dvou či více **kvalitativních statistických znaků**, z nichž alespoň jeden je znakem **množným** (znaky, které mají větší počet obměn – barva očí, stupeň vzdělání)
- vztah mezi těmito znaky je zachycen v **kontingenční tabulce**

Testování závislosti

x^2 – test o nezávislosti

H0: mezi sledovanými znaky A a B neexistuje závislost
H1: znaky A a B jsou závislé

Nij…skutečná četnost
Oij…očekávaná četnosti v i-tém řádku a j-tém sloupci tabulky

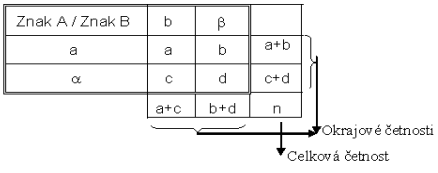
Hodnotu x2 srovnáme s kritickou hodnotou x2-rozdělení o stupních volnosti (r-1)(s-1)
Pokud $x^2 > x^2_\alpha \rightarrow$ H0 se zamítá na hladině významnosti alfa.
Závislost mezi znaky byla prokázána.

Měření síly závislosti
Pearsonův koeficient kontingence
Jsou-li zkoumané znaky nezávislé, je hodnota tohoto koeficientu 0.
Maximální hodnota dosažená při úplné závislosti je však menší než 1 a mění se podle toho, do kolika tříd byly zkoumané znaky rozděleny.
Cramerův koeficient V (Cramerovo V)
pro h=min (r,s)
Čuprovův koeficient kontingence
hodnoty uvedených měr se pohybují v intervalu (0,1)
Normalizovaný koeficient kontingence
Slouží k porovnání síly závislosti mezi několika kontingenčními tabulkami různého rozměru.

Předpoklady použití
x2 – test nelze použít
Je-li více než 20% teoretických četností menších než 5.
Je-li alespoň v jednom políčku kontingenční tabulky očekávaná četnost menší než 1.
V takových případech je možné některé sousedící skupiny sloučit. Sloučení musí být logické, věcně správně a dobře interpretovatelné. Další možnost je použití exaktních testů (např. Fisherův test)

Asociace

- zkoumá vztah dvou kvalitativních proměnných, které jsou nositeli znaků **alternativních** (znaky, které nabývají jen dvou obměn-ano/ne, muž/žena)
- Tento vztah zachycuje speciální typ kontingenční tabulky **2x2** – asociační tabulka, čtyřpolní tabulka
- A, b, c, d → skutečné (empirické četnosti)



Testování závislosti
Tabulka 2x2
x2-test nezávislosti
Pro dostatečně velké rozsahy výběru n > 40, pro 20 < n <= 40 není-li žádná očekávaná četnost menší než 5.
Pro n <= 20 x2-test nelze použít.
V případě, že nejsou splněny podmínky pro použití x2 testu používáme Fisherův faktoriálový test, který vychází z přímého výpočtu pravděpodobností.
Koeficient asociace (Cramerovo V pro r=s=2)
Yuleův koeficient asociace

Regresní a korelační analýza

- Závislost příčinná (kauzální) – jeden jev vyvolává existenci jevu druhého
- Závislost pevná – výskytu jednoho jednu nutně odpovídá výskyt druhého jevu
- Průběh závislosti lze přesně charakterizovat určitou matematickou funkcí
  - Závislost volná – jeden jev podmiňuje jev jiný jen s určitou pravděpodobností a v různé intenzitě. U této závislosti lze charakterizovat teoretický průběh závislosti a její těsnost
  - Závislost statistická – volná závislost týkající se kvantitativních statistických znaků.

- **Regrese** charakterizuje **průběh** závislosti mezi kvantitativními statistickými znaky pomocí matematického modelu (regresní funkce)
- **Korelace** měří **těsnost** (sílu, míru, intenzitu) statistické závislosti mezi kvantitativními statistickými znaky pomocí koeficientů.

Druhy závislosti:
- Podle počtu kvantitativních znaků:
  - závislost jednoduchá – dva znaky (X,Y), na závisle proměnnou Y působí jedna nezávisle proměnná X
  - závislost vícenásobná – více než dva znaky, na závisle proměnnou Y působí více nezávisle proměnných X1, X2
- Podle typu regresní funkce
  - lineární závislost
  - nelineární závislost
- Podle směru změn kvan. znaků
  - závislost pozitivní (kladná, přímá)
  - závislost negativní (záporná, nepřímá)

- V regresní analýze obecně analyzujeme vztah mezi jednou proměnnou zvanou cílová nebo **závislá proměnná (Y)** a několika dalšími, které nazýváme **nezávislé proměnné (X)**
- Závisle proměnná je spojena s nezávisle proměnnými **regresní funkcí**, jež obsahuje několik neznámých parametrů
- Základní úkoly regresní analýzy:
  - získání statistických odhadů neznámých parametrů regresní funkce na základě výběru
  - testování hypotéz o těchto parametrech
  - ověřování předpokladů regresního modelu

KORELAČNÍ ANALÝZA
- **Korelace** obecně označuje míru stupně (sílu) závislosti dvou proměnných
- Říká se, že dvě proměnné jsou korelované, jestliže určité hodnoty jedné proměnné mají tendenci se vyskytovat společně s určitými hodnotami druhé proměnné
- **Měření těsnosti – síly závislosti** – spočívá ve zjištění, jak těsně se jednotlivé skutečně napozorované hodnoty přimykají k regresní čáře, která vystihuje průběh závislosti.

Výběr regresní funkce
- **Logické posouzení daného vztahu** – věcný rozbor vztahů mezi proměnnými (využíváme zkušeností z podobných analýz apod.)
- Vycházíme z grafické analýzy dat – bodového diagramu tzv. **korelačního pole** (nutná znalost průběhu různých typů matematických funkcí)

- Využití matematicko – statistických kritérií - vztah proložíme celou řadou funkcí a jako nejvhodnější zvolíme tu, která má **nejvyšší hodnotu korelace** (lze využít několik matematicko-statistické kritéria)
- Základní model regresní závislosti
  - $y'_i = f(x_i) + e_i$
  - $f(x_i)$ **…** je regresní funkce
  - e_i **…** jsou náhodné (reziduální) chyby (odchylky)

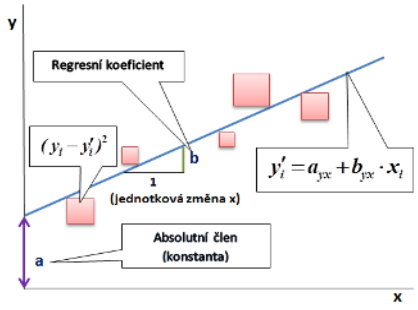
Jednoduchá lineární regrese

Model regresní přímky

- **$y_i = \alpha + \beta x_i + e_i$** **i = 1, 2, …, n**
  - X … nezávisle proměnná (vysvětlující, regresor)
  - Y … závisle proměnná (vysvětlovaná)
  - α,β … neznámé parametry modelu v ZS
  - e_i…náhodná chyba (reziduum, chyba predikce), odchylka naměřené hodnoty od hodnoty předpovídané vyrovnávací křivkou
- Parametry modelu odhadneme na základě n nezávislých pozorování, uspořádaných do dvojic číselných údajů (X1, Y1), (X2, Y2) pro proměnné
- Bodové odhady **a, b** parametrů α,β **regresní přímky** se z pozorovaných dat nejčastěji získávají **metodou nejmenších čtverců**
- **Metoda nejmenších čtverců** vychází z požadavku, aby součet čtverců odchylek pozorovaných hodnot (součet druhých mocnin reziduálních hodnot) byl minimální
- Tento postup zaručí, že **výběrová regresní funkce** bude co nejlépe **přiléhat k výběrovým hodnotám**
- Soustava normálních rovnic pro regresní přímku
- V indexu je na prvním místě vždy uváděna proměnná považovaná za závisle proměnnou (y závisí na x)
- Řešením soustavy rovnic určíme parametry výběrové **regresní přímky** (regresního modelu, regresní funkce)
- **Jednostranná závislost** – **proměnná X je nezávisle proměnná a Y pak závisle proměnná**
  - a_{yx} … absolutní člen
  - b_{yx} … regresní koeficient, vyjadřuje jednotkovou změnu (když se změní x, tak o kolik se změní y)
  - y'_i … vyrovnaná (teoretická) hodnota vysvětlované proměnné
- **Oboustranná závislost** – nelze rozhodnout, která proměnná je závislá a která (sdružená fce)

$$y'_i = a_{yx} + b_{yx} \cdot x_i \qquad \hat{y}'_i = a_{yx} + b_{yx} \cdot x_i \quad \text{nezávislá}$$

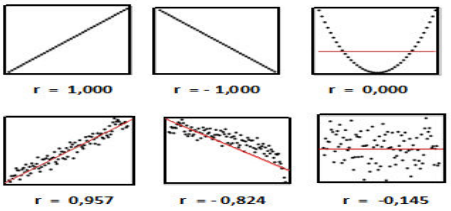
$$x'_i = a_{xy} + b_{xy} \cdot y_i$$



Korelační analýza

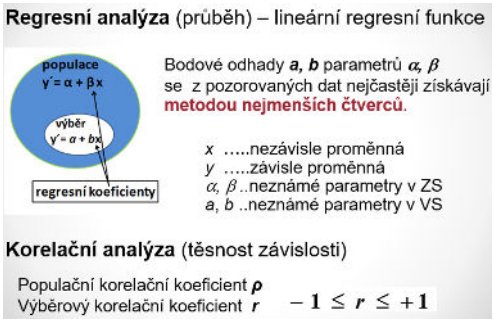
- **Korelace** obecně označuje míru stupně (sílu) závislosti dvou proměnných X a Y.
- **Měření těsnosti (síly)** závislosti – spočívá ve zjištění, jak těsně se jednotlivé skutečné napozorované hodnoty přimykají k regresní čáře, která vystihuje průběh závislosti.
- Pro jednoduchou korelaci: **párový korelační koeficient**, vyjadřuje míru lineární závislosti mezi náhodnými veličinami.
- **Pearsonův koeficient** korelace – předpokladem je, že obě náhodné proměnné X a Y, pocházejí z **normálního rozdělení** $(r_{yx} = r_{xy})$
  - Platí < -1 ≤ r ≤ +1 > … dvě náhodné proměnné jsou tím více korelovány, čím blíže je hodnota korelačního koeficientu s číslům +1 nebo -1
    - Jestliže $|r|=1$, mezi veličinami X a Y **existuje lineární funkční závislost**
    - Jestliže r = 0, proměnné X a Y jsou lineárně nezávislé, tzn. nekorelované
  - Korelační koeficient počítáme pomocí tzv. kovariance s_{yx} a směrodatných odchylek s_x a s_y obou proměnných

- **Koeficient determinace** r_{yx}^2 je druhou mocninou koeficientu korelace ($0 \leq r^2 \leq 1$)
  - Udává, z kolika % jsou změny závisle proměnné vyvolané změnami nezávisle proměnné – ověřuje správnost zvoleného regresního modelu.
    - Proložení regresní přímky korelačním polem
    - Spearmanův korelační koeficient pořadí – **neparametrický korelační koeficient**, vycházející nikoli z hodnot, ale z jejich pořadí.



Spearmanův koeficient pořadí

- Používá se u **méně rozsáhlých souborů** v případě, že chceme získat rychlou představu o intenzitě závislosti a nebo v případech, kdy není splněna normalita rozdělené náhodných výběrů X a Y
- Koeficient je rezistentní vůči odlehlým hodnotám
- Spearmanův koeficient korelace r_s nabývá hodnot z intervalu (-1, 1)
- Jestliže $r_s = 1$, resp. $r_s = -1$, párové hodnoty (x_i, y_i) leží na nějaké vzestupné, resp., klesající funkci
- Shrnutí:



Testování regresních a korelačních charakteristik

Testování hypotéz

- Regresní a korelační analýzu provádíme většinou z dat získaných pomocí náhodného výběru
- Výsledky provedených analýz jsou jen odhady pro shodné závislosti v základním (populačním) souboru
- Další částí regresní a korelační analýzy je proto testování hypotéz o hodnotách charakteristik regrese a korelace
- Podstatné testy významnosti v korelačních a regresní analýze
  - test významnosti korelačního koeficientu
  - test významnosti jednotlivých regresních parametrů
  - test významnosti regresního modelu jako celku

Testování korelačního koeficientu

- **Test významnosti** korelačního koeficientu odpovídá na otázku, zda je korelace mezi výběrovými proměnnými natolik silná, abychom ji mohli považovat za prokázanou i pro základní soubor
- Testuje se hypotéza o nulové hodnotě korelačního koeficientu základního souboru
- **Testování Pearsonova korelačního koeficientu**
  - Hypotéza předpokládá, že korelace neexistuje, tzn. veličiny X a Y jsou nezávislé
    - H0: $\rho = 0$
  - Alternativní hypotéza je postavena na existenci korelace
    - H1: $\rho \neq 0$
  - Testy hypotézy se provádí pomocí testového kritéria
    - $t = \frac{|r|}{\sqrt{1-r^2}} \cdot \sqrt{n-2}$
    - V případě, že vypočtená hodnota testového kritéria padne do kritického oboru, zamítá se nulová hypotéze **a existence lineární korelační závislosti se považuje za prokázanou.**

- **Testování Spearmanova koeficientu pořadové korelace**
  - H0: $\rho_s = 0$
  - H1: $\rho_s \neq 0$

- Hypotéza předpokládá, že korelace neexistuje, tzn. Veličiny X a Y jsou nezávislé (alternativní hypotéza je postavena na existenci korelace)
- Testování se provádí pomocí tabulek.

Testování regresního koeficientu

- Testujeme předpoklad, že výběrový regresní koeficient **b** je odhadem regresního koeficientu základního souboru β
- **Test významnosti** nulové hypotézy vychází ze skutečnosti, že regresní koeficient je roven 0 (přímka nemá směrnici, je statisticky nevýznamná)
- H0: β=0 H1: β ≠0
- Test hypotézy se provádí pomocí testového kritéria
- V případě, že se zamítá H0, je existence lineární závislosti prokázána.

$$s_b = \sqrt{\frac{s_r^2}{\sum (x_i - \bar{x})^2}}$$

Test regresního modelu

- Test významnosti celé regresní přímky (modelu) se provádí pomocí **upravené jednoduché ANOVY**
- Je-li regresní model statisticky významný, znamená to, že odhady závisle proměnné **Y'**, které z modelu dostaneme, jsou statisticky významně lepší než odhady pomocí průměru **Y'**
- **V případě lineární regresní funkce je závěr testů významnosti celého regresního modelu shodný (ekvivalentní) s testem regresního koeficientu!!!**
- **Pro rovnici s jedním prediktorem $F=t^2$.**
- Testujeme nulovou hypotézu o nulovosti všech regresních koeficientů (H0: všechna b=0 // H1: non H0)

Odhad regresních a korelačních charakteristik

Korelační charakteristiky

- Bodový odhad populačního korelačního koeficientu ρ $\hat{\rho} = \sqrt{1-(1-r)^2 \cdot \frac{n-1}{n-2}}$
- Intervalový odhad populačního korelačního koeficientu ρ
  - Postup výpočtu záleží na rozsahu výběrového souboru
    - V případě, že výběrový soubor má dostatečně velký rozsah (n > 100), lze rozdělení výběrového korelačního koeficientu aproximovat normálním rozdělením
      - Oboustranný interval spolehlivosti
      - $P\,(r - u_\alpha \,.\, s_r \le \rho \le r + u_\alpha \,.\, s_r) = 1 - \alpha$ $s_r = \frac{1-r^2}{\sqrt{n}}$
    - V případě, že výběrový soubor má rozsah n < 100, provádíme Fisherovu Z-transformaci
      - R >> Z a zpětně inverzí transformaci Z >> r (převody hodnot provádíme pomocí tabulek)

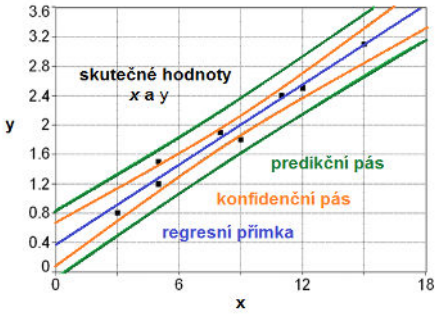
$$P\,(Z - u_\alpha \frac{1}{\sqrt{n-3}}, Z + u_\alpha \frac{1}{\sqrt{n-3}}) = 1-\alpha$$

Regresní charakteristiky

- Bodový odhad regresního koeficientu získáváme pomocí metody nejmenších čtverců, tzn. β=b
- Oboustranný interval spolehlivosti pro regresní koeficient β je vymezen následujícím vztahem

$$P\left(b - t_{\alpha(n-2)} \cdot s \le \beta \le b + t_{\alpha(n-2)} \cdot s_b\right) = 1-\alpha$$

- Výběrovou regresní přímku můžeme využít:
  - pro **odhad** podmíněné střední hodnoty závislé veličiny Y odpovídající určité konkrétní hodnotě nezávislé veličiny X
    - Konfidenční pás pro přímku
  - pro **předpověď** individuální hodnoty veličiny Y' odpovídající určité hodnotě nezávislé veličiny X'
    - Predikční pás pro jednotlivá pozorování



Shrnutí – podstatou řešení regresní analýzy je:

- Stanovit nejvhodnější tvar regresního modelu (tedy určit příslušnou rovnici, která bude popisovat závislost y na x)

- Stanovit jeho parametry (tj. stanovit konkrétní hodnoty parametrů β)
- Stanovit statistickou významnost parametru a celého modelu (tj. zda model podstatným způsobem přispěje ke zpřesnění odhadu závisle proměnné)
- Výsledky dané modelem interpretovat z hlediska zadání

Konfidenční pás pro přímku
- Kritérium pro přesnost určení přímky
- Udává rozpětí, ve kterém se budou v ZS nacházet hodnoty závislé proměnné se zvolenou pravděpodobností 1 - α

Predikční pás pro přímku
- Kritérium přesnosti predikce
- Při předpovídání jsme vystavení větší nejistotě. Interval předpovědi je širší než interval spolehlivost. Predikovaná hodnota je v hranici pásu s pravděpodobností 1 - α

Diagnostika lineárního regresního modelu

Postup tvorby lineárního modelu
1. návrh modelu (od nejjednoduššího)
2. předběžná analýza dat (sleduje se proměnlivost proměnných a možné párové vztahy)
3. odhadování parametrů modelu (metodou MNČ)
4. regresní diagnostika (identifikace odlehlých pozorování a ověření předpokladů MNČ)
5. konstrukce zpřesněného modelu
6. zhodnocení kvality modelu (testy, regresní diagnostika nového modelu)

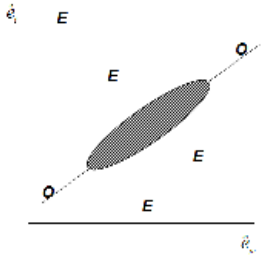
Regresní diagnostika
- Cílem regresní diagnostiky je ověřit, zda navržený regresní model koresponduje s reálnými daty a zda metoda nejmenších čtverců je vhodná pro odhad parametru
- Základní rozdíl mezi regresní diagnostikou a testy regresních a korelačních charakteristik je v tom, že není třeba přesně formulovat alternativní hypotézu
- Regresní diagnostika umožňuje interaktivní zásah uživatele do tvorby regresních modelů
- Regresní diagnostika obsahuje postupy k posouzení:
  - **kvality dat** pro navržený model (kritika a analýza vstupních dat)
  - **kvality modelu** pro daná data (kritika a analýza modelu jako celku)
  - **splnění předpokladů** požadovaných metodou **MNČ** (kritika a analýza **metody odhadů)**
- Regresní triptych = kvalita dat plus kvalita modelu plus kvalita odhad MNČ
- Z hlediska využití statistického softwaru, který je nedílným prostředkem prováděných analýz je možné regresní diagnostiky rozdělit do dvou částí:
- analýzy zaměřené na posouzení kvality dat
- analýzy zaměřené na kvalitu modelu a předpoklady MNČ

Diagnostika kvality dat
- Při posouzení kvality vstupních dat se zaměřujeme na **vlivné body**
- Vlivné body jsou takové body, jejichž vynecháním dochází **k zásadní změně regresních charakteristik** (odhadu parametru, vyrovnaných hodnot atd.) pro je nutné tyto body identifikovat
- Identifikace vlivných bodl je založena na technikách, které vycházejí z hodnocení důsledků vypuštění i-tého bodu.

Vlivné body lze rozdělit do 3 skupin:
- **hrubé chyby**, jsou způsobeny chybou měření nebo nevhodným nastavením vysvětlujícími proměnných

- **body s vysokým vlivem**, jsou <u>speciálně vybrané body</u>, které byly dobře změřeny a které obvykle rozšiřují predikční schopnosti modelu
- **zdánlivě vlivné body**, vznikají jako důsledek <u>nesprávně navrženého regresního modelu</u>

Podle polohy rozdělujeme vlivné body na:
- **vybočující pozorování** (outliers) – body, které se liší v hodnotách vysvětlované proměnné **y** (Studentizovaná rezidua)
- **extrémy** (leverage points) – body, které se liší v hodnotách vysvětlující proměnné **x** (projekční matice H)
- Ve statistických programech se na rozdíl od odborné statistické literatury nerozlišují vybočující pozorování a extrémy a obecně se hovoří o odlehlých pozorováních
- Nejjednodušší grafické vyjádření odlehlých pozorování – graf predikovaných reziduí
  - osa y … klasická rezidua $\hat{e}_i$
  - osa x … predikovaná rezidua $\hat{e}_{pi}$
  - O … vybočující bod (leží na přímce nebo v její blízkosti, ale daleko od ostatních bodů)
  - E … extrémní bod (leží mimo přímku)



Diagnostika kvality dat
- <u>Projekční matice H</u> (hat matrix) – diagnostický nástroj pro identifikaci vlivu jednotlivých pozorování
  - Zaměřuje se na hledání odlehlých hodnot v množině X
  - Za odlehlé pozorování se považuje takové, jehož diagonální prvek matice $\mathbf{h_{ii} > 2p/n}$
  - (p=počet parametrů modelu, n=počet pozorování)
- <u>Studentizovaná rezidua</u> slouží k identifikaci vlivných bodů z hlediska vysvětlované proměnné
- Zaměřuje se na hledání odlehlých hodnot v množině Y.
- Proměnné s absolutní hodnotou vyšší než dva indikují nevhodnou proměnnou **/SR/ > 2**
- Odlehlost neznamená vlivnost!

Míra vlivů kombinace proměnných x a y
- <u>Cookova D veličina</u>, která kombinuje h-hodnotu se standardizovaným reziduem
- Tato charakteristika popisuje, jak vlivné pozorování ovlivňuje celou vypočtenou rovnici všechny y‘
- Orientačně platí, že velmi vlivné body indikuje hodnota $\mathbf{D_i > 4/n}$

Míra vlivu jednotlivých pozorování
- Pro vyjádření toho, jak vlivné pozorování ovlivňuje příslušné vyrovnané pozorování (konkrétní y‘)
- Welshova-Kuhova vzdálenost DFFIT(-i)
- Přičemž platí, že pro vlivné veličiny je normovaný
- /DFFIT(-i)/ > 2 (n/bla

Analýza reziduí
- Rezidua lze zjednodušeně charakterizovat jako **lineární kombinaci všech chyb**
- Reziduum je **vyčíslená hodnota z regresního modelu** a používá se při **<u>posouzení kvality dat,</u>** ale i **<u>kvality modelu jako celku</u>**
- Systematičnost (**nenáhodnost**) zjištěná u reziduí **indikuje** nějaký zatím neidentifikovaný **nedostatek** odhadnutého regresního **modelu**.

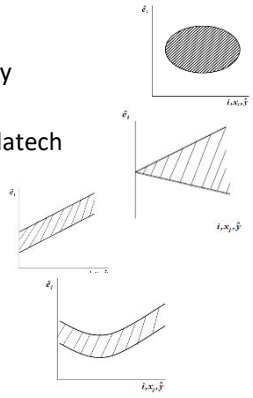
Požadavky kladené na rezidua
- reziduum $\mathbf{E_i = y_i – y_i‘}$
  - mají nulovou **střední hodnotu** $\mathbf{E(e_i)=0}$ a **konstantní rozptyl** $\mathbf{E(e_i^2)=ro2}$
  - jsou **nekorelované** (jsou nezávislé)
  - jsou **náhodná** a mají **normální rozdělení**
- Nulová střední hodnota reziduí se testuje pomocí jednovýběrového t-testu

- Nekonstantnost rozptylu nebo-li heteroskedasticita
- Pokud je splněn předpoklad o konstantním rozptylu, rezidua jsou označována jako homoskedastická
- Zatímco systematické změny variability reziduí při změnách vysvětlujících proměnných jsou známkou heteroskedasticity.

Testování heteroskedasticity reziduí např. **Whiteův test**
- Rezidua v tomto testu odhadována pomocí MNČ. Whiteův test identifikuje pouze ta rezidua, která nejsou heteroskedastická. Test proto může být statisticky významný pouze v případě, že jsou všechna rezidua homoskedastická.

- Rezidua jsou **nekorelované** – posouzení podmínky o nezávislosti reziduí se provádí pomocí **Durbin-Watsonova statistiky**. Vzájemná závislost reziduí může naznačovat autokorelaci
- Hodnoty této statistiky se pohybují od 0 do 4. Pokud je tato statistika =2, rezidua nevykazují žádnou autokorelaci, hodnoty statistiky < 2 značí pozitivní autokorelaci, hodnoty > 2 značí autokorelaci negativní
- Posouzení podmínky o **náhodnosti** reziduí se provádí pomocí **testu normality reziduí nebo graficky**
- Testy normality reziduí nemusí ještě identifikovat nesprávnost navrženého regresního modelu nebo nevhodnost dat
- Zjištěná nenormalita pravděpodobnostního rozdělení reziduí, může být způsobena přítomností vybočujících a vlivných pozorování, ale také velmi často změnou podmínek při měření.

Grafická analýza reziduí



- Pokud se v diagnostických grafech reziduí objeví tvar „mraků" bodů, je detekována správnost metody nejmenších čtverců
  - tvar mrak – správnost MNČ (jiné obrazce bodů v grafech reziduí indikují většinou nesprávnost v datech či nesprávnost modelu
  - výseč – odhaluje nekonstantnost rozptylu čili heteroskedasticitu v datech
  - pás – indikuje chybu ve výpočtu nebo se jedná o důsledek vybočujících hodnot (nebo v regresním modelu chybí absolutní člen)
  - nelineární tvar – nesprávně navržený model
- analýza reziduí a její grafické znázornění, nám také může sloužit k první identifikaci odlehlých pozorování

Kritika a analýza modelu jako celku
- Kvalitu regresního modelu lze posoudit **graficky či numericky**
- V případě jedné vysvětlující proměnné x přímo z rozptylového grafu závislosti y na x
- Numericky s využitím klasických testů (F-test, t-test) a postupů regresní diagnostiky.

Splnění předpokladů MNČ
- Regresní parametry β mohou teoreticky nabývat jakýchkoli hodnost
- Regresní model je lineární v parametrech
- Nezávislé proměnné x jsou skutečně vzájemně nezávislé
- Podmíněny rozptyl D(y/x) = ro2 je konstantní
- Náhodné chyby εi mají nulovou střední hodnotu, konečný a konstantní rozptyl a jsou nekorelované.

- Jednotlivé **nezávislé proměnné x** jsou skutečně nezávislé vzájemně, tedy jestliže mezi nimi nedochází k multikolinearitě (vzájemné lineární závislosti)
- Mezi základní kritéria pro identifikaci multikolinearity patří **jednoduché korelační koeficienty** dvojic vysvětlujících proměnných nebo **vícenásobné korelační koeficienty**
- Hodnoty korelačních koeficientů blízké **plusmínus 1** naznačují množnost **existence multikolinearity.**

Multikolinearita má za následek:

- **Nadhodnocení součtu čtverců regresních koeficientů**, což může zapříčinit, že některé vysvětlující proměnné vypadají důležitější, než ve skutečnosti jsou. (např. malá změna hodnot závisle proměnné znamená zásadní změnu parametrů)
- **Snižuje přesnost odhadů intervalů spolehlivosti** a komplikuje interpretaci individuálního vlivu jednotlivých vysvětlujících proměnných (interval spolehlivost parametrů je tak velký, že odhad parametrů ztrácí smysl)
- **Nelze** odděleně **sledovat skutečný vliv jednotlivých** vysvětlujících vstupních **proměnných** na vysvětlovanou (závislou) proměnnou
- **Podmíněný rozptyl** je konstantní tzv. je splněna podmínka **homoskedasticity** – hodnoty závisle proměnné y mají pro všechny hodnoty nezávisle proměnné **x konstantní rozptyl** (variabilitu, proměnlivost)
- **Náhodné chyby** mají **nulovou střední hodnotu** E = O. mají **konečný a konstantní rozptyl** E = rozptyl a jsou **nekorelované** (ověřujeme na základě analýzy reziduí).

Nelineární regresní funkce

- Na rozdíl od lineárních regresních modelů je třeba u nelineárních modelů počítat s řadou komplikací:
  - neodhadnutelností některých parametrů
  - existencí minima funkce jen pro některé regresní modely
  - výskytem lokálních minim a sedlových bodů
  - špatnou podmíněností parametrů v regresním modelu
  - malým rozmezím experimentálních dat (zejména u parametrů vyjadřujících limitní chování modelů)
- funkce lineární v parametrech
  - aditivní typy funkcí - kvadratická, lomená, logaritmická, kubická
- funkce nelineární v parametrech
  - multiplikativní typy funkcí - exponenciální, s-křivka, mocninná
- Při výběru typu funkce je třeba vycházet nejen z formálního hlediska **(nejvyšší hodnota indexu korelace**), ale u z hlediska **věcně logického**, podle věcné podstaty zkoumané závislosti
- Při odhadu neznámých parametrů v nelineárním modelu lze použít opět **metodu nejmenších čtverců**, i když se často volí i jiná kritéria či postupy (někdy se na základě předpokladu o typu rozdělení náhodných chyb hledají maximálně věrohodné odhady)
- Metodu nejmenších čtverců, lze použít jen u **nelineárních funkcí**, které jsou **lineární v parametrech**
- **Funkce, které nejsou lineární v parametrech, můžeme vhodnou linearizací převést na lineární** (mocninnou nebo exponenciální funkci logaritmujeme)
- Vlastnosti, které platí pro odhad regresní funkce získaný klasickou MNČ, platí pouze pro transformovanou funkci. Důsledkem toho je, že **odhady jednotlivých regresních koeficientů** užitého modelu **nesplňují podmínku nestrannosti**
- Jak již bylo uvedeno, odhad parametrů u regresních funkcí, **které nejsou lineární v parametrech, neprovádíme MNČ přímo**, protože její použití vede k soustavě nelineárních rovnic, z nichž zpravidla nedokážeme odhadnout přímo parametry ve formě vhodných výpočetních vzorců
- Používá se tedy způsob, kdy určitou regresní funkci, která je nelineární z hlediska parametrů, převedeme pomocí **linearizující transformace na funkci lineární v parametrech**

Korelace při nelineární regresi

- Využívají se při **posouzení volby typu regresního modelu** a zjištění síly závislosti mezi proměnnou x a y
- Posuzovaný vztah je tím silnější a regresní funkce tím lepší, **čím více jsou empirické hodnoty vysvětlované proměnné soustředěné kolem odhadnuté regresní funkce**, a naopak tím slabší, čím více jsou empirické hodnoty vzdáleny hodnotám vyrovnaným
- Slouží také k **posouzení přesnosti regresních odhadů**. Čím více se jednotlivé napozorované hodnoty soustřeďují kolem zvolené regresní čáry, tím je závislost těsnější a odhad přesnější

- Při konstrukci míry ukazující na sílu závislosti vycházíme ze vztahu **empirických a vyrovnaných hodnot**
  - **$S_y = S_T + S_r$**
    - Sy … součet čtverců odchylek y
    - ST … součet čtverců vyrovnaných hodnot
    - Sr … reziduální součet čtverců
- Sílu závislosti je možné měřit **indexem determinace**
  - index korelace se používá k měření těsnosti závislosti pro libovolnou regresní funkci, jejíž parametry byly odhadnuty metodou nejmenších čtverců. Má menší vypovídací schopnost než index determinace
- Nízká hodnota indexu determinace nemusí ještě znamenat nízký stupeň závislosti mezi proměnnými, ale může signalizovat chybnou volbu regresní funkce

Mnohonásobná lineární regrese a korelace

- **Cílem mnohonásobné (vícenásobné) regresní analýzy je zkoumání statistické závislosti** pomocí modelu, jenž zahrnuje jednu závisle proměnnou Y a několik nezávislých proměnných x
- **Cílem mnohonásobně korelační analýzy je určení síly závislosti**, pomocí které je hodnocen stupeň závislosti, a to jak pro případ společného vlivu všech zúčastněných proměnných, tak i pro případ jejich dílčího vlivu
- Mnohonásobná korelační závislost nám umožňuje sledovat, jak závisí proměnná y nejen na vysvětlující proměnné x1, ale také na další proměnné x2, x3…

Koeficient párový
Koeficient vícenásobné (totální) korelace
Koeficient dílčí (parciální) korelace

Mnohonásobná korelace
Sílu jednoduché lineární závislosti mezi jednou závisle proměnnou y a jednou vysvětlující proměnnou x udávají:
Párové korelační koeficienty r <-1,1>
r_{yx1} r_{yx2} r_{x1x2}

Koeficienty dílčí (parciální) korelace r
- charakterizuje sílu lineární závislosti mezi závisle proměnnou a jednou nezávisle proměnnou, jsou-li hodnoty zbývajících proměnných v modelu konstantní. r <-1,1>
- Obecně platí, že udává závislost mezi proměnnými **uvedenými před tečkou**, za předpokladu, že všechny ostatní proměnné uvedení v indexu **za tečkou** jsou konstantní
- $r_{yx1}.x_2$ parciální korelační koeficient mezi y a x_1 s vyloučením vlivu x_2 (při konstantním vlivu x_2)

- Sílu vztahu závisle proměnné y na všech vysvětlujících proměnných x udává: **Koeficient vícenásobné (totální) korelace R**, R <0,1>
- 1 = úplná závislost, 0 = nezávislost
- Měří jedna těsnost závislosti mezi proměnnými a umožňuje tím posoudit kvalitu regresního odhadu zkonstruovaného na základě vícenásobné regresní funkce, jednak jej lze použít při hodnocení volby vysvětlujících proměnných.
- R^2 Opravená hodnota (adjusted) nebere v úvahu stupně volnosti, proto je vždy v modelu s větším počtem vysvětlujících proměnných vyšší hodnota R2. Potřebujeme-li porovnat kvalitu modelů s různým počtem vysvětlujících

Mnohonásobná regresní analýza
- Je to metoda, pro modelování závislostí několika vysvětlovaných náhodných veličin (závisle proměnných) Y1, Y2… na jedné nebo několika vysvětlujících veličinách (nezávisle proměnných X1, X2 .. XK

- Smyslem regresní analýzy je nalezení vhodného regresního modelu, který nám umožní popsat a kvantifikovat závislost vysvětlovaných proměnných na vysvětlujících proměnných.

Cíle:

1. **vysvětlit rozptyl** v závisle proměnné Y (pomocí R^2)
2. **odhadnout (vypočítat) vliv** každí z nezávisle proměnných X na proměnnou závislou Y (pomocí parciálních regresních koeficientů b)
3. **predikovat** pomocí sestavené regresní rovnice pro jednotlivé případy hodnoty závisle proměnné

- Před vlastní regresní analýzou je potřeba **ověřit kvalitu dat!**
- Samotné analýze tedy musí předcházet podrobná diagnostika (analýza) vstupních proměnných.
- Model vyjadřující závislost veličiny Y na veličinách X1, X2.. lze zapsat ve tvaru **yi = f(xi1, xi2,….,xik) + ε**
  - f(xi1, xi2,….,xik) … regresní funkce (i = 1,2,…, n)
  - ε … náhodná chyba
- Nejjednodušší formou regresního modelu **je lineární vícenásobný lineární model**, který používáme jestliže závislá proměnná y je lineárně závislá na každé z vysvětlujících X1, X2..
- Vícenásobná lineární regrese je založena na Pearsonově korelačním koeficientu, takže **neexistence linearity způsobuje**, že i důležité vztahy mezi proměnnými, pokud nejsou lineární, **zůstanou neodhaleny**
- Lineární vícenásobný regresní model Y = β
- Odhadnutou regresní funkci lze zapsat ve tvaru (MMČ)
  - $y' = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$
    - b_0 … absolutní člen
    - $b_1 \ldots b_k$ … jsou dílčí parciální regresní koeficienty, které udávají změnu závisle proměnné y odpovídající jednotkové změně jedné nezávisle proměnné x, za předpokladu, že hodnoty zbývající nezávisle proměnných v modelu jsou konstantní (vyjadřuje pouze část vlivu, působících na vysvětlovanou proměnnou y)

Předpoklady modelu (viz 4.př.)

- **Vysvětlující proměnné** musí být vzájemně **nezávislé** – nesmí být korelované. Vysoké vzájemné korela jsou zdrojem **multikoliearity**. Pokud v datech existuje multikolinearita, výsledky regrese jsou nespolehlivé
- **Náhodné chyby ε** jsou nezávislé, normálně rozdělené náhodné veličiny s nulovými středními hodnotami a stejným rozptylem (**homoskedasticita**)

Hodnocení mnohonásobného modelu z hlediska testů významnosti

- Test významnosti dílčích výběrových regresních koeficient (parametrů b) provádíme pomocí **t-testů**
- Test významnosti celého regresního modelu se provádí pomocí upravené jednoduché **ANOVY – F-testů**

Výsledek F-testu	Výsledek t-testu	Hodnocení modelu
nevýznamný	Všechny nevýznamné	Posuzované proměnné jsou lineární nezávislé, model je nevhodný, nevystihuje variabilitu závisle proměnné
významný	Všechny významné	Model se považuje za vhodný k vystižení variability proměnné y, to však neznamená, že je optimálně navržen
významný	Některé nevýznamné	Model je vhodný, ale provádí se zpravidla vypuštění nevýznamných parametrů modelu
významný	Všechny nevýznamné	Zvláštní případ způsobený multikolinearitou, paradox – model je nutné upravit a nebo zcela změnit

Metody výběru prediktorů (x)

- ENTER – všechny prediktory vstoupí do rovnice (rozhodnutí uživatele)
- Metody, které vycházejí z přírůstku regresního součtu čtverců, jehož velikost je hodnocena pomocí F-testů, nebo na základě zvýšení indexu determinace
  - 1. metoda FORWARD – postupné zařazování prediktorů
  - 2. metoda BACKWARD – postupné vyřazování prediktorů

- 3. metoda STEPWISE – kombinace obou, je založena na postupném vstupu bloků proměnných (prediktorů)

ČASOVÉ ŘADY

Časová řada

= posloupnost v čase seřazených údajů, zpravidla ve směru minulost-přítomnost.

- Analýza časových řad = soubor metod, které slouží k:
  - popisu dynamiky vývoje sledovaných jevů v referenčním období (tj. období, kterého se to týká)
  - prognózování budoucího vývoje

Základní druhy časových řad

- Podle rozhodného časového hlediska
  - **Intervalové časové řady** – obsahují údaje, které se vztahují k určitému časovému intervalu k jednomu roku, měsíci, součet hodnot má věcný význam (např. HDP/rok, tržby/měsíc)
  - **Okamžikové časové řady** – sestaveny k určitému rozhodujícímu okamžiku, součet nemá smysl (např. počet pracovníků k 31.12.)
- Podle periodicity sledování
  - **Krátkodobé** – týdenní, měsíční, čtvrtletní
  - **Dlouhodobé** – roční, víceleté
- Podle druhu sledovaných ukazatelů
  - **ČŘ primárních ukazatelů** – tj. ukazatelů prvotních
  - **ČŘ sekundárních ukazatelů** – tj. ukazatelů odvozených (součtové, průměrné nebo poměrové)
- Podle způsobu vyjádření údajů
  - **ČŘ naturálních ukazatelů**
  - **ČŘ peněžních ukazatelů**

Srovnatelnost údajů v ČŘ

Každá ČŘ musí splňovat 3 hlediska srovnatelnosti:

1) hledisko **věcné** srovnatelnosti
2) hledisko **prostorové** srovnatelnosti
3) hledisko **časové** srovnatelnost

Elementární charakteristiky ČŘ

- **Elementární charakteristiky** slouží k **hodnocení vývoje** ukazatele a k rychlé informaci o **charakteru a chování** ukazatele v časové řadě
- Elementární charakteristiky je možné rozčlenit:
  1. na ukazatele, které posuzují **úrovně ČŘ**
  2. na ukazatele, které charakterizují **dynamiku (rychlost změn) vývoje ČŘ**

Ukazatele posouzení úrovně ČŘ

- Zajištění srovnatelnosti se provádí přepočítáním očištěním časové řady od kalendářních variací
- Úroveň hodnot časové řady ve sledovaném období charakterizujeme pomocí průměru řady
  - Intervalové časové řady
  - Aritmetický průměr $\bar{y} = \frac{\sum y_i}{n}$
- pomocí průměru řady
- aby nedošlo při srovnání ukazatelů za sledované období ke zkreslení, je třeba, aby se hodnoty vztahovaly ke stejně dlouhým intervalům
- okamžikové časové řady – chronologický průměr prostý – vzorec vyjadřuje prostou formu chronologického průměru za předpokladu, že délka mezi jednotlivými časovými okamžiky je stejná
- není-li délka mezi jednotlivými časovými okamžiky konstantní, je nutné jednotliví dílčí průměry vážit délkami příslušných intervalů

- zajištění srovnatelnosti se provádí přepočítáním očištěním časové řady od kalendářních variací
  - $y_t^{(0)} = y_t \frac{\bar{k}_t}{k_t}$ t = 1, 2, …, n
    - y_t … původní hodnota ČR
    - $\bar{k}_t$ … průměrný počet dní v dílčím období
    - k_t … počet dní v dílčím období (měsíc, čtvrtletí)
- Okamžikové časové řady
  - Chronologický průměr prostý
    - Vzorec vyjadřuje prostou formu chronologického průměru za předpokladu, že délka mezi jednotlivými časovými okamžiky je stejná

$$\bar{y} = \frac{\frac{1}{2}y_1 + y_2 + \cdots + y_{n-1} + \frac{1}{2}y_n}{n-1}$$

  - Chronologický průměr vážený
    - Není-li délka mezi jednotlivými časovými okamžiky konstantní, je nutné jednotlivé dílčí průměry vážit délkami příslušných intervalů
    - $$\bar{y} = \frac{\frac{y_1+y_2}{2}(t_2-t_1) + \frac{y_2+y_3}{2}(t_3-t_2) + \cdots + \frac{y_{n-1}+y_n}{2}(t_n-t_{n-1})}{t_n - t_1}$$

Ukazatele dynamiky vývoje ČŘ

Absolutní charakteristiky

- **První absolutní diference** – charakterizuje přírůstek (resp. úbytek) hodnoty ukazatele ČŘ v určitém období proti období bezprostředně předcházejícímu
  - $d_{yt}^1 = y_t - y_{t-1}$ t = 2, 3, …, n
- **Druhá absolutní diference** – vyjadřuje zrychlení (resp. zpomalení) časové řady, určí se na základě porovnání absolutních přírůstků
  - $d_{yt}^2 = d_{yt}^1 - d_{t-1}^1$ t = 3, 4, … n
- Souhrnnou absolutní charakteristikou je **průměrný absolutní přírůstek** (resp. úbytek) hodnoty ukazatele ČŘ
  - $$\bar{d} = \frac{(y_2-y_1)+(y_3-y_2)+\cdots+(y_n-y_{n-1})}{n-1} = \frac{y_n - y_1}{n-1}$$

Relativní charakteristiky

- **Koeficient růstu** (řetězový index) – charakterizuje relativní postupnou rychlost změn hodnot v ČŘ. Vyjádřený v **%** se nazývá **tempem růstu**
  - $k_t = \frac{y_t}{y_{t-1}}$ t = 2, 3, …, n
- **Průměrný koeficient růstu** – úhrnná charakteristika relativních změn pro celou ČŘ, vypočítá se jako geometrickým průměrem z jednotlivých koeficientů růstu
  - $$\bar{k} = \sqrt[n-1]{k_1 \cdot k_2 \cdot \ldots\ldots \cdot k_{n-1}} = \sqrt[n-1]{\frac{y_n}{y_1}}$$
- **Tempo přírůstku** – představuje porovnání první diference s příslušnou hodnotou časové řady

$$r_t = \frac{d_t^1}{y_{t-1}} = \frac{y_t - y_{t-1}}{y_{t-1}}$$

- **Koeficienty zrychlení** – pro vyjádření rychlosti změn v ČŘ (druhé relativní diference)

$$z_t = \frac{d_t^2}{d_{t-1}^1}$$

- **Bazický index** – vyjadřuje změny ČŘ vzhledem k základnímu období

$$BI_t = \frac{y_t}{y_0}$$

Modely časových řad

Modelování časových řad

- Klasická analýza časových řad vychází z předpokladu, že časovou řadu je možné rozdělit na 3 složky
  - **Trend (Tt)** – vyjadřuje dlouhodobou tendenci vývoje zkoumaného jevu, výsledek faktorů, které dlouhodobě působí stejným směrem

- **Periodická složka (Pt)** – je charakteru sezónního (délka periody je menší nebo rovna 1 rok), cyklického (délka periody je delší než 1 rok)
- **Náhodné kolísání (εt)** – nelze popsat žádnou funkci času, projevuje se nepravidelnými nebo ojedinělými výkyvy, ale také chybami měření atd.

Dekompozice časové řady

- **Aditivní model** – používáme tehdy má-li periodické kolísání kolem trendu přibližně stálou relativní amplitudu (rozkmit) $y_t' = T_t + P_t + \varepsilon_t$ t = 1, 2, … , n
- **Multiplikativní model** – používá se tehdy, pokud je velikost periodických kolísání úměrná úrovni trendu, logaritmickou transformaci lze převést na aditivní model $y_t' = T_t \cdot P_t \cdot \varepsilon_t$ t = 1, 2, …, n

Neperiodické časové řady

- bez periodické složky

Periodické časové řady

- obsahují periodickou složku

Analýza neperiodický ČŘ

- Hlavním úkolem analýzy neperiodických ČŘ je vystižení základní tendence jejich vývoje – **trendu**.
- Využíváme při tom metody, které nám slouží k vyrovnání ČŘ. Skutečné hodnoty ČŘ nahradíme teoretickými hodnotami, které jsou očištěné od periodického či náhodného kolísání.
- Popis trendu (trendové složky) v časových řadách:
  - **Graficky**
  - **mechanicky** (pomocí **klouzavých průměrů)**
  - **analyticky (**pomocí **trendových funkcí)**

Klouzavé průměry

- Vyrovnání pomocí klouzavých průměrů spočívá v nahrazení skutečných hodnot ČŘ průměrem z určitého počtu hodnot. **Trend v krátkých časových úsecích odhadujeme průměrem několika sousedních pozorování.**
- Při postupném výpočtu průměru postupujeme tak, že kloužeme vždy o jedno pozorování dopředu, přičemž zároveň nejstarší pozorování z té skupiny, z níž je průměr vypočítán, vypouštíme.
- Volba **klouzavé části k** (délka intervalu použitého k výpočtu průměru).
- Pokud klouzavé průměry počítáme ze sudého počtu období, vypočítaný průměr nespadá do určitého období, ale bude patřit do středu mezi dvě prostřední období.
- Abychom dostali klouzavý průměr dopovídající určité konkrétní hodnotě časové řady, musíme určit tzv. centrované klouzavé průměry.
- Proto volíme klouzavou část obvykle liché délky k, sudou délku volíme jen pro speciální případy (čtvrtletní, měsíční)

Trendové funkce

- Vyrovnání pomocí trendových funkcí
- Jde o vyjádření průběhu ČŘ matematickou funkcí, kde zkoumaný ukazatel ČŘ vystupuje jako závisle proměnná Yt a čas (časová proměnná) jako nezávisle proměnná t.
- Pro analytické vyrovnání se používá relativně nevelký okruh trendových funkcí.
- Tyto funkce by měli být z matematického hlediska jednoduché.
- Těmto vlastnostem odpovídají zejména tyto křivky: lineární, kvadratická, logaritmická, exponenciální, mocninná, odmocninná

Adaptivní modely časových řad

- Trendová složka časové řady není konstantní, ale mění se v čase, proto není možné k jejímu popisu použít jednu matematickou funkci s konstantními parametry

- Modely tohoto typu rychle reagují na strukturální změny, k nimž dochází v čase, a jsou **velmi vhodné pro prognózování průběhu** časových řad, které se vyznačují **nepravidelnostmi a zlomy v trendu.**
- Adaptivní modely vychází z předpokladu, že **pro konstrukci extrapolační prognózy budoucího vývoje mají cenu nejnovějšího pozorování časové řady.**
- **Nejnovějším pozorováním se přiřazují největší váhy**, a dřívější pozorování se buď úplně vyřazují ze zkoumání, nebo se jim přiřazují menší váhy ve srovnání s později pozorovanými hodnotami.
- **Adaptivní modely berou v úvahu stárnutí informací.**
- Skupina adaptivním modelů je rozsáhlá. Jedny z nejčastějších metod, které přináší v praktických aplikacích dobré výsledky, jsou metody exponenciálního vyrovnávání.
- Odhad trendu jako lineární kombinace všech minulých pozorování, bere v úvahu stárnutí pozorování čř, čím je pozorování starší tím má menší váhu
- Nejjednodušším případem je **jednoduché exponenciální vyrovnávání**, při kterém se trend považuje v krátkých úsecích časové řady za konstantní.
  - Jedny z nejčastěji používaných metod, které přináší v praktických aplikacích dobré výsledky, jsou metody exponenciálního vyrovnávání
- **Odhad trendu v čase t**

Metody exponenciálního vyrovnávání
- Jednoduché exponenciální vyrovnávání – trend v krátkých úsecích konstantní, jeden parametr α
- Brownovo exponenciální vyrovnávání – úroveň a trend řady, dva parametry
- Holtovo exponenciální vyrovnávání – úroveň a trend řady, dva parametry α, γ
- Exponenciální vyrovnání s tlumeným trendem – tři parametry α, γ, φ

Posouzení vhodnosti modelů ČŘ
- Základem pro rozhodování o vhodném typu trendové funkce by měla být:
  - **věcně ekonomická kritéria** – odhalují jen základní tendence ve vývoji analyzovaného ukazatele, nevedou však k volbě konkrétního typu trendové funkce
  - **míry shody** – podávají informace o stupni souladu empirických hodnot a teoretických hodnot
- Často používaným ukazatelem, který slouží k popisu stupně shody je **index determinace I^2**.
- Čím je jeho hodnota bližší 1, tím model lépe popisuje zkoumaný jev. Jestliže se hodnota I^2 blíží 0, signalizuje to stále menší soulad modelu ČŘ.
- Za nejvhodnější trendovou funkci považujeme tu, která vede k největší hodnotě I^2
- Moderní statistická metodologie standardně implementovaná v statistických programech
  - M.E. … střední chyba odhadu
  - M.S.E. … střední kvadratická chyba odhadu
  - M.A.E. … střední absolutní chyba odhadu
  - M.P.E. … střední chyba odhadu
  - M.A.P.E. … střední absolutní procentní chyba odhadu

$$\mathrm{MAPE} = \frac{100}{n}\sum \frac{|y_t - y'|}{y_t}$$

Prognóza provedená v čase t pro čas t plus 1
Jednoduché exponenciální vyrovnávání – trend v krátkých úsecích konstantní, jeden parametr α
Brownovo exponenciální vyrovnávání – úroveň a trend řady, dva parametry
Holtovo exponenciální vyrovnávání – úroveň a trend řady, dva parametry alfa a gama
Exponenciální vyrovnání s tlumeným trendem – tři parametry
Hodnocení přesnosti prognóz
Pseudoprognóza se konstruuje tak, že k vyrovnání časové řady se nevyužije několik posledních hodnot řady, které jsou tak jako by předpovídanými hodnotami.
Pro změření kvality skutečných předpovědí i pseudopředpovědí se používá **Theilův koeficient nesouladu T^2_H**.
Relativní chyba extrapolace (%) T_H
chyba predikce malá 0-5%

chyba predikce střední 5-10 %
chyba predikce velká, model pro predikci nepoužívat 10% a víc
Relativní chyba prognózy (predikce) P_t

Analýza periodických ČŘ
Periodická složka
- ≤ 1 rok … sezónní složka S_i (krátkodobé kolísání, opakování vývoje v průběhu roku u řad, které jsou měřeny měsíčně, čtvrtletně apod. – perioda 1 rok)
- > 1 rok … cyklická složka C_i (dlouhodobé kolísání, cykly s delší periodou než 1 rok)

Sezónní kolísání
- Sezónním kolísáním se rozumí soubor přímých či nepřímých příčin – vlivů, které se opakují
- Faktory (příčiny): objektivní, subjektivní
- Vždy je potřeba identifikovat, zda je sezónní kolísání skutečně statisticky významné (grafická analýza, výpočet klouzavých průměrů, autokorelační funkce, analýza periodogramu)

Popis sezónní složky
- Prokáže-li se existence sezónní složky (sezónních výkyvů) v časové řadě, provádíme její kvantifikaci neboli modelování
- Modelování sezónní složky závisí na typu rozkladu (dekompozici) časové řady.
- Aditivní model: Yij = Tij + Sij + eij
- Multiplikativní model: Yij = Tij * Sij * Eij

Intenzita sezónního kolísání
- **Aditivní model** – používáme tehdy má-li periodické kolísání kolem trendu přibližně stálou relativní amplitudu (rozkmit, konstantní sezónnost)
- Sezónní složka je v tomto případě vyjádřena pomocí **sezónních odchylek** (rozdíl mezi empirickými hodnotami a aritmetickým průměrem, resp. teoretickými-vyrovnanými hodnotami časové řady)
- **Součet sezónních odchylek=0**
- **Multiplikativní model** – používá se tehdy, pokud je velikost periodických kolísání úměrná úrovni trendu. Sezónní složka je vyjádřena pomocí **sezónních indexů**
  - $s_t = \frac{skuteč.hodnota\ řady}{vyrovnaná\ hodnota\ řady}$

Vyrovnaná (teoretická) hodnota
- **Aritmetický průměr** $\bar{y}$ skutečných hodnot za období celé periody sezónního cyklu (průměrný údaj, připadající na jedno období v rámci zkoumaného roku)
- **Vyrovnané hodnoty** y'_i - stanovené buď pomocí klouzavých průměrů, nebo některou metodou analytického vyrovnání (hodnoty vypočítané na základě trendové funkce)

Sezónní očišťování
- **Sezónní očišťování** časové řady zbavuje časovou řadu periodického kolísání, které by mohlo maskovat charakter trendu řady
- Používá se jako předběžný stupeň před analýzou trendu časové řady

Náhodné kolísání (náhodná složka)
- Náhodné (nesystematické) složky tzv. rezidua – chápeme jako výsledky působení určitých blíže nespecifikovaných (stochastických) náhodných vlivů
- Náhodnou složku εi vyjadřujeme ve tvaru $\varepsilon_i = y_i - y'_i$

- V případě, že jsme časovou řadu podrobili analýze a na jejím základě popsali systematické složky (trend, sezónnost, event. cyklus) zbývají nám odhady **náhodné (nesystematické) složky** tzv. rezidua
- Náhodnou složku chápeme jako výsledek působení blíže nespecifikovaného souboru náhodných vlivů
- Pokud jsme k odhadu parametrů systematických složek použili metodu nejmenších čtverců, předpokládáme splnění následujících předpokladů
- **Střední hodnota** náhodné složky εi se rovná nule
- Variabilit náhodných složek εi se v čase nemění, **rozptyl je konstantní**
- Jednotlivé hodnoty náhodné složky εi jsou vzájemně lineárně nezávislé (**nekorelované**)
- Jsou-li tyto předpoklady splněny, tvoří řada **tzv. bílý šum.**

Předpovědi časových řad

- Interpolace
- Extrapolace
- Bodová předpověď y'_{i+k}
- Intervalová předpověď $P\left(u_{i+k} - \Delta \le u_{n+k} \le u_{i+k} + \Delta\right) = 1 - \alpha,$
  - kde i je pořadové číslo časové proměnné v časové řadě o n členech, k – počet kroků dopředu
  - každá předpověď je spojena s určitou chybou předpovědi. Případná chyba je tím větší, čím kratší je délka časové řady, čím nedokonalejší je popis uplynulého vývoje a čím vzdálenější je horizont předpovědi

Hodnocení přesnosti prognóz

- pseudoprognóza se konstruuje tak, že k vyrovnání časové řady se nevyužije několik posledních hodnot řady, které jsou tak jako by „předpovídanými" hodnotami
- pro změření kvality skutečných předpovědí i pseudopředpovědí se používá Theilův koeficient nesouladu
  - relativní chyba extrapolace (%)
    - 0 – 5 % … chyba predikce malá
    - 5 – 10% … chyba predikce střední
    - > 10% … chyba predikce velká (model pro predikci nepoužívat)

Korelace a autokorelace časových řad

- **Existuje mezi dvěma časovými řadami závislost?**
- Zajímá nás, zda mezi 2 či více ukazateli v časových řadách existuje závislost (souvislost)
- Tj. souvislost, která by dovolovala vysvětlit změny v jedné časové řadě změnami ve druhé časové řadě, popř. v několika dalších časových řadách

Korelace časových řad

- Pro hodnocení příčinného vztahu mezi ČŘ se používají metody založené na měření těsnosti závislosti řad náhodné složky, tj. řad očištěných od trendu, popř. také od sezónní složky (jde o korelaci náhodné složky)
- Zkoumáme závislost mezi dvěma časovými řadami, z nichž hodnoty jedné ČŘ označíme symbolem y_t a druhé symbolem z_t (pro t = 1, 2, … n)
- **Odhadneme průběh trendu obou uvedených řad**. Tím dostaneme posloupnost odhadů vypočtených z trendové funkce
- Vytvoříme řady náhodných složek
- mezi kterými budeme provádět výpočet korelace.

Opožděná korelace časových řad

- Vliv určitého jevu se neprojeví ve stejném období, ale teprve **po uplynutí určitého času** (jednoho nebo i více období)
- Intenzitu opožděné korelace zkoumáme stejnými metodami jako v předchozím případě, pouze s tím rozdílem, že **posunujeme jednou časovou řadu** (závisle proměnnou) **o jedno nebo více období dále.**

Autokorelace časových řad

- Předpověď hodnot časové řady může být někdy zkomplikován jevem, kdy hodnoty ukazatele v řadě za sebou bývají vzájemně závislé
- **Autokorelace** – hodnoty časové řady jsou závislé na hodnotách této řady v předcházejících obdobích
  - **Autokorelace 1. řádu** – hodnota čř v čase t závisí na hodnotě řady v čase t – 1
  - **Autokorelace k-tého řádu** – y_t , hodnota řady v čase t závisí na y t-k (hodnotě v čase t-k)

- Např. Dnešní teplota vzduchu je závislá na teplotě včerejší, dnešní cena akcie se odvíjí od ceny včerejší
- Můžeme se setkat i s autokorekcí se záporným znaménkem, kdy dnešní vysoká hodnota vyvolává snížení příští hodnoty, např. nadbytečný nákup zásob v daném období způsobuje snížení nákupu v období příštím a naopak
- V matematickém modelu časových řad se autokorelace dat promítá do **autokorelace reziduí**
- **Příčina**: nevhodně zvolená trendová funkce
- **Důsledek:** odchylky empirických hodnot od teoretických správně nevystihují náhodnou složku časové řady, což má za následek, že reziduální odchylky nejsou v čase náhodně uspořádány
- Vhodným prostředkem ověření náhodnosti reziduí je např. **Durbin – Watsonův test autokorelace**
- Hodnoty statistiky se pohybují v intervalu od 0 do 4
  - **D = 2 řada bez autokorelace**
  - **D < 2 pozitivní autokorelace**
  - **D > 2 negativní autokorelace**

Kdy je korelace statisticky významná?

- Závisí na rozsahu výběrového souboru n a α
- tabulka kritických hodnot udává d_s a d_H.
- $D < d_s$ statisticky významná **kladná** autokorelace
- $d_s < D < d_H$ – test nerozhodnutelný
- $D > d_H$ rezidua nevykazují pozitivní autokorelaci
- Pro test negativní autokorelace použijeme tytéž hodnoty d_s a d_h odečtené od 4.
  - $D > 4 - d_s$ statisticky významná **záporná** autokorelace
  - $4 – d_s < 4 - d_H$ test nerozhodnutelný
  - $D > 4 – d_H$ rezidua nevykazují zápornou autokorelaci

Indexní analýza

- Důležitá součást analýzy sociálně ekonomických ukazatelů
- Hlediska srovnání:
  - **časové** (nejčastější srovnání roků, měsíců)
  - **prostorové** (srovnání geografických jednotek, organizačních celků)
  - **věcné** (srovnávané situace se liší např. druhem podnikatelské činnosti)

Způsob zjišťování

- **primární (prvotní) ukazatele** – neodvozené, přímo měřené (počet pracovníků k určitému dni, sklizeň určité plodiny)
- **sekundární (odvozené) ukazatele –** vznikají z primárních ukazatelů zpravidla jako jejich rozdíl či podíl (ha výnos, produktivita práce)

Hledisko dobry vyjádření

- **ukazatele okamžikové** – údaje zjišťované k určitému okamžiku (počet obyvatel k určitému datu)
- **ukazatele intervalové** – údaje sledované vždy za určité období (zisk za měsíc, dojivost za týden)

Hledisko povahy ukazatelů

- **ukazatele extenzitní – q** – vyjadřují rozsah, množství, počet nebo objem sledovaného jevu, získáváme je přímým měřením, vážením, lze je shrnovat pomocí součtů (počet pracovníků, maloobchodní obrat)
- **ukazatele intenzitní – p** – vyjadřují intenzitu nebo úroveň sledovaného jevu, jedná se o poměrové ukazatele (cena za jednotku, produktivita práce)
- **ceny – c**

Hledisko srovnání ukazatelů

ukazatele stejnorodé – prostý součet má pro daný celek smysl jako tentýž ukazatel za jednotlivé části celku (celkový objem těžby)

ukazatele nestejnorodé – prostý součet nedává smysl (sledujeme těžbu černého a hnědého uhlí)

Hledisko srovnatelnosti

- Schopnost ukazatele určit jeho celkovou hodnotu na základě jeho dílčích hodnot
  - **přímo srovnatelné**
  - **nepřímo srovnatelné**
  - **nesrovnatelné**

Hodnoty ukazatelů porovnáváme pomocí **rozdílu a podílu.**

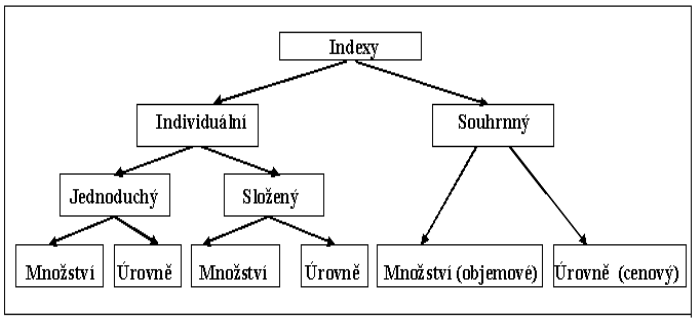
- **Absolutní změna** – rozdíl hodnot ukazatele **absolutní přírůstek**
- **Relativní změna** – podíl hodnot ukazatele **index**

Elementární srovnávání ukazatelů

$$I_{i/0} = \frac{q_i}{q_0}$$

- **Indexy bazické** – jsou vztaženy k pevně zvolenému období (bázi)
  - i … běžné období 0 … základní období
  - absolutní přírůstek Δ = qi – q0
- **Indexy řetězové** – jsou vztaženy k předcházejícímu období (koeficient růstu k)
  - $I_{i/i-1} = \frac{q_i}{q_{i-1}}$ i … běžné období i-1 … předcházející období
  - Absolutní přírůstek Δ = qi – qi-1

Druhy indexů



Individuální indexy

Jednoduché individuální indexy

- srovnáváme dvě hodnoty téhož ukazatele); (běžné období=1, základní období=0)

$$I^q = \frac{q_1}{q_0}$$

- **Jednoduchý index množství**, týká se extenzitních ukazatelů
- **Jednoduchý index úrovně**, týká se intenzitních ukazatelů

$$I^p = \frac{p_1}{p_0}$$

Složené individuální indexy

- **Složený index množství**, týká se extenzitních ukazatelů, shrnujeme pomocí součtů
- **Složený index úrovně**, týká se intenzitních ukazatelů, shrnujeme pomocí průměrů Inde proměnlivého složení IPS

Rozklad indexu proměnlivého složení

- Vyjadřujeme jako součin dvou indexů **IPS = ISS * ISTR**
  - **Indexu stálého složení ISS –** vliv změny intenzitní složky při konstantním působení složky extenzitní
  - **Index struktury ISTR** – vyjadřuje vliv změny extenzitní složky při konstantním působení složky intenzitní
- Váha jednotlivých složek mohou být fixována k období základnímu tak k období běžnému – dva způsoby rozkladu.

Souhrnné indexy

- Indexy **nestejnorodých extenzitních** ukazatelů
- Jsou nesouměřitelné, jejich součet nemá pro celek význam.
- Souměřitelnosti dosahujeme pomocí společných intenzitních ukazatelů – **souměřitelů – ceny**.
- **Souhrnné indexy úrovně – tzv. cenové indexy**
- **Souhrnné indexy množství – tzv. objemové index**

Souhrnné indexy úrovně

- **Laspeyresův index** – srovnává hodnotu produkce základního období oceněnou cenami běžného období s hodnotou téže produkce vyjádřenou v cenách základního období
- **Paascheho index** – váhou je množství z období běžného
- **Loweho index** – váhou je stále množství, množství vychází z reality nebo je uměle konstruováno např. jako průměrné = množství za uvažovaná časová období
- **Fischerův index** – geometrický průměr Laspeyresův index a Paascheho index

Souhrnné indexy množství

- Indexy nestejnorodých extenzivních ukazatelů. Vyjadřují změny objemu vytvořené či prodané produkce. Analogická konstrukce jako u cenových indexů
- **Laspeyresův index**
- **Paascheho index**
- **Loweho index**
- **Fischerův index**

Rozklad indexu

- **Index hodnotový = index cenový * index objemový**
  1. **způsob rozkladu**
  2. **způsob rozkladu**

Vícerozměrné statistické metody

- v případě, kdy místo jedné proměnné či dvojice proměnných statisticky zkoumáme a analyzujeme vztahy více proměnných zároveň a hledáme hlouběji vztahy mezi proměnnými, jde o vícerozměrné metody
- široké uplatnění metod: psychologie, sociologie potravinářství, genetika, zoologie, biologie, medicína, ekologie, atd.)

Základní skupiny metod

- metody analýzy korelačních struktur
  - vícenásobná regresní a korelační analýza
  - analýza hlavních komponent
  - faktorová analýza (latentní proměnné)
- metody srovnávání skupin nebo ošetření
  - vícenásobná analýza rozptylu (MANOVA)
  - kanonická korelace
- metody klasifikace objektů do existujících skupin
  - diskriminační analýza
  - logistická regrese
- metody analýzy podobnosti mezi objekty
  - shluková analýza
  - mnohorozměrné škálování

Metody analýzy korelačních struktur

Vícenásobná regresní a korelační analýza

- regresní model $Y = f(X1, X2, \ldots, Xn) + \varepsilon$
- korelační struktura – párové korelační koeficienty, parciální korelační koeficienty, korelační index

Analýza hlavních komponent

- Principal component analysis – PCA
- Patří mezi nejstarší a nejpoužívanější metody vícerozměrné analýzy. Zavedena Pearsonem v roce 1901 a nezávisle na tom Hotellingem v roce 1933
- Vlastnosti objektů jsou popsány pomocí velkého poštu proměnných X1, X2, … Xp
- Hlavní cíle
  - Zjistění vazeb mezi proměnnými
  - Redukce počtu proměnných a nalezení nových smysluplných proměnných
- Podstatou PCA je lineární transformace původních proměnných do menšího počtu nových fiktivních proměnných, tzv. hlavních komponent
- Vlastnosti komponent:
  - Jsou vzájemně nekorelované
  - Metoda je založena na bezezbytkovém vysvětlení celkové bariability
  - Hlavní komponeny jsou uspořádány podle velikosti vysvětleného rozptylu, nejvíce informace o variabilitě proměnných je v první komponentě, nejméně v poslední

Postup při PCA

- Počáteční analýza – zkoumáme vztahy mezi proměnnými (grafické znázornění, popisné statistiky)
- Průzkum korelační matice (redukce proměnných je možná, pokud existují významné korelace mezi původními proměnnými
- Analýza hlavních komponent, stanovení vhodného počtu hlavních komponent
- Interpretace hlavních komponent

Faktorová analýza (FA)

- Vznikla v psychologii, za zakladatele je považován Charles Edward Spearman (1904)
- Vícerozměrná analýza určená k redukci počtu proměnných, na rozdíl od PCA se snažíme vysvětlit vzájemné závislosti proměnných

- Hlavní cíle FA
  - Redukce rozsahu dat při co nejmenší ztrátě informace – nalezení menšího počtu fiktivních proměnných (faktorů) za účelem vysvětlit napozorované korelace
  - Nalezení věcného významu faktorů (interpretace)

- Předpokládáme, že každou z původních proměnných můžeme vyjádřit jako lineární kombinaci několika společných přímo neměřitelných faktorů a jedinečnosti proměnné
- Jedná se o soustavu lineárních rovnic, ve které jsou skutečné proměnné x1, x2, … xp vyjádřeny pomoc í fiktivních proměnných, a to tzv. společných faktorů
- Počet faktorů by měl být co nejmenší a nalezené závislosti by měly být vysvětleny co nejjednodušeji
- Pro interpretaci je vhodnější ta varianta, kdy každý faktor vysvětluje rozdílné skupiny vzájemně korelovaných původních znaků >> rotace faktorů
- Snahou je, aby každá proměnná dosahovala vysoké faktorové zátěže (hodnoty blízké +- 1) pouze u jediného faktoru
- Rotace faktorů - Rotace os X1 a X2 tak, abychom vystihli směr, ve kterém je variabilita dat maximální (osa Y1), při projekci na osu Y2 bude variabilita dat malá
- Data by měla být v ideálním případě spojitá (metrická), pozvolně se zvětšující nebo zmenšující, s možností lineárně je kombinovat, např. věk nebo teplota (od 0 po 100 a víc), hmotnost tělesa (od 0 po n kg), nebo vzdálenost v km
- V odborné literatuře se zpravidla v současnosti připouští změkčení těchto požadavků
- V praxi ve společenských vědách totiž taková data bývají zřídkakdy k dispozici. Musíme obvykle vystačit i s kategoriálními nebo metrickými diskrétními daty
- Dále je nevhodný příliš malý počet případů, tj. malé výběrové soubory respondentů
- FA vychází z korelací a ty se stabilizují teprve na poměrně velkých výběrových souborech (50 respondentů je velmi špatné, 100 je špatné, 200 ujde, 300 je dobré, 500 je velmi dobré a 1000+ je skvělé)
- Někteří autoři uvádějí, že respondentů (případů) musí být minimálně 5x víc než proměnných pro FA

Diskriminační analýza

- Diskriminační analýza vychází z předběžného určení počtu skupin a jejich konkrétního definování
- Jejím úkolem je nalezení takových proměnných, které mají největší schopnost rozlišit, do které ze skupin objekt patří
- Diskriminační analýza umožňuje hodnocení rozdílů mezi dvěma nebo více skupinami objektů charakterizovaných více znaky – diskriminátory
- Obyčejně se dále dělí na techniky, které interpretují rozdíly mezi předem stanovenými skupinami objektů a techniky, kde je cílem klasifikace objektů do skupin

Shluková analýza (Cluster analysis – CA)

- Cílem je nalézt v celém souboru skupiny objektů, které jsou si navzájem podobné, ale které se zároveň lidí od ostatních skupin
- Aplikace: segmentace trhu, diferenciace produktů (nabídka), typologie (hledání fenotypů – typických znaků)
- Metody shlukové analýzy
  - Hierarchické
  - Nehierarchické

Hierarchické metody

- Podstatou je tvorba shluků různé úrovně (shluky nejvyšší úrovně obsahují shluky nižší úrovně)
- Hierarchické metody vedou ke stromové struktuře, výsledky se znázorňují stromovým grafem – dendrogramem vhodné pro soubory o rozsahu do cca 150 jednotek
- Vyjádření podobnosti u kvalitativních znaků
  - Počet shodných znaků/počet všech sledovaných znaků
- Vyjádření podobnosti u kvantitativních znaků se nejčastěji používá
  - Eukleidovská vzdálenost/čtvercová Euklidovská vzdálenost
  - Vzdálenost Manhattan (Hemmningova vzdálenost)
  - Čebyševova vzdálenost
- Nestejné jednotky >> standardizace (z-skóre)

Shlukovací algoritmy hierarchické metody

- **Nejbližší soused (jednoduché spojení):** vzdálenost mezi dvěma shluky je definována jako vzdálenost dvou nejbližších členů
- **Nejvzdálenější soused (úplné spojení**): vzdálenost mezi dvěma shluky je definována jako vzdálenost dvou nejvzdálenějších členů
- **Nevážený párový průměr**: vzdálenost mezi dvěma shluky je definována jako průměrná vzdálenost mezi všemi páry, přičemž 1. Člen je z 1. Shluku a 2. Člen je z 2. Shluku
- **Vážený párový průměr**: jako předchozí, navíc se zohledňují velikosti shluků (počty objektů) jako váhy
- **Nevážený centroid**: vzdálenost mezi dvěma shluky definována jako vzdálenost centroidů (těžišť) těchto shluků. Centroid je vektor průměrů (každá souřadnice je průměr příslušných souřadnic objektů ve shluku)
- **Vážený centroid (medián**): jako předchozí, navíc se zohledňují velikosti shluků (počty objektů) jako váhy
- **Wardova metoda**: odlišná od předchozích, na určení vzdálenosti mezi shluky se používá přístup analýzy rozptylu. Shluky se tvoří tak, aby se minimalizoval vnitroshlukový součet čtverců

Nehierarchické metody

- K-průměrů (K-means) algoritmus je založen na přesouvání objektů mezi shluky, počet shluků je předem dán, buď náhodně, nebo na základě zkušeností
- Metoda K-means je vhodná pro shlukování většího množství objektů

Mnohorozměrné škálování

- Účelem mnohorozměrného škálování (MDS) je vyjádřit odlehlost objektů, popř. názory na podobnost objektů nebo preference objektů jako vzdálenosti (za objekty jsou považovány názory, podněty, produkty, uchazeči o něco, volby atd.)
- Tyto vzdálenosti jsou pak zobrazeny v mnohorozměrném prostoru