Studijní opory předmětu MT 109 APLIKOVANÁ STATISTIKA v kombinovaném studiu Vysoké školy hotelové a ekonomické s.r.o., magisterský studijní program všech oborů, ISBN: 978-80-87411-45-2

Předmět MT109 APLIKOVANÁ STATISTIKY je určen studentům kombinovaného studia všech oborů VŠHE.

Výuka předmětu "MT 109 „ Aplikovaná statistika“ v kombinovaném studiu
Výuka probíhá ve třech modulech, celkem 18 hodin, každý modul je pětihodinový (6 - 6 - 6).
Formou atestace je zkouška (5 kreditů).

Přednášející, cvičící: Dr. Ing. Sylva Skupinová
Zkoušející: Dr. Ing. Sylva Skupinová

Obsahová náplň předmětu MT109 Aplikovaná statistika

1. Vyhodnocování údajů získaných náhodným výběrem náhodným výběrem
2. Statistické testy
3. Analýza časových řad
4. Exponenciální trendy, exponenciální vyrovnání
5. Regresní a korelační analýza
6. Metoda nejmenších čtverců – další aplikace
7. Vícenásobná regrese a korelace
8. Korelace časových řad, opožděná korelace, autokorelace
9. Regresní modely
10. Lineární trendy v časových řadách
11. Sezónní časové
12. Metody vícerozměrné statistické analýzy
13. Statistické zajímavosti, historie statistiky, Český statistický úřad

Cíle výuky předmětu MT109
Studenti budou seznámeni se základními i nadstavbovými matematickými a statistickými operacemi a postupy používanými v ekonomické a hospodářské praxi.
Předmět je zaměřen rovněž na problémy vznikající při aplikacích těchto metod.
Získané poznatky studenti využijí v navazujících předmětech magisterského studia popřípadě při zpracování dat v diplomové práci.
Osvojené postupy z oblasti aplikované statistiky umožní studentovi pochopit základní principy ekonomických modelů používaných v praxi.

Předmět MT109 bezprostředně navazuje na předmět MT005 z bakalářského studia a u studentů předpokládá znalost látky, která je obsažena ve skriptech VŠHE: I. Novák, Statistika.

Požadavky ke zkoušce
Předmět MT 109 „Aplikovaná statistika“ je ukončen písemnou a ústní zkouškou. Předpokladem pro její složení je:

- aktivní účast na výuce v jednotlivých modulech (soustředění)
- prostudování základní literatury a studijních opor
- splnění korespondenčních úkolů
- úspěšné absolvování závěrečných testů a ústní části zkoušky

Organizace studia
Výuka předmětu MT 109 „Aplikovaná statistika" (semestrální kurz) je rozdělena na kontaktní a distanční část a probíhá ve třech modulech. Kontaktní výuka (18 hodin) je realizována v rámci tří soustředění, jde o 6 + 6 + 6 hodin přímé výuky. V každém soustředění se uskuteční výuka jednoho modulu, který tvoří dvě povinné části: **"tutoriál"** a **"průvodce studiem".**
Převážná část kombinovaného studia předmětu MT 109 má sice distanční formu, avšak z hlediska pedagogického přístupu ke studentům a jejich možnostem spolupracovat s vyučujícím (tutorem), jde o průběžnou výuku. Na tutoriálech a ve studijních materiálech jsou zadávány úkoly, jejichž splněním student dokládá průběžnost svého studia. Komunikace s vyučujícím je zajištěna přes Internet (skupinova@vsh.cz) a v průběhu semestru může student navštívit konzultační hodiny učitele. V případě problémového tématu má možnost navštívit přednášky či semináře prezenčního studia. Pokud mu nestačí konzultace telefonická či prostřednictvím výukového prostředí (IS VŠHE), může si student domluvit individuální (event. kolektivní) konzultaci. Administrativu studia zajišťuje příslušná referentka studijního oddělení. Všechny kontakty mezi učitelem a studujícím probíhají v rámci informačního systému VŠHE.

Časový harmonogram výuky a obsahové zaměření modulů část statistika:
1. modul (září) = Vyhodnocování údajů získaných náhodným výběrem (téma 1 - 4)
2. modul (listopad) = Regresní a korelační analýza (téma 5 - 9)
3. modul (leden) = Lineární trendy v časových řadách; Metody vícerozměrné statistické analýzy (téma 10 - 12)

Tutoriály:
Na **úvodním tutoriálu** na začátku semestru jsou studenti seznámeni, v rámci tzv. průvodce kurzu, s obsahem předmětu, s časovým rozvržením výuky jednotlivých tématických okruhů, s místem předmětu ve studijním plánu oboru, s povinnou literaturou, cílem výuky a s požadavky ke zkoušce. Je zde vysvětlen přístup k tzv. studijním oporám (studijní materiály, metodické listy) a způsob odevzdávání kontrolních úkolů (testů) v informačním systému VŠHE. Studentům je objasněn způsob hodnocení kontrolních úkolů a termíny jejich odevzdávání. Je probrána celková organizace výuky.
Na **průběžném tutoriálu** (uprostřed semestru) učitel vyhodnocuje dosavadní práci studentů. Studenti musí zaslat vyřešené úkoly elektronicky před zahájením týdne konzultací. Učitel upozorní na závažné nedostatky a v případě potřeby obtížná témata vysvětlí.
Na **závěrečném tutoriálu** na konci semestru učitel vyhodnotí uložené úkoly z minulého tutoriálu a práci studentů za celý semestr. Upozorní na problémové otázky tématických okruhů ke zkoušce. Podle potřeby proběhne společná konzultace. Studenti jsou seznámeni s časovým harmonogramem zkoušek.

Průvodce studiem:
V této kontaktní části studia je proveden metodický výklad (přednáška) daného tématického celku. Studenti jsou seznámeni s tím, co budou studovat z povinné literatury (musí být k dispozici pro studenty), jaká úskalí je čekají při samostudiu a jak jim bude učitel pomáhat při studiu. Velká pozornost je věnována jejich práci se studijními oporami, které jim nahrazují bezprostřední kontakt s vyučujícím na cvičeních (seminářích). Studijní opory jsou připraveny pro každý tématický okruh (kapitolu učebnice). Jejich součástí jsou: cíle, úvod, vlastní výklad tématu, shrnutí vyložené problematiky, klíčové pojmy, úkoly k zopakování a procvičení, odkazy na další studijní zdroje a hodnocení. Studijní opory jsou vloženy v rámci IS VŠHE do části **studijní materiály předmětu** MT109. Zpětnovazební prvky výuky (korespondenční

úkoly) vyučující vkládají v informačním systému do položky odpovědníky. Jejich zadání musí být jednoznačné a nesmí umožňovat různá řešení (pokud to ale není záměr vyučujícího). Vypracované úkoly studenti vkládají do **odevzdavárny**, event. přímo vyučujícímu.
Při studiu předmětu MT109 student využívá tři **informační zdroje**:

- metodologický výklad učitele, který vychází z předepsané literatury
- kontaktní výuku v rámci tutoriálu a samostudia;
- předepsanou literaturu a metodické materiály

Průvodce studiem jednotlivých MODULŮ

Studijní opory předmětu MT 109 APLIKOVANÁ STATISTIKA v kombinovaném studiu Vysoké školy hotelové a ekonomické s.r.o., magisterský studijní program všech oborů

Studijní literatura

Základní:
Skupinová S.: Aplikovaná statistika. Vysoká škola hotelová v Praze 8, Praha 2012, ISBN 978-80-87411-42-1.
Marek, L.; Novák, I.; Vrabec, M.: Statistika II. Vysoká škola hotelová v Praze 8, Praha, 2004, 90 stran, ISBN 80-86578-30-5
Doporučená:
Pecáková, I., Novák, I., Herzmann, J.:Pořizování a vyhodnocování dat. VŠE Praha, 2004,Oeconomica ISBN 80-245-0753-6
Hindls, R., Hronová, S., Novák, I.: Analýza dat v manažérském rozhodování. VŠE Praha, 1999 ,Grada, ISBN 80-7169-255-7
Hindls, R. a kol.: Statistika pro ekonomy. Professional Publishing, Praha 2007.

1. Modul

Modul tvoří tři tématické okruhy. Každý je probírán samostatně, jako kapitola v učebním materiálu.

Tématické okruhy:

1.1. Statistické odhady
1.2. Statistické testy
1.3. Analýza časových řad

Studijní cíle

V této kapitole se studenti seznámí se základními postupy při vyhodnocování údajů získaných náhodným výběrem. Bude objasněna teorie bodových a intervalových odhadů s důrazem na symetrické oboustranné intervaly. Dále budou studenti seznámeni s typy alternativ a statistickým testováním. Poslední tematický okruh seznámí studenty s dekompozicí časových řad a s jejími elementárními složkami s důrazem na exponenciální trendy v časových řadách.

Klíčová slova: bodový odhad, intervalový odhad, spolehlivost odhadu, statistický test, typy alternativ, testovaná a alternativní hypotéza, testové kritérium, časové řady, exponenciála, exponenciální vyrovnání

1.1. Statistické odhady

Odhady charakteristik základního souboru :

- **bodové** - jedna číselná hodnota

(průměr základního souboru μ se bodově odhaduje výběrovým průměrem) $\bar{x}$

- **intervalové** - interval hodnot

Jedná se o odhady charakteristik základního souboru takovými intervaly, v nichž lze se **zvolenou pravděpodobností** očekávat hodnoty očekávaných charakteristik.
Zvolená pravděpodobnost = spolehlivost odhadu a značí se **1-α**.

Příklad: Byla zvolena spolehlivost 95%
Hovoříme pak o 95% spolehlivosti, nebo že příslušný interval je 95%ním intervalem spolehlivosti, kdy platí, že **1-α = 0,95.** Pak existuje 5% riziko, že intervalový odhad bude chybný, tj. že hodnota odhadované charakteristiky bude mimo udaný interval.

Odhady relativní četnosti

Bodovým odhadem relativní četnosti v základním souboru Π je výběrová relativní četnost **p**.
Pro **zvolenou spolehlivost** odhadu **1-α** (například 95%) a je-li **np(1-p)>9** (což bývá při velkých výběrech obvykle splněno), je dvoustranný interval spolehlivosti vymezen nerovností:

$$p - \Delta \le \Pi \le p + \Delta, \text{ kde} \qquad \Delta = u_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

$u_{1-\frac{\alpha}{2}}$ - kvantil normovaného normálního rozdělení.

Odhady aritmetického průměru

Bodovým odhadem průměru základního souboru μ je výběrový průměr $\bar{x}$. Dvoustranný symetrický interval spolehlivosti je pak při větších výběrech (již kolem 100 jednotek a větších) vymezen nerovností:

$$\bar{x} - \Delta \le \mu \le \bar{x} + \Delta, \text{ kde} \qquad \Delta = u_{1-\frac{\alpha}{2}} \frac{S_x}{\sqrt{n}}$$

x - určitá proměnná
n - rozsah souboru
$\bar{x}$ - aritmetický průměr výběrového souboru
μ - aritmetický průměr základního souboru
S_x - výběrová směrodatná odchylka
$u_{1-\frac{\alpha}{2}}$ - kvantil normovaného normálního rozdělení.

1.2. Statistické testy

Při běžně používaných testech se proti sobě staví dvě hypotézy:
testovaná hypotéza - H_0 x **alternativní hypotéza** - H_1.
Testovaná hypotéza něco tvrdí a alternativní hypotéza to popírá. Výsledkem statistického testu je buď přijetí testované hypotézy nebo její zamítnutí, tj. přijetí alternativní hypotézy.

Chyby ve statistickém testování

Chybně může být přijata jak alternativní tak testovaná hypotéza.

Hladina významnosti - α je pravděpodobnost chybného přijetí alternativní hypotézy – chyba prvního druhu .
Hladina významnosti α se **volí**, nejčastěji $\alpha = 0,05$,tj. 5ti% hladina významnosti, tedy volíme 5ti% riziko, že na základě výběrových dat chybně přijmeme alternativní hypotézu.

Pravděpodobnost chybného přijetí testované hypotézy, tj. chyba druhého druhu β, u většiny testů **nelze** volit.

Testové kritérium

K rozhodnutí o přijetí nebo zamítnutí testované hypotézy, slouží při každém testu určitá proměnná, jejíž hodnotu lze vypočítat z výběrových dat a která má při platnosti testované hypotézy určité pravděpodobnostní rozdělení. Tato proměnná se nazývá **testové kritérium** (testová statistika).

Obor hodnot testového kritéria se rozděluje na **obor přijetí** a **kritický obor**. Hodnoty, které tyto obory od sebe oddělují se nazývají **kritické hodnoty** a lze je vyhledat ve statistických **tabulkách**.

Jestliže **hodnota testového kritéria**, vypočítaná z výběrových dat padne do **oboru přijetí**, **přijímá se testovaná hypotéza**. Jestliže vypočítaná hodnota padne do kritického oboru, zamítá se H_0 a přijímá se H_1.
Kritický obor je volen tak, aby pravděpodobnost, že hodnota testovaného kritéria padne do kritického oboru při platnosti testované hypotézy, byla rovna zvolené hladině **významnosti α**.

Výsledek testu závisí mimo jiné na zvolené hladině významnosti α, která se volí. Aby nemohlo dojít k nedorozumění, je nutné u každého testu použitou hladinu významnosti uvést. **Čím nižší je hladina významnosti, tím je kritický obor užší.**

Typy alternativ statistických testů:
- dvoustranná alternativa
- levostranná alternativa
- pravostranná alternativa.

Některé statistické testy používané v marketingových výzkumech

Všechny tyto testy vycházejí z předpokladu, že výběrový soubor je náhodným výběrem z nekonečného základního souboru nebo tzv. prostým náhodným výběrem z konečného základního souboru, jehož rozsah je mnohonásobně větší než rozsah výběrového souboru.

Každý statistický test je použitelný pouze za určitých podmínek. Nejsou-li tyto podmínky splněny, může vést jeho použití k dezinformaci.

Test hypotézy o relativní četnosti při velkém výběru (v základním souboru)

Relativní četnost základního souboru - Π
Formulace testované hypotézy H_1 závisí na tom, co chceme prokázat.

Hodnota Π_0 je hypotetická hodnotou relativní četnosti Π. Je to hodnota, kterou předpokládá testovaná hypotéza. Testovaná hypotéza se obvykle vyjadřuje zápisem:
H_0: $\Pi=\Pi_0$
Proti testované hypotéze lze podle povahy problému postavit alternativní hypotézu:

- pravostrannou H_1: $\Pi > \Pi_0$
- levostrannou H_1: $\Pi < \Pi_0$
- dvoustrannou H_1: $\Pi \neq \Pi$

Je-li rozsah výběru **n** velký používá se testové kritérium:

$$U = \frac{(p - \Pi_0)\sqrt{n}}{\sqrt{\Pi_0(1-\Pi_0)}}$$

Podmínka: součin $n\Pi_0(1-\Pi_0)$ musí být větší než 9.

U - hodnota testového kritéria
Π_0 - hypotetická hodnotou relativní četnosti Π
p - výběrová četnost
n - rozsah souboru

Vymezení kritického oboru*:

Při testu hypotézy H_0 proti pravostranné alternativní hypotéze je kritický obor vymezen nerovností: $U > u_{1-\alpha}$

Při testu hypotézy H_0 proti levostranné alternativní hypotéze je kritický obor vymezen nerovností: $U < u_{1-\alpha}$

Při testu hypotézy H_0 proti dvoustranné alternativní hypotéze je kritický obor vymezen nerovností: $/U/ > u_{1-\frac{\alpha}{2}}$

/U/ - absolutní hodnota testového kritéria
$u_{1-\alpha}$, $u_{1-\frac{\alpha}{2}}$ - kvantily normovaného normálního rozdělení .

Test hypotézy o průměru při velkém výběru:

Při testech hypotéz o průměru μ základního souboru se ověřují hypotézy, že tento průměr je větší, menší případně jiný než hypotetická hodnota μ_0.

Je-li rozsah výběru dostatečně velký (n>100), lze použít testové kritérium:

$$U = \frac{(\bar{x} - \mu_0)\sqrt{n}}{S_x}$$

$\bar{x}$ - výběrový průměr
S_x - výběrová směrodatná odchylka.

Vymezení kritického oboru je shodné s výše uvedenou definicí označenou symbolem *.

Hodnota μ_0 je hypotetická hodnotou aritmetického průměru μ. Je to hodnota, kterou předpokládá testovaná hypotéza. Testovaná hypotéza se obvykle vyjadřuje zápisem:
H_0: $\mu = \mu_0$.
V závislosti na formulaci alternativní hypotézy, lze použít některý z kritických oborů:

- pravostrannou H_1: $\mu > \mu_0$
- levostrannou H_1: $\mu < \mu_0$
- dvoustrannou H_1: $\mu \neq \mu_0$.

χ^2 test - test dobré shody:

χ^2 – test umožňuje ověření platnosti hypotézy H_0 „náhodný výběr pochází z daného rozdělení“ $\Rightarrow$ ověření hypotézy o rozdělení v základním souboru.

- H_0 – rozdělení je určitého typu
- H_1 – rozdělení je jiného typu, ale nelze specifikovat jakého.

χ^2 – test s výhodou aplikujeme při výzkumech veřejného mínění a v marketingu.

Mac Nemarův test změny názorů:

Názory dotazovaných osob na řešenou problematiku se mohou, pod vlivem určité informace (reklamní kampaň, vyjádření odborníků) nebo po provedení určitého opatření, měnit. **Cílem** testu je posoudit, zda došlo ke změně názoru v základním souboru.
H_0: tvrdí, že názory se nezměnily = nedošlo ke změně
H_1: tvrdí, že došlo ke změně, ale neříká, zda k lepšímu či k horšímu (toto lze odhadnout ze zdrojových dat).
Test se s výhodou využívá při posuzování účinnosti reklamy. Pro Mac Nemarův test se používá testové kritérium, které má při platnosti testované hypotézy přibližně χ^2 rozdělní o jednom stupni volnosti.

1.3. Analýza časových řad

Časová řada = vývojová tendence. Jedná se o zásadní a neoddělitelnou analytickou práci v ekonomické oblasti.
Předpoklad: existují data různých ukazatelů v časové řadě.
Odhad budoucích hodnot = **extrapolace** časové řady.
Prognózy do vzdálené budoucnosti předpokládají neměnný trend.
Při analýze časových řad je nutné vyžadovat věcnou, prostorovou a časovou srovnatelnost údajů. Srovnatelnost údajů je vždy nutno před jejich statistickou analýzou prověřit!

Délka časové řady se volí v závislosti na kvalitě vstupních dat.

Dekompozice časových řad

Pro dekompozice časových řad je nutné uvažovat následující předpoklad:
časovou řadu lze rozložit na systematické (a odhadnutelné) složky a na náhodnou složku.

Systematické složky:
- trendová
- sezónní
- cyklická složka.

Trendová složka - odráží dlouhodobou vývojovou tendenci (například zrychlující či zpomalující se růst či pokles), kterou lze popsat nějakou matematickou funkcí (tzv. trendovou funkcí).
Sezónní složka - popisuje pravidelně se opakující výkyvy v jednotlivých sezónách (například čtvrtletích či měsících) několika po sobě jdoucích let.
Cyklická složka - popisuje dlouhodobé výkyvy kolem trendu, tedy výkyvy opakující se vždy po několika letech.

Elementární charakteristiky časových řad

Ze zjištěných dat se velmi často počítají **roční přírůstky** a **roční koeficienty růstů**.
Velmi často se počítá i průměrný roční koeficient růstu, který je **geometrickým průměrem** jednotlivých koeficientů růstu: $\bar{k}$

$\bar{k} = \sqrt[n-1]{k_2.k_3.....k_n}$, kde $k_2 - k_n$ jsou roční koeficienty růstu.

Popis trendu časových řad ročních hodnot:
Trendové funkce jsou různé matematické funkce, kde platí následující předpoklad:
v časové řadě se projevuje pouze určitý trend a náhodné kolísání. Pak pro hodnoty časové řady platí:

$$Y_t = T_t + e_t$$

t = 1,2, …, n
T_t - je odhad trendové složky
e_t - je reziduum.

Hodnoty T_t jsou hodnotami trendové funkce T = f(t), kde f(t) je nějaká matematická funkce časové proměnné t. Může to být například přímka ($T = b_0 + b_1 t$),
hyperbola $T = b_0 + b_1 \frac{1}{t}$,
parabola ($T = b_0 + b_1 t + b_2 t^2$) aj.

b_0, b_1, b_2 - parametry, jejichž číselné hodnoty je třeba určit, aby bylo možno využít trendové funkce k odhadům do budoucna.

Při popisu trendu matematickými funkcemi jsou řešeny dvě otázky:

a) Jaká matematická funkce má být zvolena za trendovou funkci?

Podkladem pro volbu vhodné trendové funkce bývá obvykle chování některých elementárních charakteristik časové řady:

- roční absolutní přírůstek
- roční relativní přírůstek
- roční koeficienty růstu ...

Vodítkem volby matematické funkce popisující trend je grafické znázornění časové řady spojnicovým diagramem. Kvalita zvolené matematické funkce se ověřuje výpočtem reziduí.

b) Jak se určí číselné hodnoty parametrů zvolené trendové funkce?

U lineárních trendových funkcí (např. přímka, hyperbola parabola), se číselné hodnoty parametrů určují **metodou nejmenších čtverců**. Minimalizuje se součet druhých mocnin odchylek zjištěných hodnot y_t od zvolené trendové funkce:

$$S = \sum_t (y_t - T_t)^2$$

Pro výpočet parametrů nelineárních funkcí (např. exponenciála, posunutá exponenciá) **nelze** použít metodu nejmenších čtverců!
Parametry některých nelineárních funkcí lze získat metodou nejmenších čtverců až po provedení tzv. **linearizující transformace**, kdy sledovaná proměnná y je nahrazena nějakou neparametrickou funkcí y^*, například $y^* = \ln y$. Uvedená transformace se využívá například při výpočtu parametrů exponenciály $T = b_0 b_1^t$
Má-li časová řada exponenciální trend je pro ní typické stálé roční tempo růstu.

U posunuté exponenciály se využívá metoda **částečných součtů**, kdy se sledovaná časová řada rozdělí na několik částí o stejném počtu hodnot, přičemž počet částí je rovný počtu parametrů trendové funkce.

Posunutá = modifikovaná exponenciála $T = b_0 + b_1 b_2^t$

Tato funkce může vystihnout zrychlující se či zpomalující se rostoucí trend i zrychlující se či zpomalující se klesající trend.

Exponenciální vyrovnání:
Exponenciální vyrovnání je založeno na myšlence, že pro krátkodobé prognózy jsou čerstvější hodnoty časové řady důležitější než hodnoty starší. Máme-li časovou řadu ročních hodnot y_1, $y_2,\ldots y_{n-1}$, y_n, přisuzuje se největší váha hodnotě y_n, zatímco váhy ostatních hodnot postupně klesají ve směru k hodnotě y_1.
Stáří jednotlivých pozorování je vyjádřeno **proměnnou k**, která nabývá hodnot 0, 1,…n-1. Čím starší ročník tím je hodnota k vyšší. Nejmladší ročník časové řady má k = 0.
Váhu jednotlivých hodnot časové řady vyjadřujeme čísly α^k, kde α je **vyrovnávací konstanta, kde** $\alpha \in (0; 1)$. S klesající hodnotou α mají starší hodnoty menší význam!
Obvykle není zájem na příliš rychlém poklesu vah starších hodnot, volí se zpravidla hodnoty vyrovnávací konstanty α bližší 1 (α = 0,7 - 0,9).

Existují tři varianty exponenciálního vyrovnání časových řad ročních hodnot :

- ✓ Jednoduché exponenciální vyrovnání, kde se předpokládá, že v krátkých obdobích **nemá** časová řada **ani rostoucí ani klesající** trend.

- ✓ Dvojité exponenciální vyrovnání, kde se předpokládá, že v krátkých obdobích **má** časová řada **lineární trend** popsatelný přímkou.
- ✓ Trojité exponenciální vyrovnání, kde se předpokládá, že v krátkých období **má** časová řada **parabolický trend**.

Nejčastěji se používá dvojité exponenciální vyrovnání.

Chyba prognózy Δ
Po uplynutí roku n + 1 (nejmladší rok časové řady) zjistíme skutečnou hodnotu sledovaného ukazatele v tomto roce, tj. hodnotu y_{n+1}. Pak můžeme vypočítat chybu prognózy P_{n+1}:

$$\Delta_{n+1} = P_{n+1} - y_{n+1}$$

Je-li Δ kladné číslo, prognóza byla nadhodnocena, je-li Δ záporné číslo, prognóza byla podhodnocena.

Adaptivní metoda:
Na základě nově zjištěných skutečností, lze jednoduše opravovat hodnoty parametrů trendové funkce.

Shrnutí kapitoly
V kapitole byly vysvětleny základní pojmy z oblasti statistických odhadů, statistických testů a analýzy časových řad. Byly představeny bodové a intervalové odhady aritmetického průměru a relativní četnosti v základním souboru. Dále byly vysvětleny základní statistické hypotézy v rámci sadistického testování a možné typy alternativ. Byly popsány nejběžněji používané typy alternativ v testech hypotézy o relativní četnosti a průměru v základním souboru. V poslední části se problematika zaměřovala na analýzu časových řad, konkrétně na dekompozici časových řad a popis odhadnutelné složky matematickými funkcemi. Byl kladen důraz na exponenciální trendy v časových řadách a rovněž využití exponenciálního vyrovnání. Závěr kapitoly byl věnován výpočtu a chybě prognózy.

Pojmy k zapamatování:
Bodový a intervalový odhad, spolehlivost, typy alternativ, statistické testy, hladina významnosti, testovaná a alternativní hypotéza, testové kritérium, kvantily normálního a jiných rozdělení, časová řada, dekompozice časové řady, trendová funkce, typy trendů, exponenciální trend, nelineární funkce, exponenciální vyrovnání, prognóza, chyba prognózy, adaptivní metoda.

Úkoly k zopakování a procvičení

Příklad 1.1.:
Spolehlivost odhadu označujeme symbolem:
a) α
b) 1 - α
c) s
Řešení: b

Bodovým odhadem relativní četnosti základního souboru Π je:
a) relativní četnost výběrového souboru p
b) aritmetický průměr základního souboru μ
c) směrodatná odchylka základního souboru σ
Řešení: a

Při výpočtu intervalového odhadu aritmetického průměru základního souboru je nutné mít k dispozici mimo jiné hodnoty:
a) rozptylu základního souboru
b) směrodatné odchylky základního souboru
c) směrodatné odchylky výběrového souboru
Řešení: c

Příklad 1.2.:
Při běžně používaných testech se proti sobě staví dvě hypotézy:
a) pravostranná a levostranná
b) oboustranná a nestranná
c) nulová a alternativní
Řešení c

Hladina významnosti $\alpha = 0,01$ vypovídá:
a) o 99% riziku nesprávného přijetí alternativní hypotézy
b) o 1% riziku nesprávného přijetí alternativní hypotézy
c) o 1% riziku nesprávného přijetí nulové hypotézy
Řešení: b

Příklad 1.3.:
Mezi systematické složky časové řady nepatří složka:
a) cyklická
b) neodhadnutelná
c) trendová
Řešení: b

Parabola je trendová funkce:
a) s dvěma proměnnými a třemi parametry
b) s dvěma proměnnými a dvěma parametry
c) se třemi proměnnými a dvěma parametry
Řešení: a

Je li prognóza podhodnocena, pak je chyby prognózy z intervalu
a) $(-\infty; 0)$
b) $(0; \infty)$
c) $\langle -\infty; 0 \rangle$
Řešení: a

Hodnocení
Každá správná odpověď nebo výsledek výpočtu je hodnoceno jedním bodem.
Sebehodnocením je žádoucí dosáhnout alespoň 70% úspěšnost správných odpovědí. Jestliže jste nedosáhli požadované úspěšnosti, pokuste se zlepšit svůj studijní výsledek pozornějším studiem kapitoly, popřípadě se spojit s tutorem předmětu.

Korespondenční úkol:
Náhodně bylo vybráno 400 osob, z nichž 80 uvedlo, že má zájem o novou službu. Se spolehlivostí odhadu 95% proveďte intervalový odhad relativní četnosti Π základního souboru zájemců o novou službu.

Řešení: $\Pi \in \langle 0{,}1608;\ 0{,}2392 \rangle$.

2. Modul

Modul tvoří tři tematické okruhy. Každý je probírán samostatně, jako kapitola v učebním materiálu.

Tématické okruhy:

2.1. Regresní a korelační analýza
2.2. Vícenásobná regrese a korelace
2.3. Regresní modely

Studijní cíle

V této kapitole se studenti seznámí s terminologií a metodickými postupy regrese a korelace. Budou seznámeni s jednoduchou jednostrannou korelací se vzájemnou regresí a korelací a s vícenásobnou regresí a korelací. Dále budou studenti seznámeni s charakteristikami, které hodnotí kvalitu regresního modelu a s koeficienty hodnotícími těsnost studovaných závislostí včetně vícenásobného korelačního koeficientu. Studenti se rovněž seznámí s multikolinearitou, opožděnou korelací, korelací v časových řadách a autokorelací. V poslední části této kapitoly bude studentům objasněn princip klasického lineárního modelu.

Klíčová slova: jednostranná a vzájemná závislost, regrese a korelace, korelační koeficient, vícenásobná regrese a korelace, multikolinearita, korelace časových řad, opožděná korelace, autokorelace, regresní modely, klasický lineární model.

2.1. Regresní a korelační analýza

Regresní a korelační analýza se zabývá zkoumáním statistických závislostí číselných proměnných. Jsou to závislosti, kdy stejným hodnotám jedněch proměnných mohou odpovídat různé hodnoty jiných proměnných.

Z hlediska počtu proměnných v regresní a korelační analýze rozlišujeme:

- ✓ **Jednoduchou regresní a korelační analýzu**, která zkoumá závislost 2 číselných proměnných:

x – nezávisle proměnná (vysvětlující proměnná)
y – závisle proměnná (vysvětlovaná proměnná).

- ✓ **Vícenásobná regresní a korelační analýza**, která zkoumá závislost 3 a více číselných proměnných:

x, z, u, v… – nezávisle proměnné (vysvětlující proměnné)
y – závisle proměnná (vysvětlovaná proměnná).

Jednoduchá regresní a korelační analýza

Při zkoumání závislosti mezi **proměnnými** je nejdříve nutné posoudit, zda závislost **existuje**, tedy lze-li vysvětlovat změny hodnot jedné proměnné – vysvětlované = **závisle proměnné**, změnami hodnot proměnné druhé – vysvětlující = **nezávisle proměnné**.

Podle charakteru závislosti rozlišujeme u jednoduché regrese a korelace:
a) Jednostrannou závislost (*regresní analýza*), kde závisle proměnnou může být pouze jedna z řešených proměnných.
b) Vzájemnou závislost (*korelační analýza*), kde obě proměnné lze volit za závisle nebo nezávisle proměnnou.

Při zkoumání statistických závislostí řešíme dva zásadní úkoly:
1. **Nalezení vhodné matematické funkce,** tj. **regresní funkce**, pomocí nichž lze odhadovat průměrné hodnoty vysvětlované proměnné, odpovídající zvoleným hodnotám jedné nebo několika vysvětlujících proměnných.
2. **Určení síly** (intenzity, těsnosti) **závislosti** výpočtem **korelačních koeficientů** případně jiných statistických charakteristik.

Bodový diagram – poskytuje první představu o tom, jaká matematická funkce Y-f(x) by mohla být vhodnou regresní funkcí.
Každá uspořádaná dvojice **$[x_i; y_i]$** je znázorněna bodem v pravoúhlé souřadnicové soustavě. Vynesené uspořádané dvojice $[x_i;y_i]$ odpovídajících údajů vytvářejí korelační pole.

Je li matematickou funkcí přímka, hovoříme o lineární regresi. Parametry přímky se počítají již v předchozí kapitole zmiňovanou **metodou nejmenších čtverců**.
Přímka $Y = b_0 + b_1 x$, respektive y = a + bx se používá k odhadům průměrných hodnot proměnné y, odpovídajících zvoleným hodnotám proměnné x.

V případě **vzájemné lineární závislosti obou proměnných**, se spolu s regresní přímkou $\boxed{Y = b_0 + b_1 x}$ používá i regresní přímka $\boxed{X = b_0^* + b_1^* y}$, která slouží k odhadům průměrných hodnot proměnné x, odpovídajících zvoleným hodnotám proměnné y.
Čím je silnější závislost proměnných x, y, tím je úhel, který svírají obě regresní přímky, menší. Při perfektní lineární závislosti (jedna proměnná je funkcí druhé) obě regresní přímky splynou v jednu.
Parametry b_1 a b_1^* sdružených regresních přímek se nazývají **regresní koeficienty**, kde

$b_1 = b_{yx}$ - udává přírůstek průměrné hodnoty y, odpovídající jednotkovému přírůstku proměnné x.
$b_1^* = b_{xy}$ - udává přírůstek průměrné hodnoty x, odpovídající jednotkovému přírůstku proměnné y.

Při interpretaci číselných hodnot obou regresních koeficientů i při využívání obou regresních přímek k odhadům, je nutné přihlížet k tomu, v jakých jednotkách byly proměnné x, y měřeny.

Při zkoumání závislostí se používají některé regresní funkce, u nichž je metoda nejmenších čtverců použitelná až po provedení transformací, jako například u mocninné funkce, Törnquistovy křivka, exponenciály. Nelze-li u některých funkcí, pro vyčíslení parametrů, využít metody nejmenších čtverců, lze ji nahradit jinými metodami například metodou částečných součtů u posunuté exponenciály nebo metodou vybraných bodů například u

Törnquistovy křivka. Metoda vybraných bodů je jednoduchá metoda odhadu parametrů některých nelineárních funkcí. Má-li regresní funkce 2 (3) parametry, určíme ze zdrojových dat statistického šetření nějaké 2 (3) body, kterými by měla funkce procházet.

Index determinace

Determinační index I^2 je charakteristika, která se používá k posouzení vhodnosti regresní funkce, jejíž parametry byly získány metodou nejmenších čtverců, aniž by bylo nutno provést nějakou transformaci vysvětlované proměnné y. Platí, že $I^2 \in \langle 0; 1 \rangle$.
Index determinace používáme pro posouzení vhodnosti přímky, hyperboly, roviny a řady dalších regresních funkcí.
Regresní funkci lze pokládat za tím vhodnější, čím méně se zjištěné hodnoty y_i budou lišit od teoretických hodnot Y_i, tj. čím bude reziduální součet čtverců bližší 0 a teoretický součet čtverců bližší součtu $\sum (y_i - \bar{y})^2$.
Regresní funkce je považována za vhodnou, jestliže lze pomocí této funkce co nejvíce vysvětlit kolísání proměnné y, tj. regresní model je kvalitní, jestliže vysvětluje vysoké % variability hodnot proměnné y.
Hodnoty $\mathbf{I^2}$ blízké **1** svědčí o vhodnosti zvolené regresní funkce a zároveň, že proměnná y silně závisí na proměnné či proměnných, jejichž funkcí je zvolená regresní funkce.
Z hodnot determinačního indexu blízkých 0 **nelze** usuzovat na slabou závislost, ale pouze na nevhodnost zvolené regresní funkce.
Index determinace se počítá z poměru součtu čtverců odchylek:

$$I^2 = \frac{S_T}{S_y} = 1 - \frac{S_R}{S_y}$$.

Kde S_R je reziduální součet čtverců:

$$S_R = \sum (y_i - Y_i)^2$$,

součet čtverců odchylek zjištěných hodnot proměnné od jejich průměru S_Y:

$$S_y = \sum (y_i - \bar{y}_i)^2$$,

a součet čtverců odchylek teoretických hodnot proměnné od jejich skutečných průměrů S_T:

$$S_T = \sum (Y_i - \bar{y}_i)^2$$.

Korelační koeficienty:
Korelační koeficienty jsou charakteristikami síly lineární závislosti číselných proměnných. Sílu lineární závislosti proměnných x; y měří **korelační koeficient $\mathbf{r_{xy}}$ / = $\mathbf{r_{yx}}$ /**, který je poměrem kovariance obou proměnných a součinu jejich směrodatných odchylek.

Korelační koeficient **r** nabývá hodnot z intervalu <-1; 1>. Přičemž|:

- ✓ záporný korelační koeficient ukazuje na **nepřímou** závislost obou proměnných, kdy při růstu hodnoty jedné proměnné průměrné hodnoty druhé proměnné klesají,
- ✓ kladný korelační koeficient ukazuje na **přímou** závislost obou proměnných, kdy při růstu hodnot jedné proměnné rostou i průměrné hodnoty proměnné druhé,
- ✓ nulový korelační koeficient ukazuje, že obě proměnné jsou nezávislé a při růstu hodnot jedné proměnné se průměrné hodnoty druhé proměnné nemění,
- ✓ korelační koeficient je rovný 1 nebo –1 ukazuje na perfektní lineární závislost, kdy stejným hodnotám jedné proměnné odpovídají stejné hodnoty druhé proměnné.

Lineární závislost obou proměnných se považuje za tím silnější, čím je hodnota korelačního koeficientu bližší -1 nebo 1 a za tím slabší, čím je hodnota korelačního koeficientu bližší 0. Znaménko před korelačním koeficientem a regresním koeficientem se musí shodovat.

Závislost lze hodnotit podle 3-5 bodové stupnice:
r = 0 – závislost neexistuje; |r| = 1 – perfektní závislost,
$|r| \in (0; 0,3)$ - slabá závislost,
$|r| \in \langle 0,3; 0,6)$ - střední závislost,
$|r| \in \langle 0,6; 0,8)$ - silná (těsná závislost),
$|r| \in \langle 0,8; 1\ )$ - velmi silná (velmi těsná závislost).

Je-li regresní funkcí přímka, lze dokázat, že $r^2 = I^2$ (druhá mocnina korelačního koeficientu je rovna determinačnímu indexu) a pak podle hodnoty korelačního koeficientu lze posuzovat i vhodnost regresní přímky.

2.2. Vícenásobná regrese a korelace

Vícenásobná regrese a korelace řeší lineární závislosti proměnné **y** na dvou (nebo více) vysvětlujících proměnných **x; z** (u…). Měří se síla lineární závislosti proměnné každé z obou proměnných i na obou proměnných. Matematickou funkcí pro regresi y, x, z
je rovina $\boxed{y = b_0^* + b_1^* x + b_2^* z,}$ kde

$\boxed{b_1^* = b_{yx.z}}$ a $\boxed{b_2^* = b_{yz.x}}$ jsou **dílčí regresní koeficienty**.

$\boxed{b_1^* = b_{yx.z}}$ - přírůstek průměrné hodnoty proměnné y při jednotkovém přírůstku proměnné x za předpokladu, že proměnná z je konstantní.

$\boxed{b_2^* = b_{yz.x}}$ - přírůstek průměrné hodnoty proměnné y při jednotkovém přírůstku proměnné z za předpokladu, že proměnná x je konstantní.

Číselné hodnoty parametrů roviny b_0, b_1, b_2 se získávají metodou nejmenších čtverců, jež vede k třem normálním rovnicím, jejichž řešením jsou vyčíslené parametry.

Korelační koeficient ve vícenásobné regresi:
Pro hodnocení síly závislosti u vícenásobné regrese **nelze** použít **párový korelační koeficient r_{yx}**. Síla lineární závislosti proměnné y na proměnné x se posuzuje na základě **dílčího korelačního koeficientu $r_{yx.z}$**, který měří sílu lineární závislosti proměnné y na proměnné x za předpokladu, že proměnná z je konstantní.
Síla lineární závislosti proměnné y na proměnné z se posuzuje na základě **dílčího korelační koeficientu $r_{yz.x}$,** který měří sílu lineární závislosti proměnné y na proměnné z za předpokladu, že proměnná x je konstantní. Platí že:

- dílčí korelační koeficienty nabývají hodnot z intervalu <-1; 1>,
- záporné hodnoty signalizují nepřímou závislost, kladné hodnoty přímou závislost,
- závislost se považuje za tím silnější, čím jsou hodnoty dílčích korelačních koeficientů bližší -1 nebo 1.

Vícenásobný korelační koeficient $r_{y.xz}$, posuzuje sílu lineární závislosti y na obou vysvětlujících proměnných x; z. Platí že:

- ✓ nabývá hodnot z intervalu <0, 1> ,
- ✓ závislost se považuje za tím silnější, čím je jeho hodnota bližší 1,
- ✓ jeho druhá mocnina je rovna determinačnímu indexu pro rovinu, tedy vztah:

$r^2_{y.xz} = I^2$. Podle hodnoty vícenásobného korelačního koeficientu lze posuzovat vhodnost roviny jako regresní funkce

V případě více vysvětlujících proměnných, lze vypočítat všechny párové korelační koeficienty, mimo jiné i korelační koeficienty mezi všemi dvojicemi vysvětlujících proměnných, přičemž korelace vysvětlujících proměnných se nazývá **multikolinearita**.

Je-li hodnota korelačního koeficientu mezi některou dvojicí vysvětlujících (nezávislých) proměnných blízká 1 nebo -1 hovoříme o **škodlivé multikolinearitě** (alespoň jeden z párových korelačních koeficientů je větší než 0,8). Což je signál, že některá vysvětlující proměnná by se neměla brát v úvahu!

Korelace časových řad

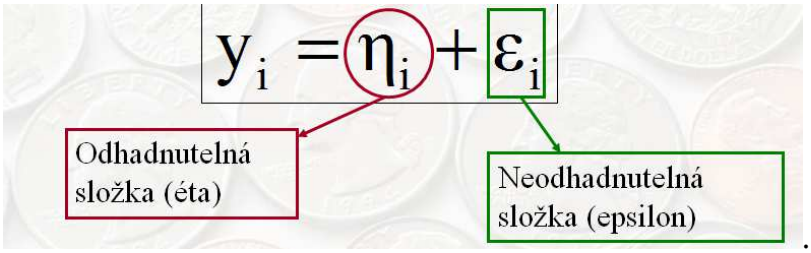
Při studiu korelací časových řad se počítají korelační koeficienty mezi hodnotami dvou časových řad, kde t je čas. Vždy je nutné korelovat odchylky od trendu a při řešení korelací v časových řadách je žádoucí získat vysoké korelační koeficienty a se stejnými znaménky.

Opožděná korelace
Příčinou změn ukazatele (proměnné) **y** jsou změny ukazatel (proměnné) **x**, ale ke změnám ukazatel **y** dochází s určitým **časovým zpožděním**. Hodnoty proměnné **y** lze odhadnout na základě proměnné **x** o rok posunuté. Odhad do budoucna je možný jen o hodnotu posunu.

Autokorelace
Jedná se o korelaci dat jedné časové řady s hodnotami téže časové řady o rok posunutými, pak hovoříme o **autokorelace prvního řádu**; o dva roky posunutými, pak hovoříme o autokorelace druhého řádu…
Vypočítané korelační koeficienty se nazývají **autokorelační koeficienty.** Ve výpočtech figuruje pouze jedna proměnná. Je-li časová řada popsána trendovou funkcí, lze počítat **rezidua** (odchylky od trendu), přičemž často se počítá autokorelace reziduí prvního až n-tého řádu. Silná autokorelace reziduí ukazuje na nekvalitní trendovou funkci. Naopak, trendová funkce se považuje za dobrou, jestliže se nezjistí autokorelace reziduí.

2.3. Regresní modely

Regresní modely se formulují z důvodů zobecňování výsledků v regresní a korelační analýze. Nejjednodušším regresním modelem je **„klasický lineární regresní model“** , kdy výstupy PC software se o tento model opírají.
O pravděpodobnostním rozdělení náhodných veličin y_1, y_2…y_n je nutné určit předpoklady:



Odhadnutelné složky – předpokládá se, že to jsou hodnoty nějaké lineární regresní funkce (přímka, parabola…).

Přímka: $\eta = \beta_0 + \beta_1 x$

Parabola: $\eta = \beta_0 + \beta_1 x + \beta_2 x^2$

Neodhadnutelné složky – v klasickém modelu se zavádějí tři předpoklady:

- ✓ jsou nezávislé,
- ✓ všechny neodhadnutelné složky mají nulové střední hodnoty a zároveň mají stejné rozptyly,
- ✓ jsou náhodné veličiny, které mají normální rozdělení.

Zjištěná data považujeme za výběrová data a na základě výběrových dat se odhadují parametry funkcí (β_0, β_1, β_2…) a hodnoty regresní funkce η_i , kde odhady β se pak značí b_0, b_1, b_2 …, které se počítají se metodou nejmenších čtverců. Lze dokázat, že parametry b_0, b_1 jsou nezkreslené odhady parametrů β_0, β_1.
Odhadujeme-li parametry modelu základního souboru na základě výběru, musíme počítat s možností chyby. O tom zda můžeme očekávat (předpokládat) velkou nebo malou chybu nás informují **směrodatné chyby odhadů.** Ve všech PC výstupech se směrodatné chyby k odhadům připojují.

Shrnutí kapitoly
V kapitole byly vysvětleny základní principy regrese a korelace a to i na úrovni vícenásobné regresní a korelační analýzy. Rovněž byly osvětleny nadstavbové metody a to korelace časových řad, opožděná korelace a autokorelace. V poslední pasáži byly vysvětleny podstatné body týkající se klasického lineárního regresního modelu včetně směrodatné chyby odhadu.

Pojmy k zapamatování:
Jednoduchá regrese a korelace. Jednostranná a vzájemná závislost. Vícenásobná regrese a korelace. Regresní koeficient. Korelační koeficient. Párový, dílčí a vícenásobný korelační koeficient. Index determinace. Multikolinearita. Korelace časových řad, opožděná korelace, autokorelace. Klasický lineární regresní model.

Úkoly k zopakování a procvičení

Příklad 2.1.:
Výstupem regrese vzájemné závislosti:
a) je parabola
b) je rovina
c) jsou sdružené regresní přímky
Řešení: c

K posouzení vhodnosti lineární regresní funkce se využívá:
a) hodnota indexu determinace
b) hodnota regresního koeficientu
c) hodnota korelačního koeficientu
Řešení: a

Korelační koeficient nabývá u jednoduché korelace hodnot z intervalu:
a) <0, 1>
b) <-1, 1>
c) (0 ,1)
Řešení: b

Příklad 2.2.:

Sílu lineární závislosti proměnné y na proměnné x za předpokladu, že proměnná z je konstantní měří:
a) párový korelační koeficient
b) dílčí korelační koeficient
c) vícenásobný korelační koeficient
Řešení: b

Běžnou matematickou funkcí používanou ve vícenásobné regresi x, y, z je:
a) přímka
b) exponenciála
c) rovina
Řešení: c

Vícenásobný korelační koeficient posuzuje sílu lineární závislosti y v modelu proměnných y, x, z:
a) na obou vysvětlujících proměnných x, z
b) na vysvětlující proměnné x
c) na vysvětlující proměnné z
Řešení: a

Příklad 2.3.:
Typem regresního modelu, o který se nejčastěji opírají PC programy je:
a) exponenciální regresní model
b) Törnquistova křivka
c) klasický lineární model
Řešení: c

Předpokládá se, že odhadnutelné složky v klasickém lineárním modelu jsou hodnoty:
a) nějaké lineární regresní funkce
b) nějaké exponenciální regresní funkce
c) polynomu vyššího stupně
Řešení: a

Neodhadnutelné složky klasického lineárního modelu značíme symbolem:
a) π
b) χ
c) ε
Řešení: c

Hodnocení
Každá správná odpověď nebo výsledek výpočtu je hodnoceno jedním bodem.
Sebehodnocením je žádoucí dosáhnout alespoň 70% úspěšnost správných odpovědí. Jestliže

jste nedosáhli požadované úspěšnosti, pokuste se zlepšit svůj studijní výsledek pozornějším studiem kapitoly, popřípadě se spojit s tutorem předmětu.

Korespondenční úkol:

Regresní analýzou byla ze zdrojových dat (proměnné x_i; y_i) získána rovnice funkce - přímka: $y = 4 + 0{,}1x$. Zdrojová data jsou uvedena v následující tabulce:

x_i	y_i
10	4
10	6
20	5
20	7
30	6
30	8

Vypočítejte hodnotu indexu determinace a posuďte kvalitu regresního modelu.
Řešení: $I^2 = 0{,}4$. Přímka není příliš vhodnou regresní funkcí.

3. Modul

Modul tvoří tři tematické okruhy. Každý je probírán samostatně, jako kapitola v učebním materiálu.

Tématické okruhy:

3.1. Lineární trendy v časových řadách
3.2. **Metody vícerozměrné statistické analýzy**
3.3. Práce českého statistického úřadu, historie statistiky

Studijní cíle

V posledním modulu jsou studenti seznámeni se všemi běžně využívanými lineárními trendy v časových řadách, kdy je kladen důraz zejména na přímku, parabolu a hyperbolu. Rovněž jsou uvedeny i příklady nelineárních trendů, které problematiku v kontrastu doplňují. Zmíněna je rovněž i problematika sezónních časových řad. Ve druhé kapitole tohoto modulu jsou metody vícerozměrné statistické analýzy klasifikovány a vysvětleny jejich základní principy. Na tohoto modulu se studenti seznámí se zásadními mezníky historie statistiky celosvětovém měřítku i na území dnešní České republiky. Tato poslední kapitola je doplněna i o výtah nejdůležitějších aktivit Českého statistického úřadu.

Klíčová slova:
Časová řada, lineární funkce, přímka, parabola, hyperbola, nelineární funkce, sezónní časové řady, metody vícerozměrné statistické klasifikace, metody analýzy korelačních struktur, historie statistiky, Český statistický úřad

3.1. Lineární trendy v časových řadách

U lineárních trendových funkcí (přímka, hyperbola, parabola) se číselné hodnoty parametrů určují **metodou nejmenších čtverců**.
Parametry se určují tak, že je minimalizován součet druhých mocnin odchylek zjištěných hodnot y_t od zvolené trendové funkce: $\boxed{S = \sum_t (y_t - T_t)^2}$.

Matematický postup metody nejmenších čtverců
Provedou se parciální derivace v součtu **S** podle jednotlivých parametrů a položí se rovny nule. Takto získáme tzv. normální rovnice, jejichž řešením se získají hodnoty parametrů.
U dvouparametrické trendové funkce jde o dvě rovnice o dvou neznámých u tříparametrické o tři rovnice o třech neznámých …

Přímkový trend - hodnoty časové řady rostou (klesají) lineárně s časem, přičemž první diference jsou přibližně konstantní, druhé diference kolísají kolem nuly.

Hyperbolický trend - tato funkce se používá při zpomalujícím se rostoucím trendu (roční přírůstky se postupně zmenšují) nebo naopak při zpomalujícím se klesajícím trendu.
Zda je pro popis trendu vhodná hyperbola určujeme z chování ročních přírůstků. Zpomalující se rostoucí trend popisuje hyperbola s parametry $b_0>0$ a $b_1<0$, zatímco zpomalující se klesající trend popisuje hyperbola s parametry $b_0>0$ a $b_1>0$.

Parabolický trend - první diference v čase jsou lineární, druhé diference přibližně konstantní, třetí diference jsou nulové. Roční přírůstky rostou, či naopak klesají, ale přírůstky ročních přírůstků (druhé diference) kolísají aniž by se systematicky zvětšovaly nebo zmenšovaly.

Obecné rovnice výše uvedených funkcí již byly zmíněny v kapitole 1.3.

Kvalita trendových funkcí se vždy ověřuje výpočtem reziduí. Regresní funkci lze pokládat za tím vhodnější, čím méně se zjištěné hodnoty y_i budou lišit od teoretických hodnot Y_i, tj. čím bude reziduální součet čtverců bližší 0.

Nelineární funkce
V praxi se za trendové funkce nevolí pouze lineární funkce, ale i různé funkce **nelineární,** kdy parametry některých z nich lze získat metodou nejmenších čtverců až po provedení **linearizující transformace**, kde **s**ledovaná proměnná y je nahrazena nějakou neparametrickou funkci y^* (například exponenciála).

Sezónní časové řady
Jedná se o časové řady, v nichž je kromě trendu patrné i **sezónní kolísání** (například časové řady čtvrtletních hodnot).

Při analýze sezónních časových řad jde o:
- vystižení jejich trendu (vhodná trendová funkce),
- vystižení jejich sezónního kolísání (využívají se sezónní odchylky nebo sezónní indexy).

K popisu konstantního **sezónního kolísání** se používají sezónní odchylky.
K popisu proporcionálního sezónního kolísání se používají **sezónní indexy**.

Existuje několik metod analýz sezónních časových řad…

Jednoduché metody:
- metoda empirických **sezónních odchylek**, užívaná pro časové řady s konstantní sezónností,
- metoda empirických **sezónních indexů**, užívaná pro časové řady s proporcionální sezónností.

Obě metody vycházejí z vyrovnání časové řady **centrovanými klouzavými průměry.**
S výhodou využíváme výpočet prognóz ze sezónně očištěných časových řad, tj. parametry trendové funkce počítáme z **očištěných** hodnot.

3.2. Metody vícerozměrné statistické analýzy

Metody umožňují hodnotit větší počet proměnných jako jeden celek, tj. umožňují komplexněji hodnotit statistické jednotky podle většího počtu proměnných, uvažovaných současně.
Výpočty u těchto metod jsou velmi náročné a často s více variantami algoritmu řešení, proto hojně využíváme výpočetní techniku a specializované počítačové programy.
Uživatel těchto metod musí znát podmínky použitelnosti těchto metod a způsob správného vysvětlení získaných výsledků.

Charakteristika metod vícerozměrné statistické analýz:
1. Metody vícerozměrné statistické klasifikace
Jedná se metody, které jsou orientovány na rozlišování a rozdělování **mnohorozměrných** statistických jednotek do dvou či více **stejnorodějších souborů**, do kterých jsou řazeny jednotky co nejvíce **navzájem podobné**, zatímco jednotky z různých souborů jsou co nejvíce navzájem odlišné: **diskriminační analýza; shluková analýza**

2. Metody analýzy korelačních struktur
Tyto metody se zabývají **racionální redukcí** dimenze (rozsahu) řešeného problému, tj. koncentrací informací obsažených ve **větším** počtu proměnných do podstatně **menšího** počtu skupin proměnných: **faktorová analýza; analýza hlavních komponent; kanonická korelační analýza**

Diskriminační analýza
Ve dvou či více předem určených souborech statistických jednotek stejného typu je sledován větší počet proměnných (měřených statistických znaků).
- ✓ Výsledkem diskriminační analýzy je **diskriminační funkce**, umožňující zařazovat s minimální chybou rozhodnutí statistické jednotky do správného souboru.
- ✓ Cílem je sestavit na základě výběrů z několika souborů (min. dvou) **diskriminační kritérium** umožňující zatřiďovat studované jednotky do těchto souborů
- ✓ Metoda umožňuje **stanovit „důležitost"** jednotlivých proměnných pro rozlišitelnost souborů od sebe, tj. určuje podíly proměnné na celkové spolehlivosti rozlišení souborů.

Shluková analýza (Cluster Analysis)
Účelem shlukové analýzy je **rozdělení** souboru do určitého počtu skupin, kde jednotlivé skupiny jsou relativně stejnorodé, tj. **jednotky uvnitř skupiny,** tzv. clusteru, se příliš **neliší,** ale **skupiny navzájem se liší hodně.**
Využití shluková analýzy:
- marketingové výzkumy,

- segmentace trhu,
- přírodní vědy (molekulární genetika → genetické vzdálenosti druhů, jedinců…).

Typickým grafickým výstupem shlukové analýzy dendrogram.

Faktorová analýza

Formálně představuje **zvláštní typ regrese**, která zkoumá závislost proměnných (v realitě pozorovatelných, měřitelných) na námi nepozorovatelných (neměřitelných, skrytých) proměnných. Tyto proměnné se označují jako **společné faktory**.

V každé vytvořené homogennější skupině proměnných lze využít výsledek metody faktorové analýzy:

- k pořadí důležitosti jednotlivých proměnných,
- k identifikaci a vyhodnocení vzájemných mnohostranných závislostí mezi proměnnými ve skupině,
- k návrhu na eventuální redukci počtu proměnných (vyřazení nejméně významných proměnných soustavy),
- k návrhu na získání agregovaných (sdružených, sloučených) informací (za celou skupinu proměnných, sdružených jedním společným faktorem).

Analýza hlavních komponent

V jediném souboru statistických jednotek sledujeme větší počet proměnných, z nichž každá obsahuje určitou část několika rozlišitelných kategorií (komponent) studované souhrnné informace.
Touto metodou vymezujeme tyto kategorie, tzv. **hlavní komponenty**.

Algoritmem metody analýzy hlavních komponent jsou **koeficienty**, které udávají, jak se na každé hlavní komponentě podílejí jednotlivé proměnné. Algoritmus metody zajišťuje sestupnou prioritu hlavních komponent.
Po formální stránce je výsledek podobný faktorové analýze.

Kanonická korelační analýza

Soustavu proměnných, na rozdíl od vícenásobné regrese a korelace (kde je jediná závisle proměnná **y** a několik nezávislých proměnných **x, z, u, v**…) **rozdělujeme na dvě podsoustavy** o větším počtu proměnných a vzájemnou závislost těchto podsoustav měříme co **nejmenším** počtem **koeficientů** (první, druhý, popřípadě třetí koeficient kanonické korelace).
Modul kanonická korelace hledá obecný lineární vztah mezi dvěma vícerozměrnými proměnnými X a Y s obecně různými dimenzemi m1, m2.

Preferenční analýza

Preferenční analýza vychází z další vícerozměrné statistické metody a to z **analýzy rozptylu** (ANOVA – **An**alysis **o**f **Va**riance). Zdrojová data jsou však **diskrétní proměnné** (stupnice).

Využití:

Jak navrhnout výrobek, aby byl přitažlivý na trhu, kdy zákazníci vyjadřují své preference pomocí stupnice.
Analýza marketingových průzkumů k detekci proměnných, které nejvíce ovlivňují volbu produktu (vůně, chuť, barva, vzhled…).

Analýza rozptylu – ANOVA (**An**alysis **o**f **Va**riance)

Princip algoritmu výpočtu analýzy rozptylu je rozklad celkového zdroje variability všech údajů na více složek, z nichž jedna je vyjádřená tzv. **reziduálním** (vnitřním) **rozptylem**, který měří přirozené kolísání hodnot sledovaného znaku v každé třídě (porovnávané výběrové soubory) okolo průměru třídy (kategorie). Další složky rozkladu měří kolísání hodnot způsobené tzv. **efektem třídění**.

Nulová hypotéza H_1, tvrdí, že rozptyl vyjadřující efekt třídění je průkazně větší než rozptyl reziduální neboli **existuje alespoň jedna dvojice průměrů** vykazující statisticky významný rozdíl na zvolené hladině významnosti α.

Průkaznost rozdílů mezi rozptyly hodnotíme pomocí **F- testu**, což je **základní část výpočtu analýzy rozptylu.**

Je-li přijata H_1 (alternativní hypotéza) detailní dvojice průměrů, které se mezi sebou liší, pak řeší metody podrobnějšího vyhodnocení analýzy rozptylu:

- T- metoda = Tukeyho metoda,
- S - metoda = Scheffého metoda aj.

Autorem metody ANOVA je Sir Ronald Aylmer Fisher (1890 -1962), anglický statistik, evoluční biolog a genetik.

T-testy

Testy se využívají pro porovnání střední hodnoty jednoho nebo dvou normálně rozdělených základních souborů.

Kritické hodnoty pro t-testy vyhledáváme ve statistických tabulkách Studentovo t –rozdělení, jejichž autorem je William Sealy Gosset přezdívaný „Student“.

3.3. Práce českého statistického úřadu, historie statistiky

Uvedená kapitola modulu vychází z internetových stránek Českého statistického úřadu a v plném rozsahu se odkazuje na samostatnou práci studentů s informacemi (včetně historie statistiky) uvedenými na internetové adrese:

http://www.czso.cz.

Studenti mají možnost navštívit studovnu Českého statistického úřadu a využít mnohé bezplatné služby této instituce se sídlem v Praze a to na adrese:

Na padesátém 81
100 82 Praha 10
Tel.: 274 051 111 (ústředna).

Shrnutí kapitoly

V kapitole byly popsány nejvýznamnější lineární trendy časových řad využívaných v ekonomické praxi v kontrastu s nelinárními trendy. Dále byly popsány základní principy sezónnosti v časových řadách. Studenti byli seznámeni se základní klasifikací nejvýznamnějších metod vícerozměrné statistické analýzy. Byly shrnuty jejich základní principy s důrazem na objasnění zásadních výpočtů u vybraných analýz. Poslední část modulu odkazuje studenty na internetové stránky Českého statistického úřadu, kde si mohou individuálně vyhledat stěžení informace o činnosti této státní instituce a rovněž o historii statistiky na celosvětové i tuzemské úrovni včetně osobností spojenými s problematikou a historií statistiky.

Pojmy k zapamatování:

Časová řada, lineární trend, přímka, parabola, hyperbola, nelineární trend, exponenciála, metody vícerozměrné statistické analýzy, metody vícerozměrné statistické klasifikace, diskriminační analýza; shluková analýza, metody analýzy korelačních struktur
faktorová analýza; analýza hlavních komponent; kanonická korelační analýza, analýza rozptylu, preferenční analýza, T-test, Český statistický úřad.

Úkoly k zopakování a procvičení

Příklad 3.1.:
Mezi lineární funkce nepatří:
a) parabola
b) hyperbola
c) exponenciála
Řešení: c

Při zpomalujícím se rostoucím trendu nebo naopak při zpomalujícím se klesajícím trendu je typickou funkcí popisující uvedený trend:
a) přímka
b) hyperbola
c) parabola
Řešení: b

K popisu konstantního sezónního kolísání se používají:
a) sezónní odchylky
b) sezónní indexy
c) regresní koeficienty
Řešení: a

Příklad 3.2.:
Mezi metody vícerozměrné statistické klasifikace nepatří:
a) preferenční analýza
b) diskriminační analýza
c) shluková analýza
Řešení: a

Nejběžnějším grafickým výstupem shlukové analýzy je:
a) polygon
b) hisogram
c) dendrogram
Řešení: c

Základní část výpočtu analýzy rozptylu tvoří:
a) F-test
b) T- test
c) χ^2 - test
Řešení: a

Příklad 3.3.:
Český statistický úřad - centrála:
a) sídlí v Brně

b) sídlí v Praze
c) nemá sídlo, pracuje pouze on-line
Řešení: b

Mezi významné statistiky patří:
a) Thomas Korrel
b) Karl Pearson
c) Peater Cluster
Řešení: b

Česká republika je z hlediska historie statistiky, využívaní a rozvoje statistických metod:
a) velmi zaostalá, stejně jako středoafrické státy
b) na nejvyšší úrovni v celosvětovém srovnání
c) zhruba na úrovni USA
Řešení: b

Korespondenční úkol:

Vyhledejte na internetových stránkách přesné iniciály a data narození, případně úmrtí zakladatele metody ANOVA.
Řešení: Sir Ronald Aylmer Fisher (1890 - 1962).

Hodnocení
Každá správná odpověď nebo výsledek výpočtu je hodnoceno jedním bodem.
Sebehodnocením je žádoucí dosáhnout alespoň 70% úspěšnost správných odpovědí, výsledků výpočtů. Jestliže jste nedosáhli požadované úspěšnosti, pokuste se zlepšit svůj studijní výsledek pozornějším studiem kapitoly, popřípadě se spojit s tutorem předmětu.

Další studijní zdroje
http://new.euromise.org/czech/tajne/ucebnice/html/html/node9.html
http://www.economics.soton.ac.uk/staff/aldrich/Figures.htm
http://www.czso.cz/