MVŠO

STATISTIKA A STATISTICKÉ ZPRACOVÁNÍ DAT

STUDIJNÍ OPORA PRO KOMBINOVANÉ STUDIUM

STATISTIKA A STATISTICKÉ ZPRACOVÁNÍ DAT

Mgr. Bc. Veronika Říhová, Ph.D.

Moravská vysoká škola Olomouc, o.p.s., 2017

Autor: Mgr. Bc. Veronika ŘÍHOVÁ, Ph.D.

Olomouc 2017

Obsah

Úvod 7

Význam a pojetí moderní statistiky 8

1.1 Pojetí moderní statistiky 9

1.2 Předmět a obsah statistiky 10

1.3 Základní statistické pojmy 11

Zpracování dat z výběrových zjišťování 13

2.1 Výběrová šetření 14

2.2 Prostý náhodný výběr 15

2.3 Výběrové charakteristiky a jejich rozdělení 16

Bodový a intervalový odhad 20

3.1 Statistika 21

3.2 Bodový odhad 21

3.3 Kritické hodnoty rozdělení 25

3.4 Intervalové odhady parametrů 26

3.5 Intervalový odhad střední hodnoty 27

3.6 Intervalový odhad rozptylu 29

Testování statistických hypotéz 32

4.1 Statistické hypotézy – základní pojmy 33

4.2 Kroky při testování hypotézy 35

Parametrické testy **39**

5.1 Úvod 40

5.2 Hypotézy o rozptylu 40

5.2.1 Test významnosti rozdílu dvou rozptylů (*F*-test) 40

5.3 Hypotézy o střední hodnotě 42

5.3.1 Test významnosti rozdílu $|m - \mu_0|$ 42

5.4 Test významnosti rozdílu dvou výběrových průměrů (*t*-test) 44

5.5 Studentův test pro párované hodnoty 46

Neparametrické testy **50**

6.1 Úvod 51

6.2 Kolmogorovův-Smirnovův test dobré shody pro jeden výběr 51

6.3 Kolmogorovův-Smirnovův test dobré shody pro dva výběry 54

6.4 Testy extrémních hodnot 56

6.4.1 Dixonův test extrémních odchylek 56

Regresní a korelační analýza **59**

7.1 Regresní funkce 60

7.2 Lineární regrese 61

Intervaly spolehlivosti a testy hypotéz v regresi a korelaci **71**

8.1 Interval spolehlivosti korelačního koeficientu (koeficientu determinace) 72

8.2 Interval spolehlivosti regresních parametrů a modelových hodnot (modelu) 73

8.3 Interval spolehlivosti Y hodnot a predikovaných hodnot (pás spolehlivosti) 74

8.4 Testy významnosti v regresní analýze 75

8.4.1 Test významnosti korelačního koeficientu R 76

8.4.2 Test významnosti regresního modelu 77

8.4.3 Test významnosti regresních parametrů 77

8.4.4 Hodnocení modelu z hlediska výsledků testů významnosti 78

8.4.5 Test shody regresních modelů 78

Statistické srovnávání ekonomických jevů **82**

9.1 Úvod 83

9.2 Ukazatel jako statistická veličina 83

9.2.1 Typy a vlastnosti ukazatelů 85

Indexy a absolutní rozdíly jako nástroj srovnávání a analýzy 89

10.1 Absolutní porovnávání 90

10.2 Relativní porovnávání 91

10.3 Indexy 92

10.4 Jednoduché individuální indexy 92

10.5 Složené individuální indexy 93

10.6 Souhrnné indexy 95

10.6.1 Cenové indexy 96

10.6.2 Objemové indexy 97

Publikace výsledků statistického zpracování dat 99

11.1 Etapy statistického zpracování dat 100

11.2 Chyby měření 101

11.3 Třídění statistických dat 103

11.3.1 Zpracování údajů statistickými postupy 103

11.3.2 Statistické programy 104

11.4 Prezentace a interpretace dat 105

Úvod

Studijní text seznamuje studenty se základními statistickými pojmy a nejdůležitějšími metodami s důrazem na porozumění smyslu statistické činnosti. Cílem výuky je vytvořit, resp. upevnit u posluchačů základy statistické gramotnosti a schopnost orientovat se ve statistických datech a ukazatelích; seznámit je se základními statistickými metodami, a to jak se zřetelem k využití v ekonomii, tak i v běžném životě.

Po absolvování kurzu student: zvládne základní statistické zpracování datového souboru ve formě tabulek, grafů a číselných charakteristik; bude schopen porozumět statistickým datům, aplikovat základní statistické postupy a správně publikovat výsledky. Současně se naučí pracovat se základními statistickými funkcemi a nástroji analýzy dat a interpretovat příslušné statistické výstupy.

Kapitola 1

Význam a pojetí moderní statistiky

Po prostudování kapitoly budete umět:

- definovat předmět a obsah statistiky;
- určit základní statistické pojmy a popsat jejich dělení.

Klíčová slova:

Statistika, deskriptivní statistika, statistická indukce, statistická jednotka, statistický znak.

1.1 Pojetí moderní statistiky

V každodenním životě je statistika na denním pořádku, aniž bychom si to uvědomovali. Jasnou příčinou je doba informací, kde informace tvoří komoditu stejně běžnou, jako tomu bylo a je například u automobilů a oblečení.

Počátek statistiky sahá až do doby starověkých říší. V té době šlo spíše o soupisy obyvatelstva, které umožňovaly přehlednější výběr daní. Samotné označení „statistika“ jako vědní obor vzniklo až v 18. století.

Největší rozmach statistika zaznamenala v 70. letech 20. století, kdy velkým příspěvkem byla výpočetní technika, která dokázala simulovat statistické prostředí. Díky tomu se v dnešní době žádný z vědeckých oborů neobejde bez práce s hromadnými daty a jejich vyhodnocování. Z těch nejčastějších oborů, které pracují se statistikou, je medicína, biologie, fyzika a jiné přírodní i technické disciplíny. Statistika ale tvoří velmi významný pilíř v marketingu, což je pro nás ekonomy velmi důležitý aspekt.

Statistika slouží nejen k úspěšné realizaci změn ve státní ekonomice, ale i ve tvorbě manažerských rozhodování, analýz příjmů a výdajů, rozborů současných trendů a jiných podnikových aktivit k flexibilnějšímu postavení na trhu. Je potřeba si uvědomit, že k úspěšným závěrům není možné jen znát pojem statistika, ale i disponovat zkušenými statistickými ekonomy.

Samotný termín statistika je dosti obecný a specifikovat ho přesně je vcelku obtížné. I když pojmu statistika rozumí každý, máme o ní rozdílné mínění. Někomu přijde až naprosto nedůvěryhodná. Důvodem podceňování statistiky u některých jedinců jsou možná zneužívání ve smyslu záměrných zkreslování statistik a poté jejich chybné vyhodnocování.

Je nesprávné jak přeceňování statistiky, protože sama o sobě všechno poznat nemůže, tak i podceňování jejich možností, protože správná statistika je velmi důležitým a ničím nezastupitelným nástrojem k získávání dat.

1.2 Předmět a obsah statistiky

„Předmětem statistiky je zjišťování a analýza vývoje stavu. V sociální oblasti to například znamená jev dotýkající se kvality života obyvatelstva a jejího vývoje[1].“

U ekonomické statistiky se statistika zabývá veškerými informacemi, které jsou spojeny s hospodářskou oblastí. Jsou to takové údaje, které se po jejich vyhodnocení dají dále využít k rozhodování či stanovení hospodářské politiky.[1]

Základní rozdělení:

- STRUKTURÁLNÍ (přesné výsledky ve větším časovém měřítku - 1 rok a více)
- KONJUNKTURÁLNÍ (co nerychlejší výsledky bez ohledu na jejich úplnou správnost)
- PODNIKOVÁ STATISTIKA (zjišťuje, vede a analyzuje ukazatele charakterizující ekonomickou vitalitu podniku)
- STÁTNÍ STATISTIKA (shromažďuje statistiky o sociálním, ekonomickém a ekologickém vývoji)

Statistické orgány na mezinárodní úrovni, v niž ČR hraje roli, jsou EUROSTAT a OSN.[2]

Abychom se vyvarovali nesprávných úsudků vyplývajících z neznalosti, je vhodné se seznámit se základy matematické statistiky a s jejími možnostmi.

Nejčastější aplikace počtu pravděpodobnosti směřují do oblasti statistiky. Její nejrozšířenější část, tzv. **matematická statistika,** se zabývá metodami získávání, zpracování a vyhodnocování hromadných dat (tzn. údajů o vlastnostech velkého počtu jedinců - osob, věcí či jevů).

Podle použitých metod práce dělíme matematickou statistiku na

- **deskriptivní, popisnou statistiku** - zabývá se efektivním získáváním ukazatelů, které poskytují obraz zkoumaného jevu;
- **statistickou indukci** (matematickou statistiku v užším smyslu) - řeší problémy zobecňování výsledků získaných popisem statistického souboru.

[1] MACEK, J. *Ekonomická a sociální statistika*. 2008, s. 6.

1.3 Základní statistické pojmy

Ve statistice se zajímáme o jevy a procesy mnoha prvků. Tyto jednotlivé prvky se nazývají **statistické jednotky**. Statistice se meze nekladou, a proto takovými jednotkami může být cokoliv. Pokud se budeme bavit o podnikové statistice, tak naší statistickou jednotkou budou například zaměstnanci a budeme zkoumat jejich výši mezd, výkonnost produkce, atd. Výsledkem tohoto zkoumání budou **statistické znaky**. Tyto znaky se nám dále dělí na číselné (**kvantitativní**) a slovní (**kvalitativní**) a v některých případech se slovní charakteristika nazývá i **kategoriální** (**akciová společnost, družstvo…).** Například mzdy můžeme označit číselnými znaky (48.000,-, 60.500,-…), u vzdělání zaměstnanců využijeme slovní charakteristiku (základní vzdělání, výuční list…).

Pokud se zkoumaný subjekt (statistická jednotka) dá označit pouze dvěma variantami, tak ho nazýváme znakem **alternativním** (rozdělení počtu zaměstnanců podniku podle pohlaví).[2]

Znak, který připouští více, než dvě varianty se nazývá znak **množný** (nejvyšší dosažené vzdělání pracovníků v podniku - ZŠ, SŠ, VŠ,) - s tímto členěním se setkáváme obvykle jen u znaků kvalitativních.

Kvalitativní znaky se nám dále rozdělují na znaky **nominální -** znaky dokážeme pouze vyjmenovat (např. rodinný stav muže: svobodný, ženatý, rozvedený, vdovec) a **pořadové -** znaky, které umožňují uspořádat jejich hodnoty podle velikosti (např. dosažené vzdělání, zkušenosti, délka praxe).

Kvantitativní znaky pak dále rozdělujeme na **spojité** (nabývají libovolných hodnot jako třeba spotřeba energií, doba vyrobení zakázky atd.) a **nespojité** (nabývají pouze některých číselných hodnot, jako například počet zmetků ve vyrobené sérii, počet pracovníků na daném úseku atd.)

Cílem úvodní kapitoly bylo zavedení a objasnění pojmu statistika, seznámení se základní statistickou terminologií a definování charakteristik statistického souboru.

1. K jakému účelu souží statistika na státní úrovni?
2. Co je obsahem ekonomické statistiky a jaké je její dělení?
3. Co je to matematická statistika a čím se zabývá?
4. Definujte základní statistické pojmy.

[2] MACEK, J. *Ekonomická a sociální statistika*. 2008, s. 6.

Literatura k tématu:

[1] HINDLS, R. *Statistika pro ekonomy*. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

[2] MACEK, J. *Ekonomická a sociální statistika*. 1. vyd. Plzeň: Západočeská univerzita v Plzni, 2008. ISBN 978-80-7043-642-4.

[3] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 1. vydání. Ostrava: Vysoká škola Báňská - Technická univerzita Ostrava, 2007 [cit. 2017-12-18]. ISBN 80-248-1194-4. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/

[4] ANDĚL, J. *Statistické metody*. 3. vyd. Praha: Matfyzpress, 2003. ISBN 978-80-867-3208-8.

[5] MAREK, L. *Statistika v příkladech*. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

Kapitola 2

Zpracování dat z výběrových zjišťování

Po prostudování kapitoly budete umět:

- definovat pojem náhodný výběr a výběrový soubor a jeho tvorbu;
- určit základní výběrové charakteristiky a jejich rozdělení.

Klíčová slova:

Výběrový soubor, odhad, výběrové charakteristiky.

2.1 Výběrová šetření

Nejdůležitějším druhem neúplného šetření je pravděpodobnostní neboli náhodný výběr. Provádí se tak, že se celý soubor nejprve rozdělí na výběrové jednotky – které jsou zpravidla totožné s elementárními (statistickými) jednotkami, ale mohou to být také jejich větší nebo menší skupiny – načež se každé výběrové jednotce přiřadí určitá (nejčastěji stejná) pravděpodobnost jejího zahrnutí do výběrového souboru. Vlastní výběr (selekce) jednotek se pak provede tak, aby o vybrání či nevybrání každé jednotky rozhodovala již jen náhoda. (Prakticky se dá takový výběr uskutečnit třeba tak, že se jednotky nebo lístky s názvy jednotek vylosují z osudí. K této technické stránce problému se ještě vrátíme.)

Obě zmíněné stránky tvorby výběrového souboru – pravděpodobnost a náhodu – je třeba vidět v jejich spojitosti. Uplatnění předem stanovených pravděpodobností vyžaduje naprostou náhodnost při vlastním vybírání, a naopak náhodnost předpokládá existenci určité soustavy pravděpodobností. Pro výraznější rozlišení obou stránek budeme mluvit o pravděpodobnosti vybrání (vytažení) – později o příbuzném pojmu pravděpodobnosti zahrnutí – jakožto vlastnostech jednotky a o náhodnosti vybírání jako metodě výběru.

Pravděpodobnostní hledisko náhodného výběru je natolik významné, že v současnosti již název "pravděpodobnostní výběr" převažuje nad starším a v praxi dosud občas užívaným názvem "náhodný výběr". Určitý význam má v tomto i ta okolnost, že pravděpodobnosti vybrání nemusí být v daném souboru u všech jednotek stejné, ale mohou se lišit. V souvislosti s tím je třeba upozornit, že v některé poválečné české (ale i cizojazyčné) literatuře byl termín "pravděpodobnostní výběr" vyhrazen pouze pro výběry s nestejnými pravděpodobnostmi.

U neodborníků se může statistik často setkat s pochybnostmi, jak je možné, že náhodný výběr může být dobrým podkladem pro zabezpečení reprezentativnosti, a tím pro usuzování z části na celek. Zdá se jim, že pokud ponecháme náhodě, které prvky budou vybrány, přestáváme řídit a ovlivňovat tvorbu výběrového souboru, stáváme se "obětí živelnosti" atd. To by ovšem platilo, kdyby se výběr prvků prováděl s různými a nám neznámými pravděpodobnostmi. Avšak tím, že všem prvkům přiřadíme předem známé pravděpodobnosti – a to buď pravděpodobnosti vybrání nebo pravděpodobnosti zahrnutí – můžeme využít výhodných stránek náhody, matematicky ovládat její zákonitosti.

V porovnání s úsudkovými (záměrnými) výběry to tedy znamená, že u pravděpodobnostních výběrů jsme schopni sestavit takové odhady, které s rostoucím rozsahem výběru konvergují k odhadované (skutečné) hodnotě – tzv. odhady konzistentní – a které často navíc při každém rozsahu výběru skutečnou hodnotu v průměru ani nenadhodnocují, ani nepodhodnocují – tzv. odhady nevychýlené.

Jejich přesnost lze při daném rozsahu výběru objektivně změřit, tj. určit střední velikost jejich výběrové chyby, popř. stanovit interval, v němž se téměř jistě nachází skutečná hodnota – tzv. intervalové odhady. Tato problematika bude podrobněji řešena v dalších kapitolách.

2.2 Prostý náhodný výběr

- jedná se o pravděpodobnostní výběr, kdy každý prvek ZS (populace) má stejnou pravděpodobnost, že se do výběru dostane.

Prostý náhodný výběr lze také definovat jako výběr o rozsahu *n*, kdy každá množina *n* prvků má stejnou pravděpodobnost, že bude vybrána.

K realizaci takového výběru musíme mít k dispozici očíslovaný seznam všech prvků základního souboru - tzv. **oporu výběru**, a dále generátor náhodných čísel, pomocí něhož vybereme očíslovaný prvek z opory výběru. Předpokládejme, že ZS má *N* prvků a výběr bude mít *n* prvků. Procedura výběru sestává z následujících kroků:

1. sestavíme oporu výběru a každému prvku přiřadíme celé číslo od 1 do *N*
2. rozhodneme, jak velký bude rozsah výběru *n*
3. vygenerujeme *n* náhodných celých čísel mezi 1 a *N*
4. získáme data od prvků identifikovaných v opoře výběru těmito náhodnými čísly

Poměr mezi rozsahem výběru *n* a velikostí ZS (populace) *N* nazýváme výběrový poměr:

$$výběrový\ poměr = \frac{rozsah\ výběru\ n}{velikost\ populace\ N}$$

Tento poměr vyjadřuje pravděpodobnost, že prvek ZS je zařazen do výběru. Výběr můžeme provádět *s vracením* nebo *bez vracení*. Vrátíme-li prvek do základního souboru, má nenulovou pravděpodobnost, že bude do výběru vybrán vícekrát. Výhodnější pro statistické odvozování různých formulí je výběr s vracením. V takovém případě je však vhodné, aby výběrový poměr byl malý (<5%).

Někdy se stává, že prostý náhodný výběr je neproveditelný nebo nákladný, hlavně v případech, kdy je ZS značně rozsáhlý. Uvádíme některé přijatelné náhradní metody výběru, jež ve výběru používají náhodný mechanismus:

- **stratifikovaný náhodný výběr** - je-li možné ZS rozdělit do dílčích oblastí, můžeme provést náhodný výběr pro každou oblast. Tyto oblasti se pak nazývají strata nebo vrstvy.

Tato technika je vhodná například, když v populaci lze stratifikovat podle pohlaví, věku, ... a výzkumník chce zajistit reprezentaci každé podskupiny;

- **systematický výběr** - ze seřazeného ZS vybereme z prvních *k* prvků náhodně jeden prvek a od něho počítajíc vybereme *k*-tý, 2*k*-tý, ... prvek;
- **vícestupňový shlukový výběr** - často se používá pro získávání informací o veřejném mínění. Chceme například zjistit názory lidí z panelových sídlišť měst určité velikosti. Postup bude takový:

    1. náhodně vybereme vzorek okresů;
    2. z každého vybraného okresu se náhodně vybere určitý počet měst požadované velikosti;
    3. pro tato města se náhodně vybere vzorek jejich sídlišť;
    4. z vybraných sídlišť se náhodně vyberou domácnosti, ve kterých se provede dotazování.

  Tato vícestupňová procedura vypadá komplikovaně, ale ve skutečnosti je velmi efektivní a méně nákladná než prostý náhodný výběr domácností ze sídlišť.

2.3 Výběrové charakteristiky a jejich rozdělení

Mějme základní soubor o rozsahu N jednotek, zajímá nás znak X (např. objem piva v lahvi). Ze základního souboru vybereme n jednotek. Výběr každé jednotky můžeme považovat za náhodný pokus. Zkoumaný znak je vlastně náhodná veličina, která je popsána buď pravděpodobnostní funkcí, nebo funkcí hustoty pravděpodobnosti. Rozdělení pravděpodobností náhodné

veličiny, kterou pozorujeme, se nazývá *statistický model*.

U každé jednotky, která se dostane do výběrového souboru, zjistíme hodnotu zkoumaného znaku x_i ($i = 1, 2, \dots, n$). Tuto hodnotu můžeme chápat jako jednu z možných hodnot náhodné veličiny X_i. Každá z těchto n náhodných veličin má stejné rozdělení jako znak X.

Náhodný výběr je tedy posloupností n nezávislých veličin se stejným rozdělením. Můžeme ho chápat jako vektor $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$. Konkrétní realizaci budeme značit $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$. Naměřené hodnoty $x_1, x_2, \dots, x_n$ nazýváme pozorování nebo také vstupní (empirická) data.

Funkce náhodných veličin X_1, X_2, … , X_n se nazývá *statistika*:

$$T = T(\boldsymbol{X}).$$

Výběrové charakteristiky jsou právě takové statistiky:

Výběrový obecný moment - K-tý výběrový obecný (počáteční) moment je dán vztahem

$$M_k' = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Výběrový průměr je

$$\overline{X} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i$$

.

Platí-li pro prostý náhodný výběr

$$EX_i = EX = \mu$$

$$DX_i = DX = \sigma^2$$

pak střední hodnota výběrového průměru

$$E\overline{X} = EX = \mu$$

a rozptyl

$$D\overline{X} = DX/n = \sigma^2/n$$

Výběrový průměr je tedy vhodný pro odhad střední hodnoty rozdělení.

Výběrový centrální moment - K-tý výběrový centrální moment je dán vztahem

$$M_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$$

Výběrový rozptyl - je dán vztahem

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Výběrová směrodatné odchylka je

$$S = \sqrt{S^2}$$

Výběrová kovariance Nechť při prostém výběru o rozsahu n jsou sledovány dva znaky X a Y. Dostaneme dva **soubory hodnot**

$$\{X_1, X_2, \ldots, X_n\}\,\{Y_1, Y_2, \ldots, Y_n\}$$

Výběrová kovariance je dáma vztahem

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

Výběrový lineární korelační koeficient

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

$r_{XY} \in \langle -1, 1 \rangle$

Úkolem výběrového šetření je podat informaci o neznámé hodnotě charakteristiky základního souboru či o parametrech rozdělení základního souboru na základě náhodného výběru. Charakteristiky základního souboru nazýváme *parametry* (příp. teoretické charakteristiky) a značíme je řeckými písmeny (μ, σ^2, Θ, …). Charakteristiky náhodného výběru nazýváme *výběrové charakteristiky nebo statistiky* a značíme je latinskými písmeny ($\bar{X}$, S_{XY} , r_{XY} ...).

1. Vysvětlete pojem náhodný výběr.
2. Jmenujte základní výběrové charakteristiky souboru.
3. Jak jsou tyto charakteristiky definovány matematicky?

Literatura k tématu:

[1] HINDLS, R. *Statistika pro ekonomy*. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

[2] MACEK, J. *Ekonomická a sociální statistika*. 1. vyd. Plzeň: Západočeská univerzita v Plzni, 2008. ISBN 978-80-7043-642-4.

[3] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 1. vydání. Ostrava: Vysoká škola Báňská - Technická univerzita Ostrava, 2007 [cit. 2017-12-18]. ISBN 80-248-1194-4. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/

[4] ANDĚL, J. *Statistické metody*. 3. vyd. Praha: Matfyzpress, 2003. ISBN 978-80-867-3208-8.

[5] MAREK, L. *Statistika v příkladech*. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

Kapitola 3

Bodový a intervalový odhad

Po prostudování kapitoly budete umět:

- aplikovat možnosti odhadování parametrů základního souboru;
- rozhodnout o volbě statistiky (metoda momentů, metoda maximální věrohodnosti).

Klíčová slova:

Statistika, bodový odhad, intervalový odhad, metoda momentů, metoda maximální věrohodnosti.

3.1 Statistika

Výběr pořizujeme proto, abychom se více dozvěděli o souboru, ze kterého jsme výběr pořídili. Zde se soustřeďme na situaci, kdy známe rozdělení souboru až na jeden nebo více parametrů. Např. víme, že rozdělení souboru je normální, ale neznáme jeho střední hodnotu, případně i rozptyl. Z výběru se snažíme hodnoty těchto neznámých parametrů odhadnout. Předpis, pomocí kterého z výběru vypočteme hodnotu neznámého parametru, se nazývá **statistika**.

Je zřejmé, že parametry základního souboru jsou konstanty, nenáhodné veličiny (které třeba ani neznáme, neboť základní soubor je možná nedostupný statistickému zpracování, popř. vůbec neexistuje), ale výběrové charakteristiky uvedené v předchozí kapitole náhodné veličiny jsou. Mění se výběr od výběru, mění se změnou rozsahu výběru, jsou to tedy tzv. **statistiky**. V tomto případě jsou to **bodové odhady** základních parametrů základního souboru.

Definice
Statistika $T = T(X)$ je funkce výběru X.
Statistika určená pro odhadování se nazývá *odhadová statistika*, pro testování *testová statistika.*

Definice neříká nic o tom, jak statistiku volit vzhledem k jejímu cílovému využití (odhad, test). Její vhodnost či nevhodnost budeme zkoumat později.

3.2 Bodový odhad

Sledujeme rozdělení s hustotou pravděpodobnosti $f(x; \mu)$ s neznámým parametrem μ. Provedli jsme realizaci náhodného výběru $x = (x_1; x_2; \dots ; x_n)$ z tohoto rozdělení a definovali statistiku $T(X)$. *Bodový odhad* parametru μ pro realizaci náhodného výběru x je hodnota statistiky T s dosazenou realizací náhodného výběru x

Bodový odhad (estimátor) parametru μ

Je tedy statistika T, která aproximuje parametr μ s předepsanou přesností.

Pro každou novou realizaci výběru obdržíme jiný bodový odhad. Odtud je zřejmé, že bodový odhad nemůže dát úplně přesnou hodnotu parametru.

Vlastní volbu statistiky jsme zatím nechali stranou. Lze pro ni použít metodu momentů nebo maximální věrohodnost, o které se ještě zmíníme.

Metoda momentů je založena na porovnání momentů základního souboru a výběru. Počet porovnávaných momentů je dán počtem parametrů rozdělení. Závisí-li rozdělení na *S* - parametrech, řešíme soustavu *S* rovnic o *S* neznámých Statistiku je také možno volit heuristicky, potom je však třeba ověřit její vlastnosti.

Oba vzorce pro bodové odhady střední hodnoty a rozptylu:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum \left(x_i - \bar{x}\right)^2 ,$$

se dají odvodit z požadavku, aby udávaly **nevychýlené odhady** příslušných parametrů:

Nevychýlený odhad parametru β je taková statistika β_n, jejíž očekávaná hodnota

$$E(\beta_n) = \beta ,$$

čili je to každá statistika, která statisticky (stochasticky) konverguje k parametru β
V opačném případě se veličina β_n nazývá **odhadem vychýleným**, a to **vpravo** nebo **vlevo**, podle toho, zda $E(\beta_n) - \beta > 0$, resp. $E(\beta_n) - \beta < 0$

V obou případech bodových odhadů střední hodnoty a rozptylu je také splněn požadavek **konzistentnosti (nespornosti) odhadu**:

Konzistentní (nesporný) odhad parametru β je taková statistika β_n, že pro *n* dosti velká je

$$P(\ |\beta_n - \beta\ | \leq \varepsilon) > 1 - \eta,$$

kde $\varepsilon > 0$, $\eta > 0$ jsou jakákoliv (libovolně malá) předem zvolená čísla.

Příklad 3.2.1. Metodou momentů určete neznámý parametr Poissonova rozdělení.

Řešení: Poissonovo rozdělení má pravděpodobnostní funkci:

$$p(x,\lambda) = \frac{\lambda^x}{x!} \cdot e^{-\lambda}$$

Vybereme n prvků $x_1, \ldots, x_n$

$$\mu_1 = \lambda$$

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mu_1 = m_1$$

Tedy:

$$\lambda = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Příklad 3.2.2. Metodou momentů určete neznámý parametr exponenciálního rozdělení.

Řešení: Exponenciální rozdělení má hustotu pravděpodobnosti:

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda \cdot e^{-\lambda x} & x \geq 0 \end{cases}$$

Vybereme n prvků $x_1, \ldots, x_n$

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mu_1 = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_{0}^{\infty} x \cdot \lambda \cdot e^{-\lambda x}dx = \lambda \cdot \int_{0}^{\infty} x \cdot e^{-\lambda x}dx = \left| \begin{matrix} u = x & v' = e^{-\lambda x} \\ u' = 1 & v = -\frac{1}{\lambda} \cdot e^{-\lambda x} \end{matrix} \right| =$$

$$= \left[-x \cdot e^{-\lambda x}\right]_0^{\infty} + \int_{0}^{\infty} e^{-\lambda x}dx = \lim_{x\to\infty} \frac{-x}{e^{\lambda x}} + 0 - \left[\frac{1}{\lambda} \cdot e^{-\lambda x}\right]_0^{\infty} = 0 + \frac{1}{\lambda} = \frac{1}{\lambda}$$

Porovnáme-li tedy opět první počáteční momenty:

$$\mu_1 = m_1$$

$$\frac{1}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\lambda = \frac{n}{\sum_{i=1}^{n} x_i}$$

Metoda maximální věrohodnosti:
Má-li základní soubor frekvenční funkci

$p(x,\theta)$, kde

$\theta = (\theta_1, \theta_2, \ldots, \theta_n)$

jsou parametry rozdělení základního souboru, pak pravděpodobnost, že výběr

$(\xi_1, \xi_2, \ldots, \xi_n)$

bude mít realizaci

$(x_1, x_2, \ldots, x_n)$

je vyjádřena vztahem:

$$P(\xi_1 = x_1, \xi_2 = x_2, \ldots, \xi_n = x_n) = p(x_1,\theta) \cdot p(x_2,\theta) \cdot \ldots \cdot p(x_n,\theta) = \prod_{i=1}^{n} p(x_i,\theta) =$$
$$= L(x_1, x_2, \ldots, x_n, \theta)$$

Funkci L nazýváme funkcí maximální věrohodnosti.
Za nejpravděpodobnější považujeme takovou hodnotu q, při níž má funkce L maximální hodnotu.

Příklad 3.2.3. Metodou maximální věrohodnosti odhadněte neznámý parametr Poissonova rozdělení.

Řešení: Poissonovo rozdělení má pravděpodobnostní funkci:

$$p(x,\lambda) = \frac{\lambda^x}{x!} \cdot e^{-\lambda}$$

Funkce maximální věrohodnosti:

$$L(x_1 x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} \cdot e^{-\lambda} \qquad |\ln$$
$$\ln L = \sum_{i=1}^{n} \left(\ln \lambda^{x_i} - \ln(x_i!) - \lambda\right)$$
$$\ln L = \sum_{i=1}^{n} \left(x_i \cdot \ln \lambda - \ln(x_i!) - \lambda\right)$$
$$\frac{d \ln L}{d \lambda} = \sum_{i=1}^{n} \left(x_i \cdot \frac{1}{\lambda} - 1\right)$$

Položíme-li derivaci rovnu 0:

$$\frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0$$

$$\frac{1}{\lambda}\sum_{i=1}^{n} x_i = n$$

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i$$

3.3 Kritické hodnoty rozdělení

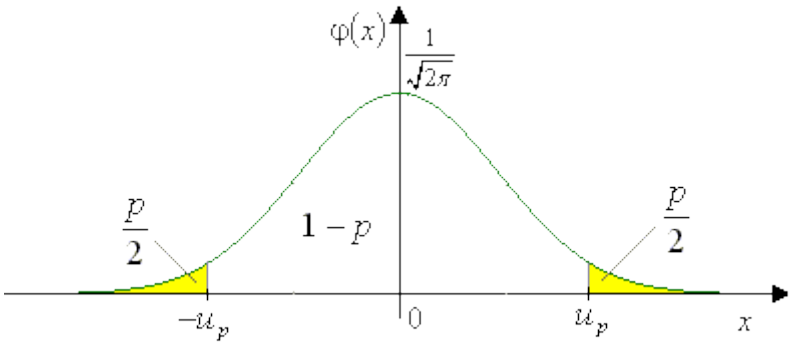
Kritické hodnoty rozdělení na hladině významnosti *p* jsou kvantily, kde index *p* vyjadřuje pravděpodobnost, že náhodná veličina (u symetrických rozdělení její absolutní hodnota), překročí tuto hodnotu.

Užívaná označení:

u_p - kritická hodnota normálního rozdělení na hladině významnosti *p*.

$P(|X| > u_p) = p$, *X*...má normované normální rozdělení $N(0,1)$

Graf kritických hodnot u rozdělení $N(0,1)$



Obrázek 3-1 Graf kritických hodnot normovaného normálního rozdělení [3]

[3] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 2007 [cit. 2017-12-18]. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/

Platí tedy:

$$\Phi(u_p) - \Phi(-u_p) = 1 - p$$
$$\Phi(u_p) - [1 - \Phi(u_p)] = 1 - p$$
$$2\Phi(u_p) = 2 - p$$
$$\Phi(u_p) = 1 - \frac{p}{2}$$

, kde u_p... $\left(1 - \frac{p}{2}\right)$-kvantil normálního rozdělení $N(0,1)$.

Odsud se určí např. $u_{0,05} = 1{,}96$.

Kritické hodnoty některých dalších rozdělení:

$\chi_p^2(n)$ - kritická hodnota rozdělení c^2 s n-stupni volnosti na hladině významnosti p.

$t_p(n)$- kritická hodnota Studentova rozdělení s n-stupni volnosti na hladině významnosti p.

$P(|X| > t_p(n)) = p$, X...má Studentovo rozdělení s n-stupni volnosti.

$F_p(m,n)$- kritická hodnota Fischerova rozdělení s m,n-stupni volnosti na hladině významnosti p.

$P(X > F_p(m,n)) = p$, X...má Fischerovo rozdělení s m,n-stupni volnosti.

3.4 Intervalové odhady parametrů

Intervalový odhad parametru β základního souboru

je interval $< B_1 ; B_2 >$, v němž leží skutečná hodnota parametru s pravděpodobností 1 - p, tzn.

$$P(B_1 \leq \beta \leq B_2) = 1 - p.$$

Interval $< B_1 ; B_2 >$ se nazývá **interval spolehlivosti (konfidenční interval)** pro parametr β **na hladině významnosti *p*** (nebo **se stupněm spolehlivosti 1 - *p***). Hodnoty B_1, B_2 jsou **kritické hodnoty** pro parametr β. Intervaly $(-\infty ; B_1)$ a $(B_2 ; +\infty)$ se nazývají **kritické intervaly**.

Hladina významnosti p je pravděpodobnost toho, že skutečná hodnota odhadovaného parametru **neleží** uvnitř intervalu spolehlivosti. Bývá zvykem volit hodnotu $p = 0{,}1$ nebo $p = 0{,}05$ nebo $p = 0{,}01$.

Stupeň spolehlivosti vyjadřuje pravděpodobnost toho, že skutečná hodnota parametru **leží** v intervalu spolehlivosti.

Interval spolehlivosti lze určit nekonečně mnoha způsoby. Nejčastěji se používá **symetrický oboustranný interval spolehlivosti**, tzn. že parametr β se vyskytuje v jednom z kritických intervalů s pravděpodobností $\frac{\gamma}{2}$.

$$P(\beta < B_1) = P(\beta > B_2) = \frac{\gamma}{2}.$$

Věnujme se nyní intervalovému odhadu nejdůležitějších statistických veličin, střední hodnoty a rozptylu. Ukazuje se, že ten se dá odvodit jako důsledek tzv. **centrální limitní věty**. Uveďme ji v jednom z několika užívaných tvarů bez důkazu:

Nechť $X = X_1 + X_2 + \ldots + X_n$ je náhodná veličina, která vznikla součtem nezávislých náhodných veličin s konečnou střední hodnotou μ a konečným rozptylem σ^2.

Pak náhodná proměnná

$$Y_n = \frac{\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

má pro $n \to \infty$ normální rozložení $N(0, 1)$.

Všimněme si hlavně toho, že o výchozím (základním) souboru není předpokládáno s výjimkou konečnosti jeho základních charakteristik vůbec nic. Hlavně se nic nepředpokládá o jeho rozložení. Přesto je tedy dokazatelné, že výběrové průměry normální rozložení mají. A jejich střední hodnota je rovna střední hodnotě základního souboru (vzpomeňme na bodový odhad střední hodnoty) a rozptyl těchto průměrů je *n-tinou* rozptylu základního souboru.

3.5 Intervalový odhad střední hodnoty

Víme tedy, že veličina

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X} - \mu}{\sigma} \cdot \sqrt{n}$$

má normované normální rozdělení pravděpodobnosti $N(0,1)$.

Nechť $u_{\frac{p}{2}}, u_{1-\frac{p}{2}}$

jsou kvantily normovaného normálního rozdělení, *p* hladina významnosti.

Pak platí:

$$P\left(u_{\frac{p}{2}} \le \frac{\overline{X}-\mu}{\sigma}\cdot\sqrt{n} \le u_{1-\frac{p}{2}}\right) = \Phi\left(u_{1-\frac{p}{2}}\right) - \Phi\left(u_{\frac{p}{2}}\right) = 1-\frac{p}{2}-\frac{p}{2} = 1-p$$

.

Využijeme-li symetrie normovaného normálního rozdělení $\left(u_{\frac{p}{2}} = -u_{1-\frac{p}{2}}\right)$, můžeme předchozí vztah upravit na tvar

$$P\left(\overline{X} - \frac{\sigma}{\sqrt{n}}\cdot u_{1-\frac{p}{2}} \le \mu \le \overline{X} + \frac{\sigma}{\sqrt{n}}\cdot u_{1-\frac{p}{2}}\right) = 1-p$$

,

což je požadovaný oboustranný interval spolehlivosti pro střední hodnotu.

Pokud není známa hodnota rozptylu základního souboru s (tak je tomu většinou), nahradíme ji bodovým odhadem. Intervalový odhad střední hodnoty je pak ve tvaru:

$$P\left(\overline{X} - \frac{s}{\sqrt{n-1}}\cdot u_{1-\frac{p}{2}} \le \mu \le \overline{X} + \frac{s}{\sqrt{n-1}}\cdot u_{1-\frac{p}{2}}\right) = 1-p$$

.

Podmínce asymptotičnosti ovšem nutno vyhovět a užívat vzorec pouze pro $n > 30$.

Pro menší vzorky platí analogický vztah, ale normální normované rozložení je nahrazeno rozložením Studentovým s $n - 1$ stupni volnosti. Kvantil u_p pak nahrazujeme kvantilem $t_p(n\text{-}1)$ Studentova *t*-rozložení:

$$P\left(\bar{x} - \frac{s}{\sqrt{n}}\cdot t_{1-\frac{p}{2}}(n-1) \le \mu \le \bar{x} + \frac{s}{\sqrt{n}}\cdot t_{1-\frac{p}{2}}(n-1)\right) = 1-p$$

Výraz

$\Delta = \frac{\sigma}{\sqrt{n}}\cdot u_{1-\frac{p}{2}} = \frac{s}{\sqrt{n-1}}\cdot u_{1-\frac{p}{2}}$, resp. $\Delta = \frac{\sigma}{\sqrt{n}}\cdot t_{1-\frac{p}{2}} = \frac{s}{\sqrt{n}}\cdot t_{1-\frac{p}{2}}$

je vlastně požadovaná přesnost pro hledaný parametr (běžný je zápis $\mu = \bar{x} \pm \Delta$), která platí pro zvolenou hladinu významnosti *p*. Ze vztahu pro výpočet D, však můžeme naopak určit *n*, které určí potřebný rozsah výběru, jehož charakteristika má požadovanou spolehlivost, např.:

$$n = \left(\frac{\sigma u_{1-\frac{p}{2}}}{\Delta} \right)^2 \quad \text{, resp.} \quad n = 1 + \left(\frac{s.u_{1-\frac{p}{2}}}{\Delta} \right)^2$$

Příklad 3.5.1. Měřili jsme průměr vačkového hřídele na 250 součástkách. Předpokládáme normální rozdělení souboru. Z výsledků měření jsme určili výběrový průměr a výběrovou disperzi x_p = 995,6, s^2 = 134,7. Určete interval spolehlivosti pro střední hodnotu základného souboru při hladině významnosti 5 %.

Řešení: Úlohu vyřešíme v Excelu - z důvodu jednoduchého výpočtu kritické hodnoty normálního rozdělení pomocí předdefinované funkce NORMSINV - v souladu s předchozí teorií:

$$\Delta = \frac{s}{\sqrt{n-1}} u_{1-\frac{p}{2}} = \frac{\sqrt{134,7}}{\sqrt{249}} .\text{NORMSINV}(0,975) = 1,441558$$

Intervalový odhad střední hodnoty je tedy:

$$\langle x_p - \Delta; x_p + \Delta \rangle = \langle 994,1584; 997,0416 \rangle$$

3.6 Intervalový odhad rozptylu

Přistupme nyní k odvození intervalového odhadu disperze. Náhodná veličina, která vznikne součtem normovaných veličin s normálním rozložením, má Pearsonovo rozložení c^2. Stejně tak často tuto součtovou veličinu i označujeme, tedy

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(x_i - \bar{x}\right)^2}{\sigma^2}$$

má rozložení c^2 s *n* stupni volnosti.

Neznáme-li střední hodnotu (a to zpravidla platí), pak náhodná veličina

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(x_i - \bar{x}\right)^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}$$

má Pearsonovo rozložení pro (*n* - 1) stupňů volnosti.

Oboustranný intervalový odhad náhodné veličiny c^2 můžeme zapsat pravděpodobnostní rovnicí:

$$P\left(\chi^2_{1-\frac{p}{2}}(n-1) \le \chi^2 \le \chi^2_{\frac{p}{2}}(n-1)\right) = 1-p$$

, čili

$$P\left(\chi^2_{1-\frac{p}{2}}(n-1) \le \frac{(n-1).s^2}{\sigma^2} \le \chi^2_{\frac{p}{2}}(n-1)\right) = 1-p$$

.

Kritické hodnoty jsou tabelovány.

Po úpravě získáme pravděpodobnostní rovnici pro **intervalový odhad rozptylu** základního souboru v praktičtějším tvaru:

$$P\left(\frac{(n-1).s^2}{\chi^2_{\frac{p}{2}}(n-1)} \le \sigma^2 \le \frac{(n-1).s^2}{\chi^2_{1-\frac{p}{2}}(n-1)}\right) = 1-p$$

Příklad 3.6.1. Určete oboustranný konfidenční interval rozptylu normálně rozloženého základního souboru pro hladiny spolehlivosti 0,90, 0,95 a 0,99, když u výběru s rozsahem $n = 12$ byl zjištěn rozptyl 0,64. Posuďte získané výsledky.

Řešení: Kritické hodnoty Pearsonova rozdělení v excelu vypočteme pomocí předdefinované funkce CHIINV.

Řešení pro spolehlivost 0,90:

$$\frac{n.s^2}{\chi^2_{\frac{p}{2}}(n-1)} \le \sigma^2 \le \frac{n.s^2}{\chi^2_{1-\frac{p}{2}}(n-1)}$$

$$\frac{12.0,64}{\text{CHIINV}(0,05;11)} \le \sigma^2 \le \frac{12.0,64}{\text{CHIINV}(0,95;11)}$$

$$0,358 \le \sigma^2 \le 1,539$$

Zbývající dva případy vyřešíme zcela analogicky.

V této kapitole jsme se věnovali problematice bodových a intervalových odhadů. Získali jsme určitá data, o kterých víme, že pocházejí z normálního rozdělení, ale neznáme hodnoty parametrů normálního rozdělení (střední hodnota a rozptyl). Chtěli bychom na základě získaných dat tyto hodnoty odhadnout. Zajímá nás, jak tento odhad zkonstruovat a, pokud jich máme několik, tak určit ten lepší.

Ve skutečnosti nemusíme pracovat zrovna s normálním rozdělením, data mohou pocházet z libovolného rozdělení, které je závislé na konkrétním parametru. Nabízí se dva způsoby, jak k odhadu parametru přistoupit. Buď nás bude zajímat jedna konkrétní hodnota, kterou budeme považovat za odhad – pak mluvíme o bodovém odhadu. Může nás ale zajímat interval, ve kterém daný parametr s určitou pravděpodobností leží. Pak konstruujeme tzv. intervalový odhad.

1. Měřil se průměr hřídele na 250 součástkách. Předpokládáme normální rozdělení souboru. Z výsledků se určil výběrový průměr a výběrová disperze: $x = 995{,}6$; $s^2 = 134{,}7$. Určete interval spolehlivosti pro střední hodnotu na hladině významnosti 5%.
2. Při měření kapacity sady kondenzátorů bylo provedeno 10 měření s výsledky: 152, 156, 148, 153, 150, 156, 140, 155, 145, 148. Odhadněte interval spolehlivosti pro kapacitu těchto kondenzátorů se spolehlivostí *a*) 90 %, *b*) 95 %.
3. Bylo zkoušeno 30 náhodně vybraných ocelových tyčí k určení meze kluzu určitého druhu oceli. Po zpracování výsledků byla určena její empirická střední hodnota 286,4 Mpa a rozptyl 121 [Mpa^2]. Určete intervalový odhad parametrů základního souboru s 95% spolehlivostí. Kolik vzorků by bylo třeba volit, aby chyba určené střední hodnoty nepřesáhla 2 Mpa?

Literatura k tématu:

[1] HINDLS, R. *Statistika pro ekonomy*. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

[2] MACEK, J. *Ekonomická a sociální statistika*. 1. vyd. Plzeň: Západočeská univerzita v Plzni, 2008. ISBN 978-80-7043-642-4.

[3] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 1. vydání. Ostrava: Vysoká škola Báňská - Technická univerzita Ostrava, 2007 [cit. 2017-12-18]. ISBN 80-248-1194-4. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/

[4] ANDĚL, J. *Statistické metody*. 3. vyd. Praha: Matfyzpress, 2003. ISBN 978-80-867-3208-8.

[5] MAREK, L. *Statistika v příkladech*. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

Kapitola 4

Testování statistických hypotéz

Po prostudování kapitoly budete umět:

- vysvětlit postup při testování statistických hypotéz;
- aplikovat některé konkrétní statistické testy.

Klíčová slova:

Hypotéza, test – parametrický, neparametrický, kritický interval, hladina významnosti, chyba prvního a druhého druhu.

4.1 Statistické hypotézy – základní pojmy

Od statistického šetření neočekáváme pouze elementární informaci o velikosti některých statistických ukazatelů. Používáme je i k ověřování našich očekávání o výsledcích nějakého procesu, k posuzování významnosti změn, které byly způsobeny změnou technologie, apod. Ukážeme, že ač formulace úloh toho typu se liší od formulace úlohy o odhadech parametrů, jde zpravidla vždy o řešení inverzní úlohy o intervalovém odhadu. Zaveďme si však napřed příslušnou terminologii.

Statistická hypotéza

je tvrzení, které se týká neznámé vlastnosti rozdělení pravděpodobnosti náhodné proměnné (i vícerozměrné) nebo jejích parametrů.

Hypotéza, jejíž platnost ověřujeme, se nazývá **nulová hypotéza H_0**.

Proti nulové hypotéze stavíme **alternativní hypotézu H_1**. Ta může být buď **oboustranná,** nebo **jednostranná**. Pak i **testy** jsou buď **oboustranné** nebo **jednostranné**.

Hypotézy se mohu týkat pouze neznámých číselných parametrů rozložení náhodné veličiny, pak jde o **testy parametrické**.

Ostatní typy jsou **testy neparametrické**.

Statistické testy

jsou postupy, jimiž prověřujeme platnost nulové hypotézy. Na základě nich pak **hypotézu buď přijmeme, nebo odmítneme**.

Testovací kritérium

je náhodná veličina závislá na náhodném výběru (též nazývaná **statistika**) mající vztah k nulové hypotéze.

Jednostranné a oboustranné testy se od sebe rozlišují z hlediska alternativní hypotézy, kterou stavíme proti prověřované nulové hypotéze a která může být dvojího druhu, jak plyne z tohoto příkladu:

Nechť nulová hypotéza předpokládá, že ***A* = *B***. V případě, že tuto hypotézu zamítneme, je buď ***A* ≠ *B***, nebo ***A* > *B*** (resp. ***A* < *B***).

a. V prvém případě (***A* ≠ *B***) nebereme zřetel na znaménko rozdílu ***A* - *B***, takže může být buď ***A* - *B* < 0** nebo ***A* - *B* > 0**. V těchto případech používáme **oboustranný test**.
b. V druhém případě, kdy proti hypotéze ***A* = *B*** klademe možnost ***A* > *B*** (resp. ***A* < *B***), používáme **jednostranných testů**.

Pro **kritické hodnoty** testovacího kritéria a_p, b_p platí:

$$P\left(a_p \le X \le b_p\right) = 1 - p .$$

Tyto hodnoty oddělují **interval prakticky možných hodnot** (interval spolehlivosti, konfidenční interval) $\langle a_p, b_p \rangle$ od **kritických intervalů**, v nichž se hodnoty veličiny *X* vyskytují s pravděpodobností *p*, které říkáme hladina významnosti. Nejčastěji volíme $p = 0{,}01$ nebo $p = 0{,}05$.

Pro oboustranné odhady volíme:

$$P\left(X < a_p\right) = P\left(X > b_p\right) = \frac{p}{2},$$

pro jednostranné buď

$$P\left(X < a_p\right) = 0, \quad P\left(X > b_p\right) = p$$

nebo

$$P\left(X < a_p\right) = p, \quad P\left(X > b_p\right) = 0$$

Porovnání hodnoty testovacího kritéria s jeho kritickými hodnotami slouží k rozhodnutí o výsledku testu. Musíme si uvědomit, že nemůžeme mluvit o **dokazování** správnosti či nesprávnosti zvolené hypotézy - to není v možnostech statistické indukce. Závěr testu pouze rozhodne mezi dvěma možnostmi:

- **hypotézu přijímáme** (zamítáme alternativní hypotézu), leží-li pozorovaná hodnota testovacího kritéria v intervalu prakticky možných hodnot. Znamená to, že rozdíl mezi pozorovanou a teoretickou hodnotou testovacího kritéria je vysvětlitelný na dané hladině významnosti *p* náhodností výběru.

- **hypotézu zamítáme** (přijímáme alternativní hypotézu), leží-li pozorovaná hodnota testovacího kritéria v kritickém oboru. Rozdíly považujeme za statisticky významné na zvolené hladině významnosti *p*, tzn., že se nedají vysvětlit pouze náhodností výběru.

Příklady otázek, na které se dá odpovídat pomocí výsledků příslušných statistických testů:

- Má základní soubor (ZS) předpokládanou střední hodnotu?
- Mají dva soubory stejnou disperzi?
- Můžeme předpokládat, že dva výběry pocházejí z téhož ZS?
- Má ZS předpokládané rozdělení?

  atd.

Těmito slovy jistě nebudou technici formulovat své otázky v konkrétním průmyslovém podniku. Bude je ale např. zajímat, zda:

- bylo dodáno uhlí deklarované kvality;
- dva měřicí přístroje pracují stejně přesně;
- se nezměnily provozní podmínky ovlivňující výrobu (např. seřízení obráběcích strojů);
- produkce zmetků v jednotlivých hodinách je rovnoměrná.

Ve shodě s běžnými zvyklostmi definujme:

Nechť ***b*** je pozorovaná, kdežto **β** teoretická hodnota statistiky ***B*** a nechť $\langle a_p, b_p \rangle$ je interval prakticky možných hodnot veličiny ***B*** na 100*p*% hladině významnosti.

Pak říkáme, že rozdíl ***b*** **- β** je

1. **náhodně vysvětlitelný**, když $b \in \langle a_{0,05} ; b_{0,05} \rangle = J_{0,05}$;
2. **statisticky významný**, když $b \in \langle a_{0,01} ; b_{0,01} \rangle = J_{0,01}$;
3. **slabě statisticky významný**, když $b \notin J_{0,05}$, ale $b \in J_{0,01}$.

4.2 Kroky při testování hypotézy

- Formulace výzkumné otázky ve formě nulové a alternativní statistické hypotézy
- Zvolení přijatelné úrovně chyby rozhodování (volba hladiny významnosti *p*)
- Volba testovacího kritéria
- Výpočet hodnoty testovacího kritéria
- Určení kritických hodnot testovacího kritéria
- Doporučení (přijmutí nebo zamítnutí nulové hypotézy H_0)

Poznámky

- *Hladina významnosti je pravděpodobnost, že se zamítne nulová hypotéza, ačkoliv ona platí. Pochopitelně se tato hodnota volí velmi malá, jak již bylo řečeno, nejčastěji 0,05 nebo 0,01.*
- *Jestliže test neindikuje zamítnutí nulové hypotézy H_0, je nesprávné přijmout nulovou hypotézu jako definitivně pravdivou. Správně můžeme pouze prohlásit, že není dostatek dokladů pro zamítnutí nulové hypotézy.*
- *Netvrďme, že data ukazují, že teorie platí/neplatí. Správnější je říct, že data podporují nebo nepodporují rozhodnutí o zamítnutí platnosti nulové hypotézy.*

Při testování hypotéz mohou nastat čtyři možnosti, které popisuje následující tabulka:

Tabulka 4-1 Závěry testování hypotéz

		Závěr testu	
		***H*₀ platí**	***H*₀ neplatí**
Skutečnost	***H*₀ platí**	správný	chyba I. druhu
	***H*₀ neplatí**	chyba II. druhu	správný

Existují tedy dvě možnosti chyby:

- **chyba I. druhu** - nulová hypotéza platí, ale zamítne se;
- **chyba II. druhu** - nulová hypotéza neplatí, ale přijme se.

Přirovnáme-li tuto situaci k medicínskému testování, pak chyba I. druhu znamená falešně pozitivní výsledek (pacient je zdráv, ale testování ukazuje na nemoc), chyba II. druhu odpovídá falešně negativnímu výsledku (pacient je nemocný, ale test to neodhalí).

Pravděpodobnost chyby I. druhu je podmíněná pravděpodobnost, že zamítneme nulovou hypotézu za předpokladu, že platí - označujeme p.

Pravděpodobnost chyby II. druhu je podmíněná pravděpodobnost, že nezamítneme nulovou hypotézu za předpokladu, že neplatí, označujeme p_0:

$$P(\text{chyba I. druhu} \mid H_0 \text{ platí}) = p$$
$$P(\text{chyba II. druhu} \mid H_1 \text{ neplatí}) = p_0$$

Konvenční hodnoty pro p_0 jsou 0,2 nebo 0,1.

Někdy můžeme také mluvit o opačných jevech k chybě I. a II. druhu, tzn. o podmíněné pravděpodobnosti, že neuděláme chybu I. druhu (spolehlivost testu) nebo že neuděláme chybu II. druhu. **Síla testu** odpovídá hodnotě (1 - p_0). Jedná se tedy o podmíněnou pravděpodobnost, že správně odhalíme testem neplatnost nulové hypotézy:

$$P(\text{neuděláme chybu I. druhu} \mid H_0 \text{ platí}) = 1 - p = \text{„spolehlivost“}$$
$$P(\text{neuděláme chybu II. druhu} \mid H_1 \text{ neplatí}) = 1 - p_0 = \text{„síla testu“}$$

Cílem při testování nulové hypotézy je omezit úrovně pravděpodobnosti chyb I. a II. druhu. Jinými slovy - usilujeme o maximalizaci spolehlivosti a síly testu.

Řešené úlohy

Příklad 4.2.1. Testování přiblížíme pomocí analogie se soudním procesem. Má padnout rozhodnutí, zda obžalovaný spáchal či nespáchal zločin.

Řešení: Soudní systém se řídí zásadou, že obžalovaný je nevinen, dokud se nepodaří prokázat opak. Formulace hypotéz má tedy tuto podobu:

H_0: Obžalovaný je nevinen.
H_1: Obžalovaný je vinen.

Různé možnosti vztahu mezi pravdou a rozhodnutím soudu vidíme v tabulce:

Tabulka 4-2 Vztah mezi pravdou a rozhodnutím

		ZÁVĚR SOUDU	
		Obžalovaný je nevinen	**Obžalovaný je vinen**
Skutečnost	Obžalovaný je nevinen	správný	chyba I. druhu
	Obžalovaný je vinen	chyba II. druhu	správný

Uvědomme si, že chyba I. druhu má pro jedince fatální následky. Proto její možnost eliminujeme na nejmenší možnou míru. Soud musí jasně prokázat vinu obžalovaného. Jeho rozhodnutí také podléhají přezkoumání vyšších instancí. Odpovídá to volbě velmi malé hladiny významnosti. V mnoha jiných případech však nevíme zcela přesně, která chyba je pro nás důležitější.

Testování statistických hypotéz umožňuje posoudit, zda experimentálně získaná data vyhovují předpokladu, který jsme před provedením testování učinili. Můžeme například posuzovat, zda platí předpoklad, že určitý lék je účinnější než jiný; nebo například, zda platí, že úroveň matematických dovedností žáků 9. tříd je nezávislá na pohlaví a na regionu. Jako statistickou hypotézu chápeme určitý předpoklad o rozdělení náhodných veličin. Jestliže se tyto předpoklady týkají hodnot parametrů rozdělení náhodné veličiny, pak hovoříme o parametrických hypotézách. V opačném případě se jedná o hypotézy neparametrické. Při testování statistických hypotéz vždy porovnáváme dvě hypotézy. První hypotéza, tzv. nulová (testovaná), je hypotéza, kterou testujeme. Značíme ji obvykle H_0. Druhou hypotézou je tzv. alternativní hypotéza, kterou obvykle značíme H_1.

1. Popište postup při testování statistické hypotézy.
2. Uveďte možnosti, které mohou nastat jako závěr testu.
3. Jaké existují chyby při testování statistických hypotéz?

Literatura k tématu:

[6] HINDLS, R. *Statistika pro ekonomy*. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

[7] MACEK, J. *Ekonomická a sociální statistika*. 1. vyd. Plzeň: Západočeská univerzita v Plzni, 2008. ISBN 978-80-7043-642-4.

[8] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 1. vydání. Ostrava: Vysoká škola Báňská - Technická univerzita Ostrava, 2007 [cit. 2017-12-18]. ISBN 80-248-1194-4. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/

[9] ANDĚL, J. *Statistické metody*. 3. vyd. Praha: Matfyzpress, 2003. ISBN 978-80-867-3208-8.

[10] MAREK, L. *Statistika v příkladech*. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

Kapitola 5

Parametrické testy

Po prostudování kapitoly budete umět:

- vysvětlit postup při testování konkrétních statistických hypotéz;
- použít parametrické testy v typových úlohách.

Klíčová slova:

Parametrický test, hypotézy o rozptylu, hypotézy o střední hodnotě, Studentův test.

5.1 Úvod

Již víme, že pomocí statistické indukce můžeme učinit závěry o populaci na základ výběrového souboru z této populace. V předcházejících kapitolách jsme se zabývali problémem, jak odhadnout prostřednictvím bodového, popř. intervalového odhadu neznámý parametr populace. V této kapitole budeme konstruovat testy, s jejichž pomocí potvrdíme nebo vyvrátíme danou hypotézu o populaci.

Parametrické hypotézy jsou hypotézy o parametrech rozdělení (populace). Můžeme se setkat se třemi typy těchto hypotéz:

1. **Hypotézy o parametru jedné populace** (o střední hodnotě, mediánu, rozptylu, relativní četnosti…)
2. **Hypotézy o parametrech dvou populací** (srovnávací testy)
3. **Hypotézy o parametrech více než dvou populací** (ANOVA …)

Parametrické hypotézy můžeme zapsat jako rovnosti (resp. nerovnosti) mezi testovaným parametrem a jeho předpokládanou hodnotou nebo jako rovnosti (resp. nerovnosti) mezi testovanými parametry. Statistickým hypotézám o jiných vlastnostech populace (tvar rozdělení, závislost proměnných…) se říká neparametrické hypotézy.

Parametrické testy se říká testům, k jejichž odvození je nutné pro daný výběr specifikovat typ rozdělení (v některých případech i některé parametry tohoto rozdělení). (Nejde tedy obecně o libovolné testy parametrických hypotéz.)

5.2 Hypotézy o rozptylu

5.2.1 Test významnosti rozdílu dvou rozptylů (*F*-test)

Předpoklady:

Jsou dány dva výběry o rozsazích n_1, n_2 s rozptyly s_1^2, s_2^2, vybrané ze dvou základních souborů s rozděleními $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$.

Nulová hypotéza:

H_0: $\sigma_1^2 = \sigma_2^2$

Alternativní hypotéza:

H_1: $\sigma_1^2 \neq \sigma_2^2$

Testovací kritérium:

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{n_1(n_2-1).s_1^2}{n_2(n_1-1).s_2^2}$$

má Fisherovo -Snedecorovo rozdělení $F(n_1 - 1, n_2 - 1)$.

Závěr:

Jestliže

$$F > F_{\frac{p}{2}}(n_1-1, n_2-1),$$

zamítáme hypotézu H_0 (přijímáme H_1). Indexy 1, 2 volíme tak, aby testovací kritérium $F > 1$.

Poznámka

V případě, že bychom chtěli prokázat hypotézu H_0 proti hypotéze H_1: $\sigma_1^2 > \sigma_2^2$, použili bychom kritickou hodnotu $F_p(n_1 - 1, n_2 - 1)$

Příklad 5.2.1. Byly sledovány výsledky běhu na 50 m (ve vteřinách) u skupiny desetiletých chlapců a dívek. Posuďte získané výsledky z hlediska vyrovnanosti výkonů v jednotlivých skupinách.

Chlapci:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
10,80	9,30	9,40	9,90	10,20	9,30	9,40	8,90	8,90	9,60	9,70	10,60	9,40	9,50	9,60	10,00	9,30

18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
9,40	8,40	9,80	8,80	9,20	9,50	9,80	9,00	10,50	9,40	9,30	9,90	9,10	9,60	8,70	8,10

Dívky:

1	2	3	4	5	6	7	8	9	10	11	12	13	14
10,70	10,80	10,00	10,60	9,20	10,20	9,90	10,00	9,30	10,20	9,80	10,00	10,00	11,00

15	16	17	18	19	20	21	22	23	24	25	26	27	28
12,00	10,00	10,00	11,20	9,40	10,70	9,30	10,10	9,10	10,20	9,30	10,00	9,40	10,90

Řešení: Hladinu významnosti zvolíme $p = 0,05$.
Určíme potřebné charakteristiky u obou skupin (prohodili jsme pořadí tak, aby vyšlo $F > 1$):

Dívky:	Chlapci:
$n_1 = 28$	$n_2 = 33$
$s_1^2 = 0,4521$	$s_2^2 = 0,3302$

Určíme hodnotu testovacího kritéria:

$$F = \frac{\widetilde{\sigma_1^2}}{\widetilde{\sigma_2^2}} = \frac{n_1(n_2-1)s_1^2}{n_2(n_1-1)s_2^2} = \frac{28.32.0,4521}{33.27.0,3302} \doteq 1,377$$

Kritická hodnota (vypočtená např. v Excelu pomocí předdefinované funkce FINV):
$F_{0,025}(27,32)$ = FINV(0,025;27;32) = 2,0689
Testovací kritérium nepřekročilo kritickou hodnotu, tudíž přijmeme H_0. Mezi rozptyly není statisticky významný rozdíl.

5.3 Hypotézy o střední hodnotě

5.3.1 Test významnosti rozdílu $|m - \mu_0|$

Předpoklady:

Je dán výběr ze základního souboru s rozdělením $N(\mu, \sigma^2)$ o rozsahu n se střední hodnotou m a disperzí s^2.

Nulová hypotéza:

H_0: $\mu = \mu_0$

Alternativní hypotéza:

H_1: $\mu \neq \mu_0$

Testovací kritérium:

$$T = \frac{m - \mu_0}{s} \cdot \sqrt{n-1}$$

má Studentovo rozdělení t (n - 1).

Závěr:

Jestliže $|T| > t_p(n - 1)$, zamítáme hypotézu H_0 (přijímáme H_1).

Poznámka

Volíme-li alternativní hypotézu H_1: $\mu > \mu_0$, pak hodnotu testovacího kritéria srovnáváme s kritickou hodnotou $t_{2p}(n - 1)$.

Příklad 5.3.1. V pivovaru došlo k opravě plnící linky. Na hladině významnosti p = 0,05 ověřte, zda se oprava zdařila, tj., zda linka plní do láhví pivo o objemu 500ml. Výsledky u vybraných vzorků (v mililitrech):

495,2	496,8	502,1	498,5	501	503	500,7
501,5	501,8	499,1	500,9	502,2	501,7	500,4
500,2	501,1	499,9	500,2	501,1	500,8	499,3

Řešení:

$\mu_0 = 500$, tudíž:

H_0: $\mu = 500$

H_1: $\mu \neq 500$

Výpočet základních charakteristik:

$n = 21$ $m = 500{,}3571$ $s = 1{,}77806$

Testovací kritérium:

$$T = \frac{m - \mu_0}{s} \cdot \sqrt{n-1} = \frac{500{,}3571 - 500}{1{,}77806} \cdot \sqrt{20} = 0{,}898$$

Kritická hodnota (vypočteme např. v Excelu pomocí předdefinované funkce TINV):
$t_{0,05}(20)$ = TINV(0,05;20) = 2,086

Závěr:
Testovací kritérium nepřekročilo kritickou hodnotu, tudíž přijmeme H_0. Oprava se zdařila, linka plní lahve správně.

5.4 Test významnosti rozdílu dvou výběrových průměrů (*t*-test)

Předpoklady:
Jsou dány dva výběry o rozsazích n_1, n_2 se středními hodnotami m_1, m_2 a disperzemi s_1^2, s_2^2, které pocházejí ze dvou základních souborů s rozděleními $N(\mu_1;\sigma_1^2)$ a $N(\mu_2;\sigma_2^2)$.
Nulová hypotéza:
H_0: $\mu_1 = \mu_2$
Alternativní hypotéza:
H_1: $\mu_1 \neq \mu_2$

a. jestliže můžeme předpokládat $\sigma_1^2 = \sigma_2^2$ (prověříme *F*-testem), volíme **testovací kritérium:**

$$T = \frac{m_1 - m_2}{\sqrt{n_1 . s_1^2 + n_2 . s_2^2}} \cdot \sqrt{\frac{n_1 . n_2 . (n_1 + n_2 - 2)}{n_1 + n_2}},$$

které má Studentovo rozdělení $t(n_1 + n_2 - 2)$.
Závěr:
Jestliže $|T| > t_p(n_1 + n_2 - 2)$, zamítneme H_0.

b. jestliže předpokládáme $\sigma_1^2 \neq \sigma_2^2$ (prověříme *F*-testem), volíme **testovací kritérium:**

$$T = \frac{m_1 - m_2}{\sqrt{(n_2 - 1) . s_1^2 + (n_1 - 1) . s_2^2}} \cdot \sqrt{(n_1 - 1) . (n_2 - 1)},$$

které má rozdělení, složené ze dvou Studentových rozdělení.
Kritické hodnoty určíme podle vzorce:

$$t_p = \frac{(n_2 - 1) \cdot s_1^2 \cdot t_p(n_1 - 1) + (n_1 - 1) \cdot s_2^2 \cdot t_p(n_2 - 1)}{(n_2 - 1) \cdot s_1^2 + (n_1 - 1) \cdot s_2^2}$$

Závěr:

Jestliže $|T| > t_p$, zamítneme H_0.

Poznámka

t-test používáme např. k ověřování následujících hypotéz:

- *Pocházejí dva vzorky z téhož základního souboru?*
- *Nedopustili jsme se při dvou měřeních, jejichž výsledkem bylo určení dvou středních hodnot m_1, m_2, systematických chyb?*
- *Má určitý faktor vliv na zkoumaný argument? Zde zkoumáme dva vzorky - jeden při působení daného faktoru, druhý bez jeho působení.*

Příklad 5.4.1. Odběratel dostává zářivky od dvou dodavatelů. Při hodnocení kvality zářivek se sleduje také počet zapojení, který snesou zářivky bez poškození. Zkoušky výrobků vedly k těmto výsledkům:

dodavatel *A*: 2139 2041 1968 1903 1952 1980 2089 1915

2389 2163 2072 1712 2018 1792 1849

dodavatel *B*: 1947 1602 1906 2031 2072

1812 1942 2074 2132

Ověřte hypotézu, že kvalita obou dodávek je stejná. Hladinu významnosti volte $p = 0{,}05$.

Řešení: V Excelu vypočteme charakteristiky obou souborů:

$n_1 = 15$ $m_1 = 1998{,}8$ $s_1^2 = 25444{,}69$

$n_2 = 9$ $m_2 = 1946{,}4$ $s_2^2 = 23554{,}25$

Nejdříve provedeme *F*-test:

Testovací kritérium:

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{n_1(n_2-1)\cdot s_1^2}{n_2(n_1-1)\cdot s_2^2} = \frac{15\cdot(9-1)\cdot 25444{,}69}{9\cdot(15-1)\cdot 23554{,}25} = 1{,}0288$$

Kritická hodnota:

$F_{0{,}025}(14,8)$ = FINV(0,025;14;8) = 4,1297

Přijmeme tedy hypotézu o shodě rozptylů $\sigma_1^2 = \sigma_2^2$.

Dále tedy postupujeme jako v případě *a*):

Testovací kritérium:

$$T = \frac{m_1 - m_2}{\sqrt{n_1 \cdot s_1^2 + n_2 \cdot s_2^2}} \cdot \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}{n_1 + n_2}} =$$

$$= \frac{1998{,}8 - 1946{,}4}{\sqrt{15 \cdot 25444{,}69 + 9 \cdot 23554{,}25}} \cdot \sqrt{\frac{15 \cdot 9 \cdot (15 + 9 - 2)}{15 + 9}} = 0{,}756$$

Kritická hodnota:

$t_{0,05}(22)$ = TINV(0,05;22) = 2,074

Závěr:

Testovací kritérium nepřekročilo kritickou hodnotu, přijmeme H_0: $\mu_1 = \mu_2$. Kvalita obou dodávek je stejná.

5.5 Studentův test pro párované hodnoty

Předpoklady:

Ze dvou normálně rozložených základních souborů s parametry μ_1, σ_1^2 a μ_2, σ_2^2 byly vybrány dva výběry se stejnými rozsahy n. Přitom každému prvku prvého výběru x_{1i} odpovídá právě jeden prvek druhého výběru x_{2i}. Vznikly tedy páry (x_{1i} ; x_{2i}), i = 1, ... n.

Nulová hypotéza:

H_0: $\mu_1 = \mu_2$, což lze jinak zapsat: $d = 0$, když d je střední hodnota rozdílů $d_i = x_{1i} - x_{2i}$, tedy:

$$\bar{d} = \frac{\sum_i (x_{1i} - x_{2i})}{n} = \bar{x}_1 - \bar{x}_2 = 0 .$$

Alternativní hypotéza:

H_1: $\mu_1 \neq \mu_2$ nebo tedy: $d \neq 0$

Testovací kritérium:

$$t = \frac{\bar{d}\sqrt{n-1}}{s_d}$$

(s_d je směrodatná odchylka hodnot d_i)

Veličina t má Studentovo rozložení s n - 1 stupni volnosti $t(n - 1)$.

Závěr:

Jestliže $|t| > t_p(n - 1)$, zamítneme hypotézu H_0.

Příklad 5.5.1 Stanovení thiocyanového iontu (SCN-) bylo paralelně provedeno dvěma metodami (Aldridge a Barker) na 12 vzorcích. Srovnejte obě metodiky otestováním výsledků. Hladina významnosti $p = 0,05$.

	1	2	3	4	5	6	7	8	9	10	11	12
Aldridge	0,38	0,56	0,45	0,49	0,38	0,41	0,6	0,36	0,26	0,41	0,43	0,4
Barker	0,39	0,58	0,44	0,52	0,41	0,45	0,59	0,37	0,28	0,42	0,42	0,38

Řešení: Nejprve vytvoříme veličinu d:

Aldridge	0,38	0,56	0,45	0,49	0,38	0,41	0,6	0,36	0,26	0,41	0,43	0,4
Barker	0,39	0,58	0,44	0,52	0,41	0,45	0,59	0,37	0,28	0,42	0,42	0,38
d_i	-0,01	-0,02	0,01	-0,03	-0,03	-0,04	0,01	-0,01	-0,02	-0,01	0,01	0,02

Z tabulky jednoduše vypočteme potřebné charakteristiky:

$$\bar{d} = \frac{\sum_i d_i}{n} = \frac{-0,12}{12} = -0,01$$

(nebo v Excelu pomocí funkce PRŮMĚR)

Obdobně směrodatnou odchylku:

$s_d = 0,018257$

Testovací kritérium:

$$t = \frac{|\bar{d}| \cdot \sqrt{n-1}}{s_d} = \frac{0,01 \cdot \sqrt{11}}{0,018257} \doteq 1,8166$$

Kritická hodnota:

$t_{0,05}(12 - 1) = \text{TINV}(0,05;11) = 2,201$

Testovací kritérium nepřekročilo kritickou hodnotu, přijmeme H_0. Obě metodiky dávají stejné výsledky.

Σ

Při analýze experimentálních dat provádíme nejčastěji testování rozdílů mezi výběrovými soubory za účelem zjištění, zda existuje rozdíl mezi populacemi, z kterých výběry pocházejí. U populací, které odpovídají Gaussovu normálnímu rozdělení, testujeme hypotézy o parametrech μ a σ tohoto rozdělení pomocí parametrických testů. Základní otázkou, kterou klademe obvykle při parametrickém testování experimentálních dat, je otázka, zda se dva výběry shodují ve svém průměru (tj. zda pocházejí z populace s toutéž střední hodnotou), nebo zda sledovaný výběr má určitou konkrétní hodnotu průměru (tj. zda pochází z populace s touto konkrétní střední hodnotou. Další otázkou kladenou při parametrickém testování mohou být dále hypotézy týkající se rozdílu rozptylů mezi dvěma populacemi při hodnocení vlivu pokusných zásahů na variabilitu sledované veličiny. Pro použití parametrických testů je nutno splnit předpoklad normality dat sledovaných veličin. Mezi parametrické testy se řadí především Studentův t-test pro testování rozdílu dvou středních hodnot a F-test pro testování rozdílu dvou rozptylů.

?

1. Dva automaty vyrábějí součástky téhož druhu. Ze součástek vyrobených na prvním automatu jsme změřili $n_1 = 9$ součástek, ze součástek vyrobených na druhém automatu $n_2 = 12$ součástek. Výběrové disperze měřené délky jsou $s_1^2 = 6$ μm, $s_2^2 = 23$ μm. Můžeme přijmout hypotézu o rovnosti disperzí na hladině významnosti 0,05?
2. Každé ze dvou polí bylo rozděleno na 10 lánů a zaseto obilí. Přitom na lánech prvního pole bylo použito speciální americké hnojivo. Výnosy z lánů prvního a druhého pole měly průměry $x_1 = 6$; $x_2 = 5,7$ a rozptyly $s_1^2 = 0,064$; $s_2^2 = 0,024$. Zjistěte na 5 % hladině významnosti, jestli hnojení mělo průkazný vliv na výnosy.
3. U dvou vzorků byly změřeny základní charakteristiky: $n_1 = 10$, $x_1 = 26,5$; $s_1^2 = 4,5$; $n_2 = 5$, $x_2 = 28$; $s_2^2 = 5,8$. Jsou střední hodnoty obou vzorků významně odlišné na hladině významnosti 5 %?

Literatura k tématu:

[1] HINDLS, R. *Statistika pro ekonomy*. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

[2] MACEK, J. *Ekonomická a sociální statistika*. 1. vyd. Plzeň: Západočeská univerzita v Plzni, 2008. ISBN 978-80-7043-642-4.

[3] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 1. vydání. Ostrava: Vysoká škola Báňská - Technická univerzita Ostrava, 2007 [cit. 2017-12-18]. ISBN 80-248-1194-4. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/

[4] ANDĚL, J. *Statistické metody*. 3. vyd. Praha: Matfyzpress, 2003. ISBN 978-80-867-3208-8.

[5] MAREK, L. *Statistika v příkladech*. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

Kapitola 6

Neparametrické testy

Po prostudování kapitoly budete umět:

- vysvětlit postup při testování konkrétních statistických hypotéz;
- použít neparametrické testy v typových úlohách.

Klíčová slova:

Neparametrický test, Komogorovův-Smirnovův test, testy extrémních hodnot.

6.1 Úvod

Jak již bylo zmíněno v předchozích kapitolách, statistické testy dělíme na parametrické a neparametrické. Parametrickým testem rozumíme takový test, pro jehož odvození je nutno specifikovat typ rozdělení, případně jeho parametry. Nejčastěji se setkáváme s předpokladem normality dat. Neparametrická hypotéza je hypotéza o jiných vlastnostech základního souboru (tvar rozdělení, závislost proměnných atd.). Neparametrickým testem tedy rozumíme takový test, pro jehož odvození není nutno specifikovat typ rozdělení.

Neparametrické testy:

- Nemají předpoklady o rozložení vstupujících dat, lze je tedy použít i při asymetrickém rozložení, odlehlých hodnotách, či nedetekovatelném rozložení
- Snížená síla těchto testů je způsobena redukcí informační hodnoty původních dat, kdy neparametrické testy nevyužívají původní hodnoty, ale nejčastěji pouze jejich **pořadí**

Rovněž i zde při testování neparametrických hypotéz proti sobě stojí 2 hypotézy –nulová a alternativní. Nulová hypotéza vyjadřuje tvrzení o základním souboru, které je bráno jako předpoklad při testování (rovnovážný stav). V následujících částech kapitoly uvedeme pouze některé vybrané typy neparametrických testů.

6.2 Kolmogorovův-Smirnovův test dobré shody pro jeden výběr

Předpoklady:
Nechť výsledky pozorování jsou roztříděny do k skupin a v každé skupině je zjištěna skupinová četnost n_{ej} (četnosti experimentální). Uvažujme určité rozdělení, které budeme považovat za model pro náš výběr. Pro každou třídu určíme teoretické, modelové, očekávané četnosti n_{oj} (**j** = 1,...,**k**). Pro empirické i teoretické očekávané rozdělení stanovíme kumulativní četnosti N_{ej} a N_{oj}, **j** = 1,...,k.
Nulová hypotéza:

H$_0$: Základní soubor má očekávané rozložení, tzn. že četnosti N_{ej} a N_{oj} (j = 1,...,k) se liší pouze náhodně.

Testovací kritérium:

$$D_1 = \frac{1}{n} \cdot \max_j \left| N_{ej} - N_{oj} \right| , \quad j = 1, \cdots, k$$

Tato veličina má speciální rozložení, jehož kritické hodnoty jsou tabelovány pro **n** < 40.

Pro **n** ≥ 40 se počítají podle přibližných vzorců.

Pro hladinu významnosti **p** = 0,05 je

$$D_{1;0,05}(n) = \frac{1,36}{\sqrt{n}},$$

pro hladinu významnosti **p** = 0,01 je

$$D_{1;0,01}(n) = \frac{1,63}{\sqrt{n}}$$

Závěr:

Jestliže $\mathbf{D_1} \geq D_{1;\mathbf{p}}$, zamítneme hypotézu $\mathbf{H_0}$.

Příklad 6.2.1

Je dán statistický soubor:

i	1	2	3	4	5	6	7	8	9	10	11	12
obsah Al_2O_3	8-9	9-10	10-11	11-12	12-13	13-14	14-15	15-16	16-17	17-18	18-19	19-20
f_{ei}	2	5	7	19	52	57	72	61	19	14	4	1

Na hladině významnosti 5 % otestujte hypotézu, že soubor má normální rozdělení.

Řešení: Úlohu vyřešíme pomocí Kolmogorovova - Smirnovova testu pro jeden výběr. Nejdříve vypočteme příslušné charakteristiky, tj. parametry normálního rozdělení - střední hodnotu a rozptyl.

Střední hodnota:

$$m = \frac{1}{N}\sum_i x_i f_i = \frac{4417,5}{313} = 14,11342$$

Rozptyl:

$$s^2 = \hat{n}_2 = n_2 - \frac{h^2}{12} = \frac{1}{N}\sum_i (x_i - m)^2 f_i - \frac{h^2}{12} =$$

$$= \frac{1050,224}{313} - \frac{1}{12} = 3,272014$$

Směrodatná odchylka:

$s = \sqrt{3,272014} = 1,808871$

Pomocí parametrů normálního rozdělení můžeme vypočítat očekávané četnosti f_{oi}:

Uvedeme např. výpočet f_o1:

$f_o1 = N.P(8 \le X \le 9) = 313.(F(9) - F(8))$ = (v Excelu) =

= 313*(NORMDIST(9;14,11342;1,808871;1) -

- NORMDIST(8;14,11342;1,808871;1)) =

= 0,6220961

Zbylé očekávané četnosti vypočteme analogicky, viz. tabulka:

i	obsah Al_2O_3	f_{ei}	f_{oi}
1	8 - 9	2	0,6220961
2	9 - 10	5	2,8582712
3	10 - 11	7	9,7422953
4	11 - 12	19	24,64009
5	12 - 13	52	46,25248
6	13 - 14	57	64,446882
7	14 - 15	72	66,661732
8	15 - 16	61	51,187338
9	16 - 17	19	29,176478
10	17 - 18	14	12,343305
11	18 - 19	4	3,8750334
12	19 - 20	1	0,9025231

Dále stačí dopočítat kumulativní četnosti a testovací kritérium:

i	obsah Al_2O_3	n_{ei}	n_{ei} po sloučení	n_{oi}	N_{ei}	N_{oi}	$N_{ei} - N_{oi}$
1	8 - 9	2	7	3,480367	7	3,480367	3,519633
2	9 - 10	5					
3	10 - 11	7	7	9,742295	14	13,22266	0,777337
4	11 - 12	19	19	24,64009	33	37,86275	-4,86275
5	12 - 13	52	52	46,25248	85	84,11523	0,884767
6	13 - 14	57	57	64,44688	142	148,5621	-6,56212
7	14 - 15	72	72	66,66173	214	215,2238	-1,22385
8	15 - 16	61	61	51,18734	275	266,4112	**8,588815**
9	16 - 17	19	19	29,17648	294	295,5877	-1,58766
10	17 - 18	14	14	12,34331	308	307,931	0,069032
11	18 - 19	4	5	4,777557	313	312,7085	0,291476
12	19 - 20	1					
		n =	313			TK:	**0,02744**

Testovací kritérium:

$$D_1 = \frac{1}{n}.\max\left|N_{ei} - N_{oi}\right| = \frac{8,588815}{313} = 0,02744$$

Kritická hodnota:

$$D_{1;0,05}(313) = \frac{1,36}{\sqrt{313}} = 0,076872$$

.

Testovací kritérium nepřekročilo kritickou hodnotu. Daný soubor má normální rozdělení.

Předchozí test ověřoval, zda rozložení výběru neodporuje předpokladu o určitém rozložení základního souboru. Následující test bude ověřovat, shodu rozložení dvou výběrů.

6.3 Kolmogorovův-Smirnovův test dobré shody pro dva výběry

Předpoklady:

U dvou výběrových souborů s rozsahy $\mathbf{n_1}$ a $\mathbf{n_2}$ bylo provedeno roztřídění do **k** skupin a zjištěny kumulativní třídní četnosti pro každou třídu: $\mathbf{N_{1,j}}$ a $\mathbf{N_{2,j}}$.

Nulová hypotéza:

Oba výběrové soubory mají totéž rozložení (pocházejí tedy z téhož základního souboru).

Testovací kritérium:

a) $\mathbf{n_1} = \mathbf{n_2} \leq 40$

$$D_2 = \max_j \left| N_{1j} - N_{2j} \right| \ , \quad j = 1, \cdots, k$$

má speciální rozložení, jeho kritické hodnoty se vyčtou z příslušných tabulek,

b) $\mathbf{n_1} > 40$ a $\mathbf{n_2} > 40$ (i různě velké):

$$D_2 = \max_j \left| F_{1j} - F_{2j} \right|$$

.

Kritické hodnoty se počítají podle vzorců:

pro **p** = 0,05 je

$$D_{2;0,05} = 1,36 \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}$$

a

pro **p** = 0,01 je

$$D_{2:0,01} = 1{,}63 \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}.$$

Závěr:

Jestliže $\mathbf{D}_2 \geq \mathbf{D}_{2:\mathbf{p}}(\mathbf{n}_1,\mathbf{n}_2)$, zamítneme nulovou hypotézu $\mathbf{H}_0$.

Příklad 6.3.1.

Ve dvaceti vybraných závodech byly zkoušeny dva typy filtrů odpadních vod. Bylo zjišťováno, jaké procento nečistot filtr zadrží, a to tak, že nejprve byly instalovány filtry 1. typu a po určité době filtry 2. typu. Výsledky jsou v tabulce:

Množství zadržených nečistot (v %)	10	20	30	40	50	60	70
$n_{1,j}$	1	2	3	8	5	1	0
$n_{2,j}$	0	2	3	2	3	7	3

Zjistěte, jestli se porovnávané filtry kvalitativně liší.

Řešení:

H_0: Dva základní soubory mají totéž rozdělení (porovnávané filtry se kvalitativně neliší).

Volíme hladinu významnosti p = 0,05

množství zadržených nečistot (v %)	$n_{1,j}$	$n_{2,j}$	$N_{1,j}$	$N_{2,j}$	$\lvert N_{1,j} - N_{2,j} \rvert$
10	1	0	1	0	1
20	2	2	3	2	1
30	3	3	6	5	1
40	8	2	14	7	7
50	5	3	19	10	9
60	1	7	20	17	3
70	0	3	20	20	0
Σ =	20	20			

Z tabulky vidíme, že $n_1 = n_2 < 40$, tudíž testovací kritérium:

$$D_2 = \max_j \left| N_{1,j} - N_{2,j} \right| = 9$$

Kritická hodnota:

$D_{2;0,05}(20) = 9$

Závěr:

$D_2 = D_{2;0,05}(20) = 9$, zamítneme H_0.

Filtry se kvalitativně liší.

Existují i neparametrické testy, které neověřují rozložení výběrového souboru. Uveďme test, který se snaží zjistit, zda výběrový soubor neobsahuje údaj zatížený hrubou chybou měření, popř. chybou v zápise. Jde o jeden z **testů extrémních odchylek**.

6.4 Testy extrémních hodnot

6.4.1 Dixonův test extrémních odchylek

Předpoklady:

Ve výběrovém souboru o rozsahu n je $x_1 = \min(x_i)$, resp. $x_n = \max(x_i)$ (např. hodnoty jsou seřazeny podle velikosti od x_1 do x_n).

Nulová hypotéza:

H_0: Hodnota x_1 (nejmenší hodnota), resp. x_n (největší hodnota) se neliší významně od ostatních hodnot souboru.

Testovací kritérium:

$$Q_1 = \frac{x_2 - x_1}{x_n - x_1} \text{, nebo } Q_n = \frac{x_n - x_{n-1}}{x_n - x_1},$$

podle toho, testujeme-li minimální nebo maximální hodnotu ve výběru. Kritické hodnoty $Q_{1;p}$, resp. $Q_{n;p}$ se vyčtou z příslušných tabulek.

Závěr:

Jestliže $Q_1 > Q_{1;p}$, resp. $Q_n > Q_{n;p}$, zamítneme nulovou hypotézu H_0.

Příklad 6.4.1.

Při kalibraci titrační metody k stanovení krevního cukru bylo provedeno 12 paralelních analýz z jednoho vzorku s těmito výsledky:

83	88	84	78	82	82
86	81	98	83	85	80

Otestujte, zda hodnota 98 není chybná.

Řešení:
Dixonovým testem:

x_1 = 78 (nejmenší hodnota)

x_n - 1 = 88 (druhá největší hodnota)

Testovací kritérium:

$$Q_n = \frac{x_n - x_{n-1}}{x_n - x_1} = \frac{98 - 88}{98 - 78} = 0,5$$

Kritická hodnota:

$Q_{12;0,05}$ = 0,376;
$Q_{12;0,01}$ = 0,482

Závěr:
Testovací kritérium překročilo kritickou hodnotu (pro obě zkoumané hladiny významnosti). Zamítáme nulovou hypotézu H_0.

Hodnota 98 se významně liší od ostatních hodnot.

V této kapitole jsme se seznámili s neparametricými testy, kterými testujeme jinou hypotézu o rozdělení základního souboru než je hypotéza o jeho parametru. Tyto typy testů mají nižší sílu (tedy schopnost správně zamítnout ve skutečnosti neplatnou nulovou hypotézu) než testy parametrické, mají vyšší tendenci „nezamítnout“ nulovou hypotézu (v hraničních případech – kdy je testové kritérium velmi blízké kritické hodnotě – mohou vést k nezamítnutí nulové hypotézy, zatímco parametrický test pro stejná data nulovou hypotézu zamítne). Pro stejnou sílu testu je nutná větší velikost výběru než u parametrických testů. Tyto testy mají širší použití než parametrické (lze testovat většinou i ZS hodnot slovních znaků, především

ordinálních, tj. rozlišujících dle relace (např. pořadové testy), některé dokonce i pro hodnoty nominálních znaků, tj. zařazujících jen do skupin. Neparametrické testy jsou nezávislé na rozdělení a na přítomnosti extrémních hodnot a vhodné pro malé výběry. Rovněž všechny obvyklé parametrické testy mají své neparametrické „obdoby“

1. Prověřte na 5% hladině významnosti, zda soubor má rovnoměrné rozdělení, když pro náhodný výběr byly zjištěny tyto četnosti jednotlivých tříd: 10, 21, 0, 8, 12, 6, 8, 13, 11, 11.
2. Při sériové výrobě určitého předmětu byly na podkladě kontrolních měření zjišťovány vadné výrobky vyrobené v každé hodině během jedné směny. Ověřte, zda výskyt vadných výrobků během směny je rovnoměrný.

hodina výroby	1	2	3	4	5	6	7	8
počet zmetků	29	7	27	61	87	110	101	42

Literatura k tématu:

[1] HINDLS, R. *Statistika pro ekonomy*. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

[2] MACEK, J. *Ekonomická a sociální statistika*. 1. vyd. Plzeň: Západočeská univerzita v Plzni, 2008. ISBN 978-80-7043-642-4.

[3] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 1. vydání. Ostrava: Vysoká škola Báňská - Technická univerzita Ostrava, 2007 [cit. 2017-12-18]. ISBN 80-248-1194-4. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/

[4] ANDĚL, J. *Statistické metody*. 3. vyd. Praha: Matfyzpress, 2003. ISBN 978-80-867-3208-8.

[5] MAREK, L. *Statistika v příkladech*. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

Kapitola 7

Regresní a korelační analýza

Po prostudování kapitoly budete umět:

- vysvětlit základní metodiku regresní a korelační analýzy a její smysl;
- zvolit vhodnou regresní funkci a ověřit vhodnost jejího použití.

Klíčová slova:

Regrese, korelace, regresní funkce, metoda nejmenších čtverců odchylek, index korelace.

7.1 Regresní funkce

Důležitou statistickou úlohou je hledání a zkoumání závislostí proměnných, jejichž hodnoty získáme při realizaci experimentů. Vzhledem k jejich náhodnému charakteru reprezentuje nezávisle proměnné náhodný vektor $X = (X_1,...,X_k)$ a závisle proměnnou náhodná veličina Y . Vektor X může být i nenáhodný, jak bývá v aplikacích časté, anebo jsou rozptyly všech složek $X_1,...,X_k$ zanedbatelné vůči rozptylu náhodné veličiny Y .

K popisu a vyšetřování závislosti Y na X užíváme regresní analýzu, přičemž tuto závislost vyjadřuje regresní funkce

$$y = \phi(x,\beta) = E\,(Y|X = x),$$

kde $x = (x_1,...,x_k)$ je vektor nezávisle proměnných (hodnota náhodného vektoru X),

y je závisle proměnná (hodnota náhodné veličiny Y),

$\beta = (\beta_1,...,\beta_m)$ je vektor parametrů, tzv. regresních koeficientů βj, j = 1,...,m, a

$E\,(Y|X = x)$ je podmíněná střední hodnota.

Při vyšetřování závislosti Y na X získáme realizací n experimentů (k +1)-rozměrný statistický soubor $((x_1,y_1),...,(x_n,y_n))$ s rozsahem n, kde y_i je pozorovaná hodnota náhodné veličiny Y_i a x_i je pozorovaná hodnota vektoru nezávisle proměnných X, i = 1,...,n.

Pro určení odhadů neznámých regresních koeficientů βj minimalizujeme tzv. reziduální součet čtverců

$$S^* = \sum_{i=1}^{n} [y_i - \phi(x_i, \beta)]^2$$

a hovoříme o tzv. metodě nejmenších čtverců.

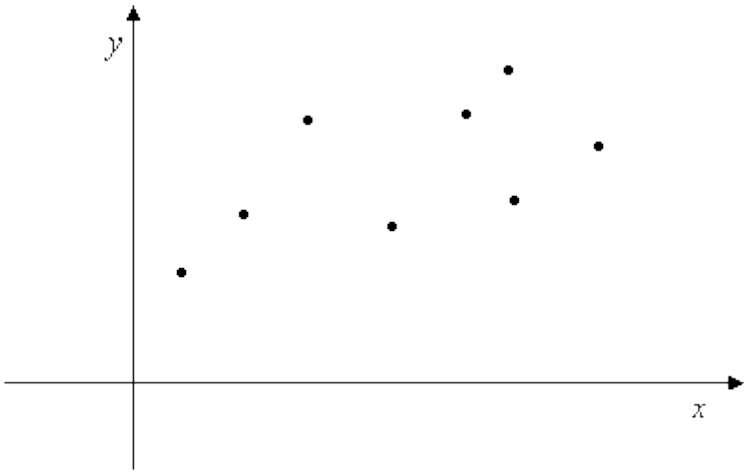
Pro aplikaci regresní analýzy je nezbytné znát tvar (předpis) regresní funkce. Obvykle jej volíme tak, aby co nejvíce odpovídal vyšetřované nebo uvažované závislosti. Bývá zvykem volit regresní funkci s co nejmenším počtem regresních koeficientů, avšak dostatečně flexibilní a s požadovanými vlastnostmi: monotonie, předepsané hodnoty, asymptoty aj. Vychází se přitom povětšinou ze zkušenosti, avšak v současné době se při realizaci regresní analýzy na PC dají často úspěšně použít vhodné databáze regresních funkcí.

7.2 Lineární regrese

Představme si výběr párových hodnot (x_1, y_1), (x_2, y_2), (x_3, y_3),..., (x_n, y_n), získaných (např. změřených) na statistických jednotkách základního souboru. Zde jsou x_i hodnotami závisle proměnné a y_i jsou hodnotami nezávisle proměnné. Zmíněné párové hodnoty můžeme získat zejména dvojím způsobem:

- Hodnoty nezávisle proměnné jsme předem pevně zvolili a k nim jsme „změřili“ příslušné párové hodnoty. V této situaci jsou hodnoty znaku Y pevné (nenáhodné), zatímco hodnoty znaku X považujeme za náhodné veličiny.
- Párové hodnoty „změříme“ na n náhodně zvolených jednotkách základního souboru. V této situaci jak hodnoty znaku X, tak hodnoty znaku Y považujeme za náhodné veličiny.

Graficky lze zobrazit dvojrozměrnou náhodnou veličinu, statistický soubor s dvěma statistickými znaky $(\mathbf{x_i},\mathbf{y_i})$; $\mathbf{i} = 1,2,...,\mathbf{n}$ (korelační pole) například takto:



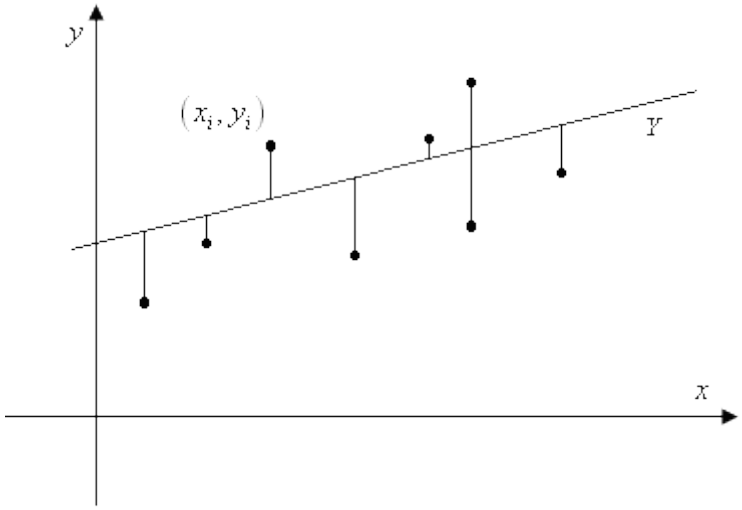
Obrázek 7-1 Korelační pole[4]

Hledejme vyjádření této "**statistické**" závislosti "nejlepším" **funkčním** předpisem. A pro začátek předpokládejme tento předpis lineární:

$$Y = a + bx$$

[4] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 2007 [cit. 2017-12-18]. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/

Jako kritérium pro "nejlepší" funkční předpis vezměme z určitých důvodů (známých už např. Gaussovi v počtu pravděpodobnosti i např. proto, že se takový přístup úspěšně uplatňuje i v jiných situacích) minimalizaci sumy kvadrátů odchylek empirických hodnot *y* od teoretických hodnot získaných pomocí předpisu *Y*:



Obrázek 7-2 Korelační pole – metoda nejmenších čtverců[5]

$$S(a,b) = \sum_{i=1}^{n} (Y_i - y_i)^2 = \sum_{i=1}^{n} (a + bx_i - y_i)^2 = \min$$

Hodnota veličiny *S* závisí na volitelných hodnotách *a* a *b* a je to tedy funkce dvou proměnných. Její extrém se najde nulováním parciálních derivací podle těchto proměnných.

$$\frac{\partial S}{\partial a} = 2.\sum_{i=1}^{n} (a + bx_i - y_i).1 = 0$$
$$\frac{\partial S}{\partial b} = 2.\sum_{i=1}^{n} (a + bx_i - y_i).x_i = 0$$

Po úpravě dojdeme k soustavě lineárních rovnic pro určení *a* a *b*. (V dalším textu budeme někdy zjednodušovat zápis sumační symboliky.)

[5] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 2007 [cit. 2017-12-18]. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/.

$$n.a + b.\sum_i x_i = \sum_i y_i$$

$$a.\sum_i x_i + b.\sum_i x_i^{\,2} = \sum_i x_i y_i$$

Tuto soustavu můžeme vyřešit mnoha způsoby. Například pomocí determinantu matice soustavy, který lze upravit na vyjádření pomocí rozptylu:

$$D = n.\sum_i x_i^{\,2} - \left(\sum_i x_i\right)^2 = n^2.s_x^2,$$

takže koeficienty rovnice přímky nakonec jsou:

$$a = \frac{n \cdot \sum_i y_i \cdot \sum_i x_i^2 - \sum_i x_1 \cdot \sum_i x_i y_i}{n^2 \cdot s_x^{\,2}}$$

$$b = \frac{n \cdot \sum_i x_i y_i - \sum_i x_i \cdot \sum_i y_i}{n^2 \cdot s_x^{\,2}}$$

Po poněkud pracnějších úpravách (s využitím vyjádření centrálních momentů pomocí momentů počátečních):

$$Y = \frac{\sum_i y_i \cdot \sum_i x_i^2 - \sum_i x_i \cdot \sum_i x_i y_i}{n^2.s_x^2} + \frac{n.\sum_i x_i y_i - \sum_i x_i \cdot \sum_i y_i}{n^2.s_x^2}.x$$

$$Y = \frac{1}{s_x^2}.\left(\frac{\sum_i y_i}{n}.\frac{\sum_i x_i^2}{n} - \frac{\sum_i x_i}{n}.\frac{\sum_i x_i \cdot y_i}{n} + \frac{\sum_i x_i \cdot y_i}{n}.x - x.\frac{\sum_i x_i}{n}.\frac{\sum_i y_i}{n}\right)$$

$$Y = \frac{1}{s_x^2}.\left(\bar{y}.\frac{\sum_i x_i^2}{n} - \bar{y}.\left(\frac{\sum_i x_i}{n}\right)^2 + \bar{y}.\left(\frac{\sum_i x_i}{n}\right)^2 - \bar{x}.\frac{\sum_i x_i \cdot y_i}{n} + x.\frac{\sum_i x_i \cdot y_i}{n} - x.\bar{x}.\bar{y}\right)$$

$$Y = \frac{1}{s_x^2}.\left(\bar{y}.s_x^2 + \frac{\sum_i x_i \cdot y_i}{n}\left(x - \bar{x}\right) - \bar{x}.\bar{y}.\left(x - \bar{x}\right)\right)$$

$$Y = \bar{y} + \frac{1}{s_x^2}.\left(\frac{\sum_i x_i \cdot y_i}{n} - \bar{x}.\bar{y}\right).\left(x - \bar{x}\right)$$

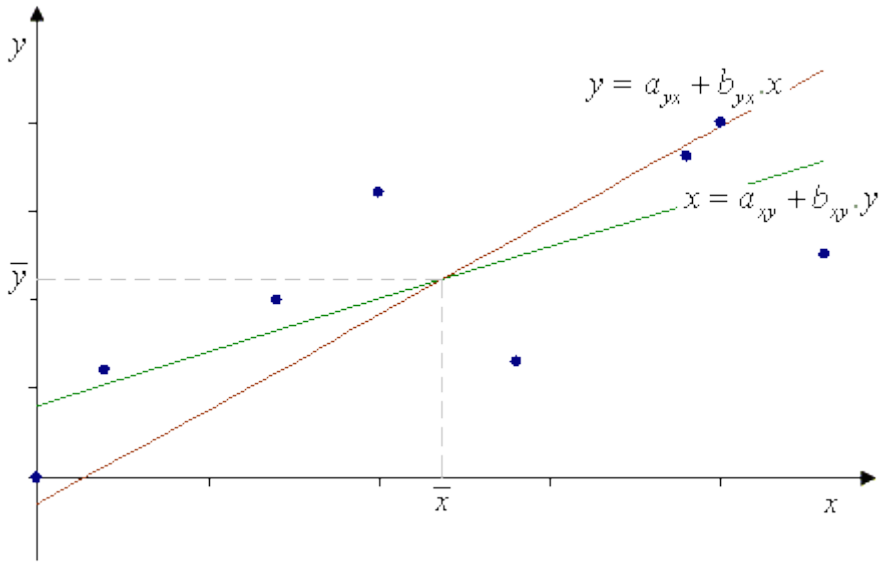
dostáváme jinou podobu **rovnice regresní přímky**, z níž vyplývá, že tato přímka prochází tzv. centrálním bodem $[\bar{x}, \bar{y}]$ (**x**, **y** jsou střední hodnoty proměnných **x**, **y**) a že směrnici přímky, tzv. koeficient regrese, ovlivňuje jak kovariance, tak rozptyl té proměnné, která byla prohlášena za nezávislou:

$$y - \bar{y} = \frac{\mathrm{cov}\,xy}{s_x^2}.(x - \bar{x})$$

Tuto volbu můžeme pochopitelně změnit a tak se dojde analogickou cestou k jiné regresní přímce:

$$x - \bar{x} = \frac{\mathrm{cov}\,xy}{s_y^2}.(y - \bar{y})$$

Vykreslíme-li obě takto získané přímky do jedné souřadnicové soustavy, dostaneme tzv. regresní nůžky:



Obrázek 7-3 Regresní nůžky[6]

.

[6] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 2007 [cit. 2017-12-18]. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/

Směrnice obou regresních přímek $b_{yx} = \frac{\mathrm{cov}\,xy}{s_x^2}$ a $b_{xy} = \frac{\mathrm{cov}\,xy}{s_y^2}$ nazýváme **regresní koeficienty** při závislosti *y* na *x*, resp. *x* na *y* a mají velmi důležitou praktickou interpretaci: udávají **přírůstek závisle proměnné při jednotkové změně nezávisle proměnné**. (Dokažte!) Zároveň umožňují vypočíst koeficient lineární korelace, který jsme výše definovali jako normovaný smíšený moment druhého stupně, vypočíst jiným způsobem:

$$b_{yx}.b_{xy} = \frac{\left(\mathrm{cov}\,xy\right)^2}{s_x^2.s_y^2} = r^2$$

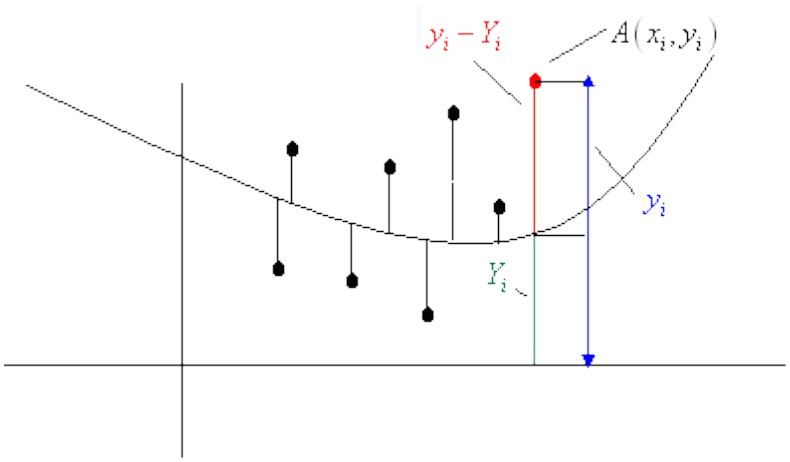
Znaménko přidělíme podle znaménka kteréhokoliv regresního koeficientu, např.:

$$r = sign\left(b_{yx}\right).\sqrt{b_{yx}\cdot b_{xy}}$$

Dá se dokázat, že tento koeficient nabývá hodnoty z intervalu $\langle -1,1 \rangle$ a měří vhodnost lineární funkce vyjádřit statistickou závislost mezi veličinami *x* a *y*. Čím je hodnota koeficientu blíže krajním hodnotám, tím je náhrada těsnější. V případě, že tento koeficient nabývá hodnoty 1 nebo -1, leží všechny body na regresní přímce a závislost veličin *x* a *y* je přesně lineární.

Stanovit stupnici oceňující závislost (závislost "slabá", "střední", "silná") není úkol pro matematika, ale pro profesního odborníka. Podobné stupnice bývají součástí oborových norem.

Lineární průběh nemusí vždy vystihovat vzájemné chování obou složek dvojrozměrné náhodné veličiny. Nic ale nestojí v cestě přirozenému zobecnění předešlých úvah a postupů.



Obrázek 7-4 Korelační pole – nelineární průběh[7]

Uvažujme jako výše korelační pole ($\mathbf{x_i}$,$\mathbf{y_i}$); **i** = 1,2,...,**n** a funkci (kterou volíme pouze jejím charakterem, ale nikoliv jejími parametry, které určují detailně průběh funkce)

$$Y = f(x, a_0, a_1, \ldots, a_k),$$

která by měla vyjádřit vztah mezi složkami *x* a *y*. A hledejme množinu koeficientů a_i tak, aby byl splněn požadavek MNČ (metody nejmenších čtverců):

$$S(a_0, a_1, \cdots, a_k) = \sum_{i=1}^{n} (f(x_i, a_0, \cdots, a_k) - y_i)^2 = \min$$

Řešením soustavy rovnic:

$$\frac{\partial S(a_0, \cdots, a_k)}{\partial a_j} = 0 \quad , \quad j = 0, \cdots, k$$

vzniklé nulováním parciálních derivací funkce *S* podle jednotlivých hledaných koeficientů, dostaneme hledanou regresní funkci. Mohou však nastat problémy algebraického charakteru. Vzniklá soustava rovnic může být velmi nesnadno řešitelná (zvlášť bez použití výpočetní techniky). Proto se zpravidla hledají vhodné regresní funkce pouze mezi tzv. adičními funkcemi:

[7] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 2007 [cit. 2017-12-18]. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/

$$f(x, a_0, \cdots, a_k) = a_0 + a_1 \cdot f_1(x) + \cdots + a_k \cdot f_k(x)$$

Ty totiž vedou k řešení soustavy lineárních rovnic, jak lze snadno ukázat.
Na případy adičních funkcí se často převádějí i funkce multiplikativní, jako je např. funkce mocninná či exponenciální. Linearizace logaritmováním funkčního předpisu však obecně dává pouze suboptimální řešení z hlediska MNČ.
Postup ukážeme na regresní funkci

$$\mathbf{Y = a.e^{bx}}$$

Tuto funkci použijeme za předpokladu, že rychlost růstu závisle proměnné je přímo úměrná její velikosti.
Při určování konstant **a**, **b** zlogaritmujeme funkci:

$$\ln\mathbf{Y} = \ln\mathbf{a} + \mathbf{bx}$$

Jestliže nyní položíme $\mathbf{Z} = \ln\mathbf{Y}$, $\mathbf{a}_1 = \ln\mathbf{a}$, je funkce

$$\mathbf{Z} = \mathbf{a}_1 + \mathbf{bx}$$

lineární v parametrech a můžeme použít již známého postupu. Hledáme tedy minimum funkce

$$\sum_i (a_1 + bx_i - z_i)^2 .$$

Po sestavení soustavy rovnic se můžeme vrátit k původním proměnným. Soustava bude mít tedy tvar:

$$N \cdot \ln a + b \cdot \sum_i x_i = \sum_i \ln y_i$$
$$\ln a \cdot \sum_i x_i + b \cdot \sum_i x_i^2 = \sum_i x_i \cdot \ln y_i$$

Podobně postupujeme např. pro funkci $\mathbf{Y = a.x^b}$ (kde **b** není přirozené číslo) nebo

$$Y = \frac{1}{a + b \cdot \Phi(x)}$$

Poznámka

Hledisko numerické náročnosti regresní analýzy se stává v současné době druhořadé, neboť standardní počítačové programy nabízejí automatizované řešení této úlohy.

Podstatnější problém nastává při měření vhodnosti regresní funkce. Koeficient lineární korelace tu ztrácí svůj význam a je třeba najít jinou míru těsnosti uvažovaného vztahu a daného korelačního pole.

Zaveďme tato označení pro speciálním způsobem definované rozptyly:

$$s_y^2 = \frac{1}{n} \cdot \sum_i \left(y_i - \bar{y} \right)^2$$

$$s_Y^2 = \frac{1}{n} \cdot \sum_i \left(Y_i - \bar{y} \right)^2$$

$$s_{y.x}^2 = \frac{1}{n} \cdot \sum_i \left(y_i - Y_i \right)^2 ,$$

když Y_i je funkční hodnota regresní funkce příslušná *i*-té *x*-ové složce.
Všimněme si, jaký mezi nimi existuje vztah:

$$\begin{aligned} s_y^2 &= \frac{1}{n} . \sum \left(y_i - \bar{y} \right)^2 = \frac{1}{n} . \sum \left(\left(y_i - Y_i \right) + \left(Y_i - \bar{y} \right) \right)^2 \\ &= \frac{1}{n} . \sum \left(\left(y_i - Y_i \right)^2 + \left(Y_i - \bar{y} \right)^2 + 2 . \left(y_i - Y_i \right) . \left(Y_i - \bar{y} \right) \right) = \\ &= s_{y.x}^2 + s_Y^2 + \frac{2}{n} . \sum \left(y_i - Y_i \right) . \left(Y_i - \bar{y} \right) \end{aligned}$$

Dá se dokázat, že poslední výraz na pravé straně je roven nule.

Pak $s_y{}^2 = s_{y.x}{}^2 + s_Y{}^2$ a podíl $\frac{s_Y^2}{s_y^2} = 1 - \frac{s_{ym}^2}{s_y^2} \in \langle 0; 1 \rangle$ bývá používán jako míra těsnosti, vhodnosti regresní funkce (**koeficient determinace**). Udává vlastně, jaká část disperze znaku *y* je způsobena závislostí na *x*. Doplněk koeficientu determinace do jedné znamená podíl náhodné složky na disperzi. Odmocnina

$$I_{yx} = \frac{s_Y}{s_y} = \sqrt{1 - \frac{s_{yx}^2}{s_y^2}}$$

(**index korelace**) má analogickou interpretaci jako koeficient korelace (pro lineární regresní vztah jde o zcela totožný výsledek).

Poznámky

K posouzení míry vhodnosti regresní funkce může sloužit také pouze hodnota

$$s_{y.x}^2 = \frac{1}{n} \cdot \sum_i (y_i - Y_i)^2$$

***reziduální (zbytkový) součet čtverců (rozptyl)**. Nejvhodnější regresní funkcí je pak samozřejmě ta funkce, která má reziduální součet čtverců nejnižší.*

V regresní analýze studujeme vztah mezi jedinou proměnnou (hodnotami statistického znaku) nazývanou závisle proměnnou (někdy vysvětlovanou proměnnou) a obecně několika proměnnými (hodnotami statistických znaků), které nazýváme nezávisle proměnné (někdy vysvětlující proměnné). V případě závislosti dvou znaků mluvíme o jednorozměrné regresi (případně jednoduché regresi). Cílem regresní analýzy je tedy stanovení formy (trendu, tvaru, průběhu) této závislosti pomocí vhodné funkce vystihnout pomocí regresní funkce průběh (trend) závislosti mezi X a Y na základě znalosti dvojic empirických hodnot.

Korelace znamená vzájemný vztah mezi dvěma procesy nebo veličinami. Pokud se mezi dvěma procesy ukáže korelace, je pravděpodobné, že na sobě závisejí, nelze z toho však ještě usoudit, že by jeden z nich musel být příčinou a druhý následkem. To samotná korelace nedovoluje rozhodnout.

V určitějším slova smyslu se pojem korelace užívá ve statistice, kde znamená vzájemný lineární vztah mezi znaky či veličinami *x* a *y*. Tento vztah může být kladný, pokud (přibližně) platí $y = kx$, nebo záporný ($y = -kx$). Míru korelace pak vyjadřuje korelační koeficient, který může nabývat hodnot od −1 až po +1.

Hodnota korelačního koeficientu −1 značí zcela nepřímou závislost, tedy čím více se zvětší hodnoty v první skupině znaků, tím více se zmenší hodnoty v druhé skupině znaků, např. vztah mezi uplynulým a zbývajícím časem. Hodnota korelačního koeficientu +1 značí zcela přímou závislost, např. vztah mezi rychlostí běhu a běžeckou frekvencí kroků sprintera. Pokud je korelační koeficient roven 0, pak mezi znaky není žádná statisticky zjistitelná lineární závislost. Je dobré si uvědomit, že i při nulovém korelačním koeficientu na sobě veličiny mohou záviset, pouze tento vztah nelze vyjádřit lineární funkcí, a to ani přibližně. Může jít např. o nelineární závislost. Z nekorelovanosti náhodných veličin striktně nevyplývá jejich nezávislost, ale naopak z jejich nezávislosti vyplývá i jejich nekorelovanost

1. Vyrovnejte data v tabulce regresní přímkou

x	5	15	25	35	45	55	65
y	3,5	5,2	5,5	6,1	5,9	6,4	7,8

2. Charakterizujte závislost proměnné *y* na *x* regresní funkcí ve tvaru hyperboly

$$y = a + \frac{b}{x}$$

x	55	55	55	65	65	65	75	75	75	85	85	95	95	95
y	3	3,6	4,2	1,8	2,4	3	1,8	2,4	3	1,8	2,4	1,8	2,4	3

Pozn. k řešení použijte vhodný matematický software.

Literatura k tématu:

[1] HINDLS, R. *Statistika pro ekonomy*. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

[2] MACEK, J. *Ekonomická a sociální statistika*. 1. vyd. Plzeň: Západočeská univerzita v Plzni, 2008. ISBN 978-80-7043-642-4.

[3] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 1. vydání. Ostrava: Vysoká škola Báňská - Technická univerzita Ostrava, 2007 [cit. 2017-12-18]. ISBN 80-248-1194-4. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/

[4] ANDĚL, J. *Statistické metody*. 3. vyd. Praha: Matfyzpress, 2003. ISBN 978-80-867-3208-8.

[5] MAREK, L. *Statistika v příkladech*. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

Kapitola 8

Intervaly spolehlivosti a testy hypotéz v regresi a korelaci

Po prostudování kapitoly budete umět:

- určit intervaly spolehlivosti korelačního koeficientu, regresních parametrů a regresního modelu;
- určit pás spolehlivosti (interval spolehlivosti predikovaných hodnot);
- použít testy významnosti v regresní analýze.

Klíčová slova:

Interval spolehlivosti, pás spolehlivosti, regresní model, test významnosti.

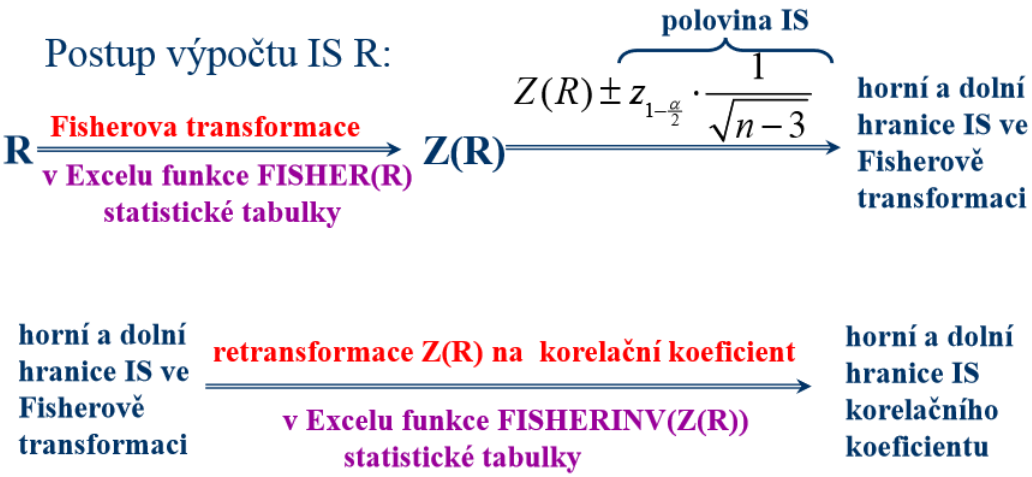
8.1 Interval spolehlivosti korelačního koeficientu (koeficientu determinace)

IS vymezuje interval možných hodnot korelačního koeficientu základního souboru ρ (s pravděpodobností 1 - α).

Jako každou výběrovou statistiku je i výběrový korelační koeficient *r* vhodné doplnit intervalem spolehlivosti, který nám dá informaci o variabilitě tohoto odhadu. Na rozdíl od výpočtu bodového odhadu, který lze vypočítat na datech z různých rozdělení, je však v případě, že chceme rozhodovat o vlastnostech korelačního koeficientu (např. konstruovat interval spolehlivosti pro r nebo testovat hypotézy o r), nutné učinit předpoklad o normalitě náhodných veličin X a Y. Jinými slovy, při výpočtu r předpokládáme realizaci dvourozměrného náhodného vektoru z dvourozměrného normálního rozdělení o rozsahu n. Dalším problémem při konstrukci intervalu spolehlivosti pro r je fakt, že výběrové rozdělení výběrového korelačního koeficientu není normální. Abychom byli schopni interval spolehlivosti zkonstruovat, je třeba použít Fisherovu transformaci na náhodnou veličinu, přičemž transformace je následující:

$$Z(R) = \operatorname{arctgh}(R) = 0.5 \ln \frac{1+R}{1-R}$$

která má přibližně normální rozdělení se střední hodnotou E(Z) = Z(ρ) a rozptylem D(Z) = 1/(n-3).



Obrázek 8-1 Postup výpočtu intervalu spolehlivosti[8]

8.2 Interval spolehlivosti regresních parametrů a modelových hodnot (modelu)

Interval spolehlivosti vyjadřuje úsek na číselné ose, ve kterém se s pravděpodobností 1 - α vyskytuje neznámý parametr β základního souboru

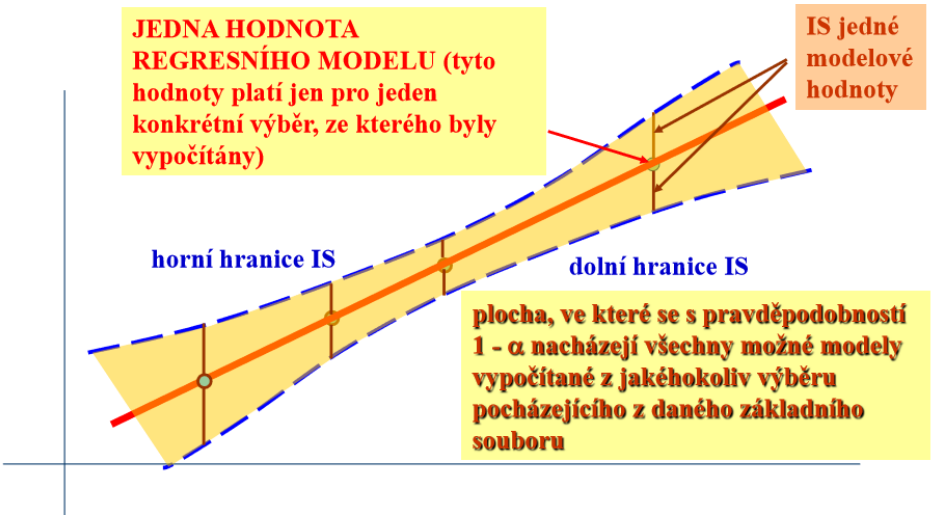
$$\beta_j = b_j \pm t_{\frac{\alpha}{2},n-m} \cdot s_{b_j}$$

Pokud IS obsahuje nulu – tedy dolní hranice je záporná a horní kladná - je daný parametr statisticky nevýznamný. Směrodatné odchylky pro přímku:

$$s_a = \frac{s_{yx}}{\sqrt{n-2}} \cdot \sqrt{1+\frac{\overline{x}^2}{s_x^2}} \qquad s_b = \frac{s_{xy}}{s_x\sqrt{n-2}}$$

Oboustranný konfidenční interval (interval spolehlivosti) je část roviny náhodných proměnných ohraničená dvěma křivkami symetricky položenými kolem regresní přímky.

[8] DRÁPELA, K. Korelace a regrese[online]. 2007 [cit. 2017-12-18]. Dostupné z: user.mendelu.cz/drapela/Statisticke_metody/Prezentace/zakladni/KorelaceRegrese.ppt



Pro model přímky: směr.odchylka reziduí

$$\mu_{y'} = y'_i \pm t_{\frac{\alpha}{2},n-2} \cdot \frac{\sigma}{\sqrt{n-2}} \cdot \sqrt{1 + \frac{n(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

modelová hodnota

polovina IS modelu přímky

Obrázek 8-2 Interval spolehlivosti modelových hodnot[9]

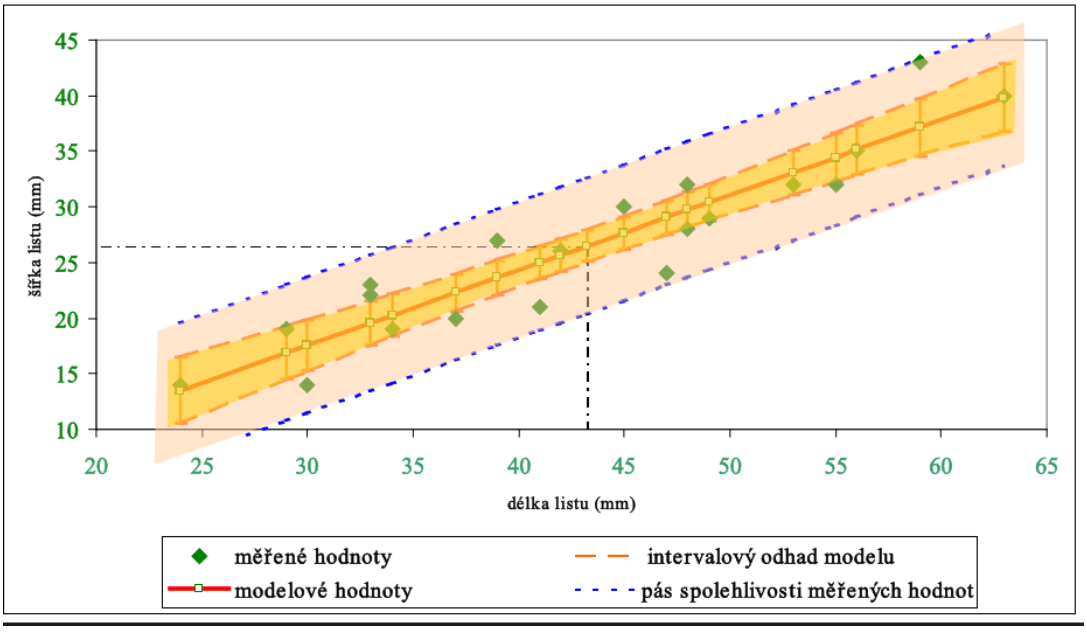
8.3 Interval spolehlivosti Y hodnot a predikovaných hodnot (pás spolehlivosti)

Interval spolehlivost Y hodnot udává rozpětí, ve kterém se budou v základním souboru nacházet hodnoty závisle (vysvětlované) proměnné se zvolenou pravděpodobností 1 - α

[9] DRÁPELA, K. Korelace a regrese[online]. 2007 [cit. 2017-12-18]. Dostupné z: user.mendelu.cz/drapela/Statisticke_metody/Prezentace/zakladni/KorelaceRegrese.ppt.

$$y_{i(min,max)} = y'_i \pm t_{\frac{\alpha}{2};n-m} \cdot \sigma$$

Interval spolehlivosti predikovaných hodnot (pás spolehlivosti)



Obrázek 8-3 Pás spolehlivosti[10]

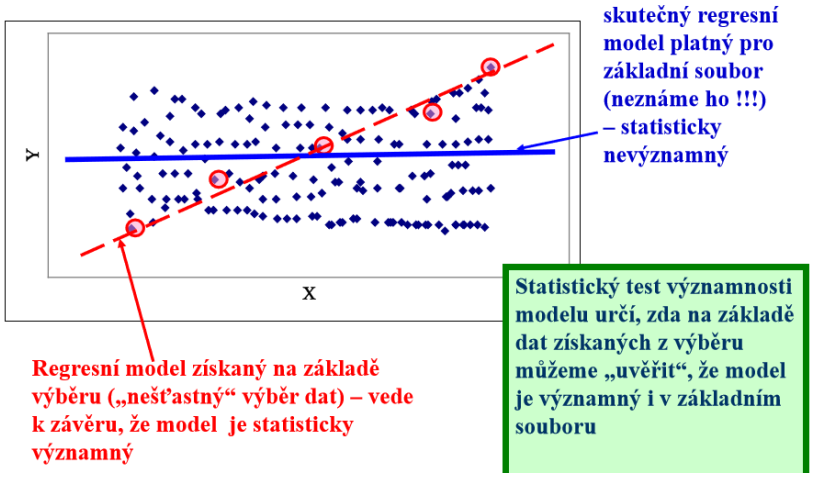
8.4 Testy významnosti v regresní analýze

Testy významnosti rozdělíme na následující varianty:

1. test významnosti korelačního koeficientu
2. test významnosti modelu jako celku
3. test významnosti jednotlivých regresních parametrů
4. test shody lineárních regresních modelů

Následující obrázek odpovídá na otázku proč je nutné testovat významnost v regresní analýze.

[10] DRÁPELA, K. Korelace a regrese[online]. 2007 [cit. 2017-12-18]. Dostupné z: user.mendelu.cz/drapela/Statisticke_metody/Prezentace/zakladni/KorelaceRegrese.ppt



Obrázek 8-4 Testy významnosti v regresní analýze[11]

8.4.1 Test významnosti korelačního koeficientu R

Test významnosti odpovídá na otázku, zda je korelace mezi výběrovými proměnnými (R) natolik silná, abychom mohli tuto korelaci považovat za prokázanou i pro základní soubor (ρ).

Pro párový R:

$$t_R = \frac{R \cdot \sqrt{n-2}}{\sqrt{1-R^2}}$$

KH: $t_{\alpha,n-2}$, n – počet hodnot výběru

Pro násobný R:

$$F_R = \frac{R^2(n-m)}{(1-R^2)(m-1)}$$

KH: $t_{\alpha,n-m}$, m – počet proměnných

Pro parciální R:

$$t_R = \frac{R \cdot \sqrt{n-k-2}}{\sqrt{1-R^2}}$$

[11] DRÁPELA, K. Korelace a regrese[online]. 2007 [cit. 2017-12-18]. Dostupné z: user.mendelu.cz/drapela/Statisticke_metody/Prezentace/zakladni/KorelaceRegrese.ppt

KH: $t_{\alpha,n-k-2}$, m – počet proměnných, k – počet „vyloučených“ proměnných, n – počet hodnot výběru

8.4.2 Test významnosti regresního modelu

V prvé řadě testujeme model jako celek (zda příslušná kombinace nezávisle proměnných statisticky významně zpřesní odhad závisle proměnné oproti použití jejího průměru).

Pomocí analýzy rozptylu:

Tabulka 8-1 Test významnosti regresního modelu jako celku pomocí analýzy rozptylu[12]

Zdroj variability	Součet čtverců odchylek	Počet stupňů volnosti	Průměrný čtverec odchylek (rozptyl)	Testové kritérium
regresní model	$S_{REG} = \sum_{i=1}^{n} (y'_i - y)^2$	$DF_{REG} = m - 1$	$M_{REG} = \frac{S_{REG}}{DF_{REG}}$	$F = \frac{M_{REG}}{M_R}$
reziduum (nevysvětleno regresním modelem)	$S_R = \sum_{i=1}^{n} (y_i - y'_i)^2$	$DF_R = n - m$	$M_R = \frac{S_R}{DF_R}$	
Celkový	$S_C = \sum_{i=1}^{n} (y_i - y)^2$	$DF_C = n - 1$		

Testové kritérium F se porovná s kritickou hodnotou $F_{\alpha,m-1,n-m}$

Dále testujeme jednotlivé parametry modelu $b_0, b_1, b_2, \ldots b_m$ (jestliže je daný parametr nevýznamný, příslušná proměnná x_j nijak nepřispívá ke zpřesnění odhadu závisle proměnné a je v modelu zbytečná)

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_mx_m$$

8.4.3 Test významnosti regresních parametrů

Nulovou hypotézu tohoto testu formulujeme takto:

H_0: $\beta_j = 0$,

tj. j-tý regresní parametr je nevýznamný

[12] DRÁPELA, K. Korelace a regrese[online]. 2007 [cit. 2017-12-18]. Dostupné z: user.mendelu.cz/drapela/Statisticke_metody/Prezentace/zakladni/KorelaceRegrese.ppt

$$t = \frac{b_j - \beta_j}{s_b} \quad \text{pro } \beta_j = 0 \Rightarrow \quad t = \frac{b_j}{s_b}$$

Pokud platí, že $|t| > t_{\alpha 2;n-m}$, potom je j-tý regresní parametr statisticky významný a příslušná proměnná musí zůstat v modelu.

8.4.4 Hodnocení modelu z hlediska výsledků testů významnosti

V rámci hodnocení modelu jako celku mohou nastat tyto případy:

- Je-li výsledek F testu (celého modelu) nevýznamný a současně jsou všechny testované parametry modelu (výsledek T testu) nevýznamné, pak jsou posuzované veličiny lineárně nezávislé nebo je model jako celek nevhodný (nevystihuje variabilitu závisle proměnné).
- Je-li výsledek F testu (celého modelu) významný a současně jsou všechny testované parametry modelu (výsledek T testu) významné, pak je model vhodný (ale nemusí být optimálně navržen).
- Je-li výsledek F testu (celého modelu) významný a současně jsou některé testované parametry modelu (výsledek T testu) nevýznamné, pak je model vhodný (je možné vypustit nevýznamné členy modelu).
- Je-li výsledek F testu (celého modelu) významný a současně jsou všechny testované parametry modelu (výsledek T testu) nevýznamné, pak jde o zvláštní případ způsobený multikolinearitou a je nutné upravit nebo zcela změnit model.

8.4.5 Test shody regresních modelů

V rámci tohoto testu porovnáváme

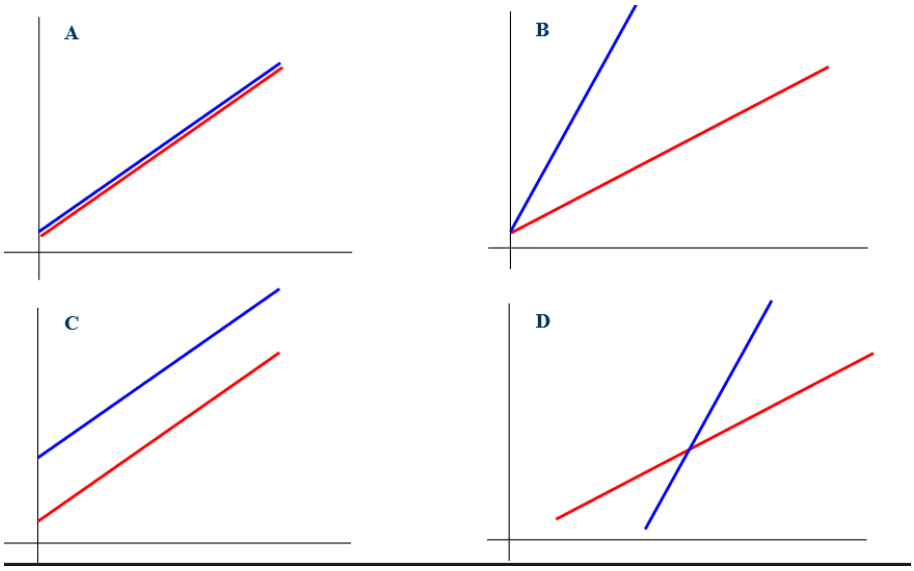
- empirický model (modely) s teoretickým
- dva nebo více empirických modelů mezi sebou

Nulovou hypotézu pak formulujeme tímto způsobem:

H_0: Porovnávané modely jsou shodné (tj. shodují se ve směrnici i v úseku).

Graficky lze celou situaci vyjádřit takto:

Tabulka 8-2 Test shody regresních modelů[13]



Shoda empirického a teoretického modelu:

H_0: Empirický model $y' = a + bx$ pochází ze základního souboru, jehož model $y' = \alpha + \beta x$ je shodný s teoretickým modelem $y'_0 = \alpha_0 + \beta_0 x$, tj. platí $\alpha = \alpha_0$, $\beta = \beta_0$.

Testové kritérium vypočteme takto:

$$t = \frac{a - \alpha_0}{s_a} \qquad t = \frac{b - \beta_0}{s_b}$$

Shoda dvou empirických modelů:

H_0: $\beta_{j,1} = \beta_{j,2}$, tj. regresní koeficienty obou modelů jsou v základním souboru shodné.

Vycházíme z testování shody regresních parametrů dvou lineárních modelů

$y_1 = X_1\beta_1 + \varepsilon_1$ a $y_2 = X_2\beta_2 + \varepsilon_2$

Při tomto testu využijeme tzv. složeného modelu, tj. oba porovnávané výběry sloučíme do jednoho a také pro něj stanovíme parametry stejného modelu jako pro oba dílčí výběry

Testové kritérium vypočteme takto:

[13] DRÁPELA, K. Korelace a regrese[online]. 2007 [cit. 2017-12-18]. Dostupné z: user.mendelu.cz/drapela/Statisticke_metody/Prezentace/zakladni/KorelaceRegrese.ppt.

$$F_C = \frac{(RSC_s - RSC_1 - RSC_2)(n - 2m)}{(RSC_1 + RSC_2) \cdot m}$$

Kde n je celkový počet prvků obou výběrů, tj. $n_1 + n_2$,
RSC_s reziduální součet čtverců složeného modelu,
RSC_1 reziduální součet čtverců prvního modelu,
RSC_2 reziduální součet čtverců druhého modelu.

Na rozdíl od dosud probraných statistik nejsou sdružené regresní přímky skalární veličiny, ale mají tvar funkce. Pro směrnice regresních přímek — regresní koeficienty — lze konstruovat konfidenční intervaly a testovat hypotézy, podobně jako tomu je u dalších statistik skalárního charakteru. Pokud jde o regresní koeficienty, tentokrát jsme se omezili pouze na konfidenční intervaly a testy hypotéz o jednom koeficientu. Pro podmíněnou střední hodnotu lze zkonstruovat oboustranný konfidenční interval, který na rozdíl od předchozích má tvar části roviny omezené dvěma křivkami, jejíž osa souměrnosti je odhadnutá přímka. Pro pozorované hodnoty závisle proměnné lze konstruovat oboustranný tzv. pás spolehlivosti, což je část roviny omezená přímkami, kam pozorovaná hodnota závisle proměnné padá s předem zvolenou pravděpodobností.

1. Co je cílem regresní a korelační analýzy?
2. Pomocí metody nejmenších čtverců odvoďte rovnici regresní přímky.
3. Popište význam intervalu spolehlivosti v regresní analýze.
4. Jaké znáte testy významnosti v regresní analýze?

Literatura k tématu:

[1] HINDLS, R. *Statistika pro ekonomy*. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

[2] MACEK, J. *Ekonomická a sociální statistika*. 1. vyd. Plzeň: Západočeská univerzita v Plzni, 2008. ISBN 978-80-7043-642-4.

[3] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 1. vydání. Ostrava: Vysoká škola Báňská - Technická univerzita Ostrava, 2007 [cit. 2017-12-18]. ISBN 80-248-1194-4. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/

[4] ANDĚL, J. *Statistické metody*. 3. vyd. Praha: Matfyzpress, 2003. ISBN 978-80-867-3208-8.

[5] DRÁPELA, Karel a Jan ZACH. *Statistické metody I*. Brno: Mendelova zemědělská a lesnická univerzita, 1999. ISBN 80-7157-416-3.

[6] DRÁPELA, Karel. *Statistické metody II.* Brno: Mendelova zemědělská a lesnická univerzita, 2000. ISBN 80-7157-474-0.

[7] MAREK, L. *Statistika v příkladech*. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

Kapitola 9

Statistické srovnávání ekonomických jevů

Po prostudování kapitoly budete umět:

- vysvětlit základní pojmy, které se týkají metod statistického srovnávání ekonomických jevů;
- rozdělit a popsat vybrané ukazatele jako statistické veličiny.

Klíčová slova:

Ukazatel (primární, sekundární, absolutní, relativní), ekonomická veličina, statistická metoda.

9.1 Úvod

Popisem a analýzou ekonomických jevů a procesů pomocí ukazatelů se zabývá hospodářská statistika, která je speciální oblastí statistiky. Jejím cílem je nalézt způsoby měření ekonomické skutečnosti (ve formě ukazatelů) a jejího vyhodnocení (např. měření inflace, dynamiky produkce, vývoje kurzů akcií, atd.).

Ukazatele jsou veličiny, s nimiž se denně setkáváme (tisk, TV, rozhlas,…) Seznamujeme se s takovými pojmy jako HDP, průměrná mzda, dovoz, vývoz, produktivita práce, atd. Tyto pojmy jsou vždy doprovázeny čísly, která charakterizují velikost či vývoj příslušného ekonomického jevu. Můžeme se dozvědět, že např. hrubý domácí produkt vzrostl o 3,6%, průměrná nominální mzda vzrostla o 5% a zároveň se zpravidla seznamujeme s tím, zda tyto hodnoty lze hodnotit kladně či záporně, v jakých souvislostech a za jakých podmínek. Všechna tato čísla, hodnocení a předpoklady budoucího vývoje jsou výsledkem práce statistiků v hospodářství.

Tato kapitola není po formální stránce tak náročná jak předchozí text, ale o to více je důležitá v souvislosti se správnou volbou metod analýzy a případných chybných úsudků a souvislostí v probíhajících hospodářských procesech.

9.2 Ukazatel jako statistická veličina

Ukazatel je specifickou statistickou veličinou popisující určitou sociálně ekonomickou skutečnost. Každý ukazatel má tedy svůj věcný obsah a zároveň svoji formálně logickou konstrukci, která ho řadí mezi statistické veličiny. Chceme-li ukazatele definovat, musíme se zaměřit na jeho předmětnou, ale i formálně logickou definici.

Podíváme-li se na ukazatele z **předmětného (obsahového) hlediska,** je zřejmé, že se jedná o pojmy, které používá i ekonomická teorie. Ekonomická teorie definuje své kategorie a jejich vztahy verbálně, často bez ohledu na to, zda jsou tyto pojmy a vztahy kvantifikovatelné či nikoli. Statistika ale naopak potřebuje reálně existující ekonomické jevy a procesy měřit, vyjádřit jejich velikost, intenzitu pomocí číselné charakteristiky tak, aby daný ukazatel co nejlépe odrážel skutečnost, popisovanou daným pojmem. Při konstrukci ukazatelů obsahové náplni pojmů ekonomické teorie.

Logický postup „pojem →" není však v praxi vždy uplatňován. Často se můžeme setkat s opačným postupem, kdy uměle vytvořenému ukazateli se přiřadí název a vypovídající schopnost, která ne vždy odpovídá podstatě měřeného pojmu. Využití takového ukazatele v praxi je velmi problematické, neboť dezorientuje uživatele, který zpravidla nezkoumá konstrukci ukazatele, ale

podle názvu usuzuje na jeho vypovídající schopnost. Kvalita vztahu pojmu ekonomické teorie a statistického ukazatele zároveň předpokládá, že ekonomické teorie se bude zamýšlet nad možnostmi kvantifikace kategorií, které definuje. Je zřejmé, že v praxi sledování ekonomických jevů a procesů nevystačíme s pojmy, které nám nabízí ekonomické teorie, neboť v řadě případů je nezbytné kvantifikovat jevy specifické, detailní. Vždy je však třeba dbát na to, aby existoval soulad názvu a vypovídající schopnosti ukazatele a aby tudíž konstrukce ukazatele byla smysluplná.

Formálně logická definice ukazatele nám dovádí k problému vztahu základních statistických pojmů, jako jsou statistická jednotka, statistický znak, statistický soubor a pojmu ukazatele. Je zřejmé, že tyto pojmy spolu souvisejí, ale jejich vzájemný vztah není zřetelný.

Statistický ukazatel je statistickou charakteristikou, je tedy funkcí hodnot znaku definovaných na statistických jednotkách, popř. je funkcí těchto charakteristik.

Statistický ukazatel je ale specifickým typem statistických charakteristik, neboť využívá jen omezeného počtu funkčních předpisů (nejčastěji úhrnu), statistických jednotek, statistických znaků, a to těch, které mají sociálně ekonomický charakter. To ostatně plyne ze specifického postavení pojmů „ukazatele" v české terminologii. Ukazatelem se v tomto smyslu rozumí veličina, vypovídající o nějaké sociálně ekonomické hromadné skutečnosti. V ostatních disciplínách se pojmu ukazatel nepoužívá.

V terminologii „západní" statistiky můžeme najít v překladech odpovídající termín „indicator", nebo „indicateur", jejichž význam však není totožný s obsahem našeho pojmu „ukazatel". Výše uvedených cizojazyčných termínů se používá spíše ve významu rozhodující veličiny pro charakterizování určitého stavu nebo jevu. Pro veličiny odpovídající významově našemu ukazateli nepoužívá „západní" statistika speciálních termínů, ale obecných termínů typu veličina, statistika, proměnná apod.

Z toho, co jsme uvedli o podstatě ukazatele jako statistické charakteristiky a o pojetí termínu ukazatel v „západní" statistice, plyne, že ukazatel je proměnnou veličinou. Zároveň víme, že každá proměnná veličina nabývá vždy určitých hodnot v závislosti na své definici a že o ukazatelích se vždy hovoří v souvislosti s číselnými hodnotami. Vzniká tedy otázka, jak z ukazatele jako proměnné veličiny získáme číslo, tj. konkrétní hodnotu ukazatele neboli **údaj.**

Statistický ukazatel je statistickou charakteristikou a je zřejmé, že toto konstatování implicitně předpokládá, že statistický soubor je obecně prostorově a časově vymezen. Vezmeme-li např. ukazatel „odpracovaná doba", pak tento ukazatel je v metodických předpisech vymezen, jakou úhrn doby odpracované pracovníky (popř. dělníky) podniku (popř. závodu) v měsíci (popř. čtvrtletí, roce). Jde tedy o popis ukazatele, kde je obecně definován čas – měsíc a prostor – podnik. Jestliže přesně

definujeme tento čas a prostor (např. únor 2005, podnik Alfa), dostaneme konkrétní hodnotu ukazatele, tj. údaj.

9.2.1 Typy a vlastnosti ukazatelů

Vraťme se ještě k definici ukazatele. Uvedli jsme, že ukazatel je statistickou charakteristikou, tj. funkcí hodnot znaku definovaných na statistických jednotkách, popř. funkcí těchto charakteristik. Tyto dvě části definice jsou významné pro základní členění ukazatelů na primární – prvotní a sekundární – odvozené.

Primární ukazatele jsou ukazatele přímo zjišťované, neodvozené, např. odpracovaná doba, počet pracovníků k určitému datu, stav zásob apod. Jedná se o ukazatele, kde lze jednoznačně určit typ charakteristiky, statické jednotky i statistického znaku.

Druhým typem ukazatelů jsou ukazatele sekundární, odvozené, které mohou vznikat trojím způsobem:

- Jako funkce (zpravidla rozdíl nebo podíl) různých primárních ukazatelů. Např. zisk, přidaná hodnota, doba obratu zásob apod.,
- Jako funkce různých hodnot téhož primárního ukazatele. Zde je možné jmenovat všechny časové průměry, ukazatele struktury, hrubého obratu,
- Jako funkce dvou primárních ukazatelů, kde alespoň u jednoho pracujeme s více hodnotami, resp. Jako funkce více než dvou primárních ukazatelů (tj. kombinací předchozích postupů). Jako příklad lze uvést relativní ukazatele, kde alespoň jeden je časovým průměrem (produktivita práce na pracovníka, vybavenost práce apod.), resp. Funkcí více primárních ukazatelů (ziskovost produkce, podíl přidané hodnoty na celkové produkci apod.).

V souvislosti s členěním ukazatelů na primární a sekundární vzniká často otázka, kam zařadit indexy, absolutní rozdíly a jiné podobné míry rozdílnosti. Jsou tyto veličiny také ukazateli či nikoli?

Indexy, absolutní rozdíly a další míry rozdílnosti jsou nástroji srovnávání a nástroji analýzy výsledků srovnání. Ukazatele samy o sobě vypovídají o nějaké skutečnosti, ale nehodnotí ji, zatímco indexy a absolutní přírůstky měří rozdílnost dvou hodnot téhož ukazatele, analytické míry rozdílnosti pak tuto odlišnost vyhodnocují.

Poznámka: Pokud někteří autoři považují indexy, absolutní rozdíly a analytické míry rozdílnosti za ukazatele, vymezují jim zpravidla specifické místo, např. členěním ukazatelů na pravé a nepravé. Pravé ukazatele popisují určitý ekonomický jev, určitou skutečnost. Nepravé ukazatele slouží k srovnání hodnot pravých ukazatelů a vyhodnocování zjištěné rozdílnosti.

Nepravé ukazatele jsou pak nástrojem srovnání a analýzy rozdílnosti hodnot pravých ukazatelů. Nepravé ukazatele jsou vždy sekundární.

Vedle třídění ukazatelů na primární a sekundární je důležité i členění ukazatelů na **absolutní** a **relativní** (podrobněji tyto druhy ukazatelů probereme v následující kapitole). Absolutní ukazatele vyjadřují velikost určitého jevu bez vztahu k jinému jevu. Do této skupiny patří všechny ukazatele primární (resp. Ty, které jsou úhrnem hodnot znaku), ale i některé ukazatele sekundární (časové průměry, ukazatele hrubého obratu, rozdílové ukazatele jako zisk, přidaná hodnota apod.). Relativní ukazatele vyjadřují velikost jednoho jevu na měrovou jednotku jiného jevu. Relativní ukazatele jsou vždy sekundární, neboť vznikají jako podíl absolutních (primárních i sekundárních) ukazatelů.

Jestliže členění ukazatelů na primární a sekundární, resp. na absolutní a relativní je vyčerpávajícím, pak členění stejných veličin na **extenzivní** a **intenzivní** opomíjí skupinu tzv. strukturních ukazatelů. Extenzivní ukazatele (ukazatele množství) jsou ukazatele absolutní, intenzivní ukazatele (ukazatele úrovně) však nepokrývají celou skupinu relativních ukazatelů, ale pouze jen ty, které vyjadřují intenzitu určitého jevu. Vyčerpávající popis ukazatelů tedy získáme, připojíme-li k extenzivním a intenzivním ukazatelům ještě ukazatele struktury. Členění ukazatelů na extenzivní a intenzivní je důležité především v indexní teorii.

Ukazatele se dále zpravidla třídí na okamžikové a intervalové. Toto členění již definuje vlastnost ukazatele a předurčuje způsob jeho shrnování v čase. Třídění ukazatelů na okamžikové a intervalové není opět vyčerpávající, týká se jednoznačně pouze primárních ukazatelů a rozdílových sekundárních ukazatelů (tj. absolutních ukazatelů). U ostatních sekundárních ukazatelů (relativních ukazatelů) nelze definovat, zda ukazatel je okamžikový či intervalový, ale pouze určit jeho chování v čase, tzn., zda s prodlužováním časového intervalu se bude jeho hodnota měnit (růst nebo klesat) či nikoli. To záleží na chování v čase primárních, resp. Sekundárních ukazatelů, z nichž je příslušný relativní ukazatel složen (např. hodnota ukazatele doby obratu zásob, definovaného jako podíl průměrného stavu zásob /jehož hodnota se s prodlužováním časového intervalu nemění/ a nákladů /jejichž hodnota s prodlužováním časového intervalu roste/ s prodlužováním časového intervalu klesá).

Výše uvedené členění ukazatelů z hlediska jejich chování v čase bývá již spíše považováno za popis vlastnosti ukazatele, neboť skutečnost, zda ukazatel je okamžikový či intervalový, je důležitá pro operace s ukazateli. Za typickou vlastnost ukazatelů je však uváděna jejich **stejnorodost, srovnatelnost a shrnovatelnost.**

Stejnorodost statistických ukazatelů je vlastnost, kterou zdůrazňujeme především v indexní teorii, ale je zřejmé, že má širší význam, neboť je první a výchozí podmínkou možnosti shrnování dílčích hodnot určitého ukazatele.

Stejnorodost statistických ukazatelů je dána povahou statistických jednotek. Kritériem stejnorodosti je pak statistický znak, který na daných jednotkách sledujeme. Stejnorodost statistických ukazatelů je relativní a závisí na způsobu vymezení souboru jednotek pro daný účel zkoumání. To, co se v jedné situaci jeví jako soubor homogenních jednotek, je v jiné situaci souborem nestejnorodých jednotek. Obecně je možné říci, že absolutní ukazatel je stejnorodý tehdy, jestliže má věcný smysl shrnovat jeho dílčí hodnoty součtem. Relativní ukazatel je stejnorodý jen tehdy, když jsou stejnorodé oba absolutní ukazatele, z nichž se skládá, resp. Lze-li dílčí hodnoty relativního ukazatele shrnovat průměrem. Pokud toto neplatí, je ukazatel nestejnorodý.

Srovnatelnost statistických ukazatelů je vlastnost, která má vazbu na tvorbu relativních ukazatelů a indexů. Za srovnatelné považujeme takové ukazatele, jejichž srovnáním (resp. Srovnáním jejich hodnot) získáme smysluplnou veličinu, tj. smysluplný relativní ukazatel, resp. Index (např. ukazatel produktivity práce, strukturní ukazatele, časové, prostorové a druhové indexy apod.) Za nesrovnatelné tedy považujeme takové ukazatele, jejichž srovnání, resp. Srovnání jejich hodnot nemá smysl z hlediska rozdílného druhového, časového či prostorového vymezení statistických jednotek (např. nemá smysl konstruovat relativní ukazatel srovnávající počet narozených dětí a obrat zahraničního obchodu, srovnávat cenu dvou naprosto rozdílných výrobků apod.)

Shrnovatelnost je poslední, ale určitě ne nejméně důležitou vlastností ukazatelů. Shrnovatelnost bezprostředně souvisí se stejnorodostí, jež je základním předpokladem smysluplnosti shrnování dílčích hodnot určitého ukazatele. Shrnovatelnost vyjadřuje schopnost ukazatele určit jeho celkovou hodnotu na základě jeho dílčích hodnot. Z tohoto hlediska potom rozlišujeme ukazatele **přímo shrnovatelné, nepřímo shrnovatelné a neshrnovatelné.**

Přímo shrnovatelné jsou takové ukazatele, jejichž souhrnnou hodnotu můžeme určit výlučně z dílčích hodnot daného ukazatele (např. odpracovanou dobu za rok určíme jednoznačně na základě znalosti měsíčních hodnot).

Nepřímo shrnovatelnými rozumíme takové ukazatele, kde k určení souhrnné hodnoty daného ukazatele musíme znát nejen dílčí hodnoty tohoto ukazatele, ale i dílčí hodnoty jiného ukazatele (typické pro všechny relativní ukazatele).

Za neshrnovatelné považujeme takové ukazatele, kde souhrnnou hodnotu daného ukazatele nelze určit ani při znalosti dílčích hodnot daného ukazatele, ale ani dalších ukazatelů. Souhrnnou hodnotu ukazatele můžeme určit výlučně na základě znalosti individuálních dat (jedná se o malou skupinu ukazatelů, kde jakou charakteristika vystupuje např. medián).

Z podstaty časového, prostorového a věcného (druhového) vymezení ukazatele, resp. Jeho hodnoty plyne, že rozlišujeme časové, prostorové a druhové shrnování hodnot ukazatelů a zároveň platí, že neexistuje obecný princip shrnování hodnot určitého ukazatele, ale že dílčí hodnoty se mohou shrnovat rozdílně v čase, v prostoru či druhově (např. okamžikové ukazatele se v čase shrnují průměrem, v prostoru součtem).

Vedle elementárního statistického zpracování dat se hromadné jevy analyzují tzv. srovnáváním různých ukazatelů. Statistický ukazatel - proměnná veličina (znak), která kvantitativně popisuje hromadný jev. Z věcného hlediska se ukazatele se dělí na: extenzitní ukazatele - objem množství, velikost, hodnota, … a intenzitní ukazatele což je poměr 2 extenzitních ukazatelů (prům. náklady = celkové náklady / objem produkce, produktivita = hodnota produkce / počet pracovníků, tržba = cena x velikost produkce. Pro srovnávání hodnot ukazatelů je důležitá jejich stejnorodost z hlediska jejich věcného obsahu. Jako prostředky ke srovnávání hodnot ukazatelů slouží indexy a diference. Rozlišujeme absolutní srovnání (pomocí rozdílů) a relativní srovnání (pomocí podílů - indexů).

1. Co je cílem statistického srovnávání ekonomických jevů?
2. Jaké znáte typy ukazatelů a jak se dále dělí?
3. Co znamená stejnorodost, srovnatelnost a shrnovatelnost ukazatele?

Literatura k tématu:

[1] HINDLS, R. *Statistika pro ekonomy*. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

[2] MACEK, J. *Ekonomická a sociální statistika*. 1. vyd. Plzeň: Západočeská univerzita v Plzni, 2008. ISBN 978-80-7043-642-4.

[3] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 1. vydání. Ostrava: Vysoká škola Báňská - Technická univerzita Ostrava, 2007 [cit. 2017-12-18]. ISBN 80-248-1194-4. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/

[4] ANDĚL, J. *Statistické metody*. 3. vyd. Praha: Matfyzpress, 2003. ISBN 978-80-867-3208-8.

[5] MAREK, L. *Statistika v příkladech*. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

Kapitola 10

Indexy a absolutní rozdíly jako nástroj srovnávání a analýzy

Po prostudování kapitoly budete umět:

- základy použití indexní analýzy;
- vysvětlit způsoby porovnávání získaných ekonomických hodnot a určit míru a směr jednotlivých vlivů.

Klíčová slova:

Porovnávání (absolutní, relativní), index (jednoduchý, složený, souhrnný).

Indexní analýza se používá při analyzování sociálně ekonomických ukazatelů. Pomocí indexů můžeme porovnávat vzájemně odlišné údaje a určovat jejich vzájemné vztahy. Jak bylo uvedeno v předchozí kapitole, ukazatel je statistická veličina popisující některý ze sociálně ekonomických jevů.

Ukazatele rozdělujeme na:

- extenzivní - Jsou to ukazatele vyjadřující velikost zkoumaného jevu. Charakterizují množství, rozsah, objem, a podobně. Tyto ukazatele můžeme při počítání souhrnů sčítat. Pro označení těchto ukazatelů používáme q (případně Q). Také je označujeme jako ukazatele množství.
- intenzivní - Jedná se o ukazatele vzniklé jako podíl dvou extenzivních ukazatelů. Může se jednat o ukazatele uvádějící cenu za kus, průměrnou mzdu, výnos na hektar, atd. Pro jejich označení se používá znak p, nebo c. Označují se také jako ukazatele úrovně. Tyto ukazatele se nedají při souhrnech sčítat.

Hodnoty ukazatele se v čase mění, a proto je rozlišujeme z časového hlediska. Ukazatele číslujeme dolními indexy, takže 0 značí původní časový údaj, nebo období, vzhledem ke kterému se porovnávání děje (báze), nazýváme jej ukazatel bazický. Jednotlivé ukazatele jsou tedy ve tvaru p_0, p_1, p_2, ..., q_0, q_1, q_2, ... ,c_0, c_1, c_2, ...

Budeme používat dva druhy porovnávání těchto ukazatelů a, to pomocí diference a pomocí indexů.

10.1 Absolutní porovnávání

Absolutní porovnávání provádíme pomocí diference (rozdílů). Porovnání můžeme provádět vzhledem k bázi, pak tedy získáme absolutní přírůstek

$$\Delta = q_n - q_0 ,$$

nazýváme jej také bazická diference.

Absolutní přírůstek určuje absolutní množství, které přibylo (v případě záporného čísla ubylo) od původního období.

Také můžeme vytvářet tzv. řetězové absolutní srovnávání. Toto srovnávání spočívá ve vypočtení n-1 různých diferencí, kde n je počet časových momentů. Tyto diference pak spočítáme jako

$$\Delta_i = q_{i-1} - q_i,$$

pro $i \in \{1,\ldots,n\}$. Každá z těchto diferencí určuje přírůstek (úbytek) ukazatele za jedno období. Z řetězových diferencí můžeme bazickou diferenci získat, pokud sečteme všechny řetězové diference

$$\Delta = \sum_{i=1}^{n} \Delta_i$$

10.2 Relativní porovnávání

Častější než absolutní porovnávání je porovnávání relativní. Relativní porovnávání se provádí pomocí indexů. Indexy jsou hodnoty, které získáme podělením ukazatelů.

Indexy nás informují o poměru ukazatelů, a tedy po vynásobení stem o tempu růstu v procentech. Opět rozlišujeme, zda provádíme porovnání vzhledem k bázi, pak dostáváme bazické indexy.

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

Bazické indexy značí růst za celé zvolené období. Druhou možností je porovnání vzhledem k sousedním obdobím, čímž získáme řetězové indexy.

$$I_{\frac{i}{i-1}} = \frac{q_i}{q_i - 1}$$

Řetězové indexy značí tempo růstu (poklesu) za každý interval zvlášť. Používáme pro ně též označení koeficienty růstu (poklesu), proto je také označujeme

$$k_i = I_{i/(i-1)}$$

Z koeficientů růstu také můžeme odvodit koeficient přírůstku (úbytku) ki −1, který po vynásobení stem značí v procentech, jaký byl přírůstek (úbytek) ukazatele za určité období. Často nás zajímá průměrný koeficient růstu. Ten vypočítáme pomocí geometrického průměru

$$k = \sqrt[n]{k_1 \cdot k_2 \cdot \ldots \cdot k_n}$$

Pro výpočet průměrného koeficientu růstu platí vztah:

$$k = \sqrt[n]{\frac{q_n}{q_0}}$$

Z toho vyplývá, že průměrný koeficient růstu závisí pouze na počtu období a pak na první a poslední hodnotě. To je třeba mít na paměti, neboť pokud ukazatel velmi kolísá, pak nemusí mít průměrný koeficient růstu žádný reálný smysl a může se velmi lišit podle toho, které okamžiky vezmeme jako bazický a poslední. V této části jsme se zabývali jen nejjednoduššími indexy, ale celkově je indexů celá řada jak uvidíme dále.

10.3 Indexy

Indexy rozdělujeme na individuální a souhrnné. Pomocí individuálních indexů srovnáváme stejnorodé ukazatele, což jsou takové, jejichž součet má stejný smysl jako stejný ukazatel za jednotlivé části, např. součet zisků za měsíc můžeme sečíst a výsledek má stejný smysl jen za jiné období. Naproti tomu souhrnné indexy jsou indexy nesourodých ukazatelů, tedy takových, jejichž součet nemá pro celek význam, např. součet produkce různých výrobků.

10.4 Jednoduché individuální indexy

Pomocí těchto indexů srovnáváme dvě hodnoty téhož ukazatele. Postupujeme způsobem popsaným výše. Pokud porovnáváme extenzivní ukazatel, pak počítáme jednoduchý individuální index množství

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

Intenzivní ukazatel porovnáváme pomocí jednoduchého individuálního indexu úrovně

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

odpovídající absolutní přírůstek pak bude

$$\Delta_p = p_1 - p_0$$

Vidíme, že u jednoduchých individuálních indexů nehraje roli, zda je ukazatel extenzivní, či intenzivní.

Individuální jednoduché indexy (zde výlučně časové) se často vyskytují sdružené do delších časových řad. V takovém případě mohou být příslušné indexy počítané vždy ke stejnému základu (např. k nejstarší hodnotě v časové řadě původních pozorování), nebo k proměnlivému základu (k bezprostředně předcházejícímu pozorování v časové řadě původních hodnot). V prvním případě, kdy základ srovnání je vždy stejný, hovoříme o tzv. bazických indexech, ve druhém případě, kdy srovnáváme vždy dvě za sebou jdoucí hodnoty v časové řadě, konstruujeme tzv., řetězové indexy. Bazické a řetězové indexy lze vzájemně přepočítávat, tzn., že násobením řetězových indexů získáme indexy bazické a naopak dělením bazických indexů indexy řetězové.

10.5 Složené individuální indexy

Tyto indexy využíváme v případech, kdy máme stejnorodé ukazatele několika částí a chceme je shrnout za celek. Například pokud známe jednotlivé údaje v několika pobočkách a chceme srovnat vývoj za celou společnost.

V případě výpočtů složených individuálních indexů je třeba důkladně hlídat, zda počítáme složené indexy extenzivních, nebo intenzivních ukazatelů.

Extenzivní ukazatele

Pokud porovnáváme extenzivní ukazatele, pak je shrnujeme pomocí součtu. Pro shrnutí využijeme složený individuální index množství

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

Absolutní přírůstek vypočteme opět jako rozdíl hodnot

$$\Delta_q = \sum q_1 - \sum q_0$$

Intenzivní ukazatele

Intenzivní ukazatele p_i nemůžeme shrnout tak lehce jako extenzivní, ale musíme použít vážený aritmetický průměr. Abychom určili váhy jednotlivých hodnot, musíme nejdříve určit hodnoty extenzivního ukazatele q_i. Pro tento extenzivní ukazatel musí existovat jiný extenzivní ukazatel Q_i, tak že platí:

$$p_i = \frac{Q_i}{q_i}$$

Index, který takto získáme, se nazývá index proměnlivého složení, a vypočítáme jej pomocí následujícího vzorce.

$$I^p = \frac{\overline{p_1}}{\overline{p_0}} = \frac{\frac{\sum p_1 \cdot q_1}{\sum q_1}}{\frac{\sum p_0 \cdot q_0}{\sum q_0}}$$

Absolutní přírůstek pak má tvar

$$\Delta_p = \overline{p_1} - \overline{p_0} = \frac{\sum p_1 \cdot q_1}{\sum q_1} - \frac{\sum p_0 \cdot q_0}{\sum q_0}.$$

Index proměnlivého složení tedy zachycuje změny jak intenzivního ukazatele, tak extenzivního ukazatele. Někdy potřebujeme znát pouze změnu jedné z těchto složek, při konstantní hodnotě druhé složky. Abychom potlačili vliv extenzivního ukazatele, musíme výpočet vztáhnout k jednomu období. Což je buď základní q_0, nebo běžné q_1. Poté počítáme index stálého složení, který vyjadřuje vliv změny intenzivní složky při konstantním působení složky extenzivní, značíme jej I^{ss}.

Pro běžné období platí:

$$I^{ss} = \frac{\frac{\sum p_1 \cdot q_1}{\sum q_1}}{\frac{\sum p_0 \cdot q_1}{\sum q_1}} = \frac{\sum p_1 \cdot q_1}{\sum p_0 \cdot q_1}$$

Pro základní období platí:

$$I^{ss} = \frac{\frac{\sum p_1 \cdot q_0}{\sum q_0}}{\frac{\sum p_0 \cdot q_0}{\sum q_0}} = \frac{\sum p_1 \cdot q_0}{\sum p_0 \cdot q_0}$$

Indexy stálého složení nám udávají, jaký by byl růst či úbytek pokud by byl extenzivní ukazatel na stálé hodnotě a to buď té, která byla v běžném období, nebo té v základním období. Podobně můžeme určit index struktury, který charakterizuje vliv extenzivního ukazatele při konstantním působení intenzivního ukazatele, značíme jej I^{str}. Také máme dvě možnosti podle období.

Pro běžné období platí:

$$I^{str} = \frac{\frac{\sum p_1 \cdot q_1}{\sum q_1}}{\frac{\sum p_1 \cdot q_0}{\sum q_0}}$$

Pro základní období platí:

$$I^{str} = \frac{\frac{\sum p_0 \cdot q_1}{\sum q_1}}{\frac{\sum p_0 \cdot q_0}{\sum q_0}}$$

10.6 Souhrnné indexy

Souhrnné indexy jsou indexy nestejnorodých extenzivních ukazatelů. Nestejnorodý ukazatel značí, že hodnoty mohou pro různé části mít různou jednotku. Pokud je tedy chceme srovnávat, není možné hodnoty pouze sečíst. Pro srovnání musíme použít nějaký společný intenzivní ukazatel, vzhledem k němuž můžeme provést souhrn nestejnorodých ukazatelů. Příkladem takového extenzivního ukazatele může být množství prodaných výrobků, v případě, kdy se některé prodávají

v kilogramech, jiné v litrech, kusech, či metrech. Společným intenzivním ukazatelem pak mohou být ceny za jednu jednotku. Souhrnný index pak udává změnu vytvořené hodnoty (např. tržba) a nazýváme jej hodnotový index.

$$I^{cq} = \frac{\sum c_1 \cdot q_1}{\sum c_0 \cdot q_0}$$

Absolutně pak vývoj tržeb vyjádříme jako:

$$\Delta_{cq} = \sum c_1 \cdot q_1 - \sum c_0 \cdot q_0$$

Stejně jako v případě indexu proměnlivého složení, udává hodnotový index změnu ovlivněnou dvěma vlivy. V praxi je ovšem vhodné získat údaje o vlivu jednotlivých ukazatelů. Abychom určili vliv jednoho ukazatele na vývoj vytvořené hodnoty, musíme u druhého ukazatele předpokládat, že je konstantní v čase. V případě cenových indexů předpokládáme, že je konstantní extenzivní ukazatel, naopak v případě objemových indexů předpokládáme, že je konstantní intenzivní ukazatel.

10.6.1 Cenové indexy

Cenové indexy se také nazývají souhrnné indexy úrovně a udávají vliv změny ceny na hodnotový index, za předpokladu že extenzivní ukazatel se nemění. Těchto indexů je celá řada, ale nejčastěji se využívají následující čtyři.

Laspeyresův cenový index využívá jako váhu množství základního období:

$$I_c^L = \frac{\sum c_1 \cdot q_0}{\sum c_0 \cdot q_0}$$

Paascheho cenový index využívá jako váhu množství běžného období:

$$I_c^P = \frac{\sum c_1 \cdot q_1}{\sum c_0 \cdot q_1}$$

Loweho cenový index využívá jako váhu předem zvolené číslo q, toto číslo může být dané, nebo vypočítané, například jako aritmetický průměr extenzivního ukazatele v základním a běžném období:

$$I_c^{Lowe} = \frac{\sum c_1 \cdot q}{\sum c_0 \cdot q}$$

Fisherův cenový index je počítán jako geometrický průměr Laspeyresova a Paascheho indexu:

$$I_c^F = \sqrt{\frac{\sum c_1 \cdot q_0}{\sum c_0 \cdot q_0} \cdot \frac{\sum c_1 \cdot q_1}{\sum c_0 \cdot q_1}}$$

Ve vzorcích cenových indexů je pro označení intenzivního ukazatele použito označení ci, což ale neznamená, že se indexy nemohou použít i v případech, kdy intenzivní ukazatel udává jinou hodnotu než cenu za jednotku.

10.6.2 Objemové indexy

Podobně jako cenové indexy udávají objemové indexy vliv změny množství na hodnotový index, za předpokladu konstantní ceny. Objemové indexy se také nazývají souhrnné indexy množství.

Laspeyresův objemový index využívá jako váhu cenu základního období:

$$I_q^L = \frac{\sum c_0 \cdot q_1}{\sum c_0 \cdot q_0}$$

Paascheho objemový index využívá jako váhu cenu běžného období:

$$I_q^P = \frac{\sum c_1 \cdot q_1}{\sum c_1 \cdot q_0}$$

Loweho objemový index využívá jako váhu předem zvolené číslo c, toto číslo může být dané, nebo vypočítané, například jako aritmetický průměr intenzivního ukazatele v základním a běžném období:

$$I_q^{Lowe} = \frac{\sum c \cdot q_1}{\sum c \cdot q_0}$$

Fisherův objemový index je počítán jako geometrický průměr Laspeyresova a Paascheho indexu:

$$I_q^F = \sqrt{\frac{\sum c_0 \cdot q_1}{\sum c_0 \cdot q_0} \cdot \frac{\sum c_1 \cdot q_1}{\sum c_1 \cdot q_0}}$$

V této kapitole jsme objasnili základy použití indexní analýzy a uvedli způsoby porovnávání získaných ekonomických hodnot (absolutní a relativní). Jako prostředky ke srovnávání hodnot ukazatelů slouží indexy a diference. Rozlišujeme absolutní srovnání (pomocí rozdílů) a relativní srovnání (pomocí podílů - indexů). Podle věcného obsahu ukazatelů dělíme indexy a diference na: objemové → srovnávají 2 hodnoty extenzitního ukazatele a úrovňové → srovnávají 2 hodnoty intenzitního ukazatele. Dále lze indexy a diference dělit na individuální (jednoduché a složené) pro stejnorodé ukazatele a souhrnné (jednoduché a složené) pro různorodé ukazatele.

1. Tabulka uvádí průměrnou dobu výroby výrobků v letech 2016 a 2017. Porovnejte celkovou dobu výroby v jednotlivých letech a určete, jaký vliv na této změně měla změna výrobního postupu (doba výroby) a jaký vliv mělo různé množství vyráběných výrobků.

Výrobek	Množství		Doba výroby	
	2016	2017	2016	2017
A	100	120	5	4
B	150	130	3	2
C	210	150	7	10

2. Jak dělíme indexy? Uveďte příklady.
3. Co je to absolutní porovnávání?
4. Co je to relativní porovnávání?

Literatura k tématu:

[1] HINDLS, R. *Statistika pro ekonomy*. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

[2] KŘÁPEK, Milan. *Statistika I*. [online]. 1. vyd. Znojmo: Soukromá vysoká škola ekonomická Znojmo, 2013 [cit. 2018-01-10]. ISBN 978-80-87314-40-1. Dostupné z: http://www.svse.cz/uploads/File/statistika%20Ix.pdf

[3] MACEK, J. *Ekonomická a sociální statistika*. 1. vyd. Plzeň: Západočeská univerzita v Plzni, 2008. ISBN 978-80-7043-642-4.

[4] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 1. vydání. Ostrava: Vysoká škola Báňská - Technická univerzita Ostrava, 2007 [cit. 2017-12-18]. ISBN 80-248-1194-4. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/

[5] ANDĚL, J. *Statistické metody*. 3. vyd. Praha: Matfyzpress, 2003. ISBN 978-80-867-3208-8.

Kapitola 11

Publikace výsledků statistického zpracování dat

Po prostudování kapitoly budete umět:

- popsat a vysvětlit etapy statistického zpracování dat;
- identifikovat nežádoucí chyby měření;
- využít předložený postup prezentace a interpretace dat v praxi.

Klíčová slova:

Postup, fáze, vyhodnocení, chyba, interpretace, prezentace, software.

11.1 Etapy statistického zpracování dat

Jedná se o pracovní fáze shodné s postupem při experimentech vědeckého zkoumání.

Než začneme zkoumat či měřit nám dosud nedostatečně známý jev či zákonitost, je třeba se v literatuře seznámit s teoreticky či prakticky dosaženými výsledky.

1. Statistické zjišťování

znamená získání dat, tedy souboru naměřených hodnot. Je to nejdůležitější část experimentální práce, neboť chybně získané nebo neúplné údaje již dalšími sebepečlivějšími postupy nelze opravit. Zde se plně uplatní pečlivost záznamu všech podmínek, za kterých byla data získána.

2. Zpracování statistických údajů

představuje třídění naměřených hodnot, jejich sestavení do tabulek a grafů, výpočet statistických charakteristik, korelační a regresní analýza apod.

Tabulky všech naměřených hodnot, výsledné statistické charakteristiky a grafy představují v odborných a vědeckých publikacích části označované jako výsledky.

3. Vyhodnocení zpracovaných údajů a jejich analýza

je konečnou fází statistické práce. V publikacích bývá označena jako diskuse a závěr. Představuje porovnání námi získaných výsledků s teoretickými předpoklady, s výsledky uvedenými v literatuře apod. V případě odchylek pak jejich zdůvodnění, případně návrh postupu dalšího šetření.

V závěru se uvede konečný výsledek - nejčastěji střední hodnota a míra její variability, která bývá vyjádřena intervalem shody pro určitou pravděpodobnost (např. v ekonomii většinou 95 %, přičemž u souborů s „normálním“ rozložením se nejčastěji používá aritmetický průměr a směrodatná nebo relativní směrodatná odchylka jako míra variability našeho měření, směrodatná odchylka průměru pak při porovnání průměrů základního a výběrových souborů. Pokud nemůžeme zamítnout nulovou hypotézu v testu normality, je správnější jako míru polohy uvádět medián anebo modus; jako míru variability, vzhledem k jednoduchosti výpočtu nějaké rozpětí).

Právě vzhledem k pravděpodobnostnímu charakteru zkoumaných jevů není matematicky možno přesně vystihnout všechny kvantitativní i kvalitativní vlastnosti. Tato skutečnost vyžaduje v praxi vždy určitou opatrnost při formulování závěrů a především rozsahu jejich platnosti.

11.2 Chyby měření

Vzhledem k variabilitě daného jevu je třeba vždy provádět co největší rozsah (počet) měření. Můžeme-li změřit všechny existující prvky, mluvíme o tzv. základním souboru. Jeho rozsah (počet měření) se označuje N. To je případ velice vzácný. Běžně pracujeme pouze s určitou částí tohoto souboru, neboť základní soubor bývá příliš rozsáhlý a z praktických důvodů (např. časových) neměřitelný. Proto pracujeme pouze s náhodně vybranou částí jeho prvků označovanou jako náhodný výběrový soubor. Jeho rozsah (počet měření) označujeme n. Při opakovaném měření nedostaneme zcela shodné výsledky vzhledem k nepřesnosti přístrojů, nedokonalosti smyslů experimentátora a obtížnosti dodržet přesně stejné podmínky během měření.

Takto vzniklé chyby dělíme do tří skupin:

1. **Chyby hrubé**

jsou nevhodnou volbou metody (resp. postupu) nebo hrubou nedbalostí experimentátora. Výskyt hrubé chyby vždy negativně ovlivní správnost konečného výsledku. Takové výsledky proto ze souboru vylučujeme a nepoužijeme je ke zpracování statistických údajů.

Abychom mohli odlišit výsledky zatížené hrubou chybou od krajních hodnot, které ještě patří do souboru, je vhodné použít statistických postupů, tzv. testů, pomocí nichž je možno tyto odlehlé (odlišné) výsledky testovat. Při malém počtu měření umožňuje takovéto objektivní posouzení např. Q test (Deanův a Dixonův test), který využívá variační rozpětí

$$R=x_{max}-x_{min}$$

U Q testu seřadíme naměřené hodnoty podle velikosti od nejmenší po největší. Testovací kriterium Q pak vypočteme:

$$Q = \frac{[x_e - x_s]}{R}$$

kde
x_e je zkoumaná extrémní odlehlá hodnota (nejvyšší nebo nejnižší z celého souboru)
x_s je hodnota s x_e sousedící při uspořádání podle velikosti
R je variační rozpětí

Vypočtenou hodnotu Q porovnáme s níže uvedenou tabelovanou tzv. kritickou hodnotou Q_T podle požadované pravděpodobnosti a rozsahu souboru (počtu měření) n. Je-li vypočtená hodnota Q větší než tabelovaná Q_T, je třeba testovanou extrémní hodnotu vyloučit pro další výpočty. Tím se nám logicky v další práci sníží rozsah souboru n o 1. Pokud je Q menší než Q_T, zůstává testovaná extrémní hodnota v souboru a předpokládáme, že se jedná pouze o náhodný vliv. Vyloučení extrémních

hodnot uvedeme v protokolu o měření. V publikaci se spokojíme pouze s konstatováním, že extrémní hodnoty byly testovány Deanovým - Dixonovým testem a případné odlehlé hodnoty s určitou pravděpodobností P (např.95 %) byly vyloučeny.

Tabulka 11-1 Tabulka hodnot Q_T[14]

	Q	Q_t
n	P = 95%	P = 99%
3	0,914	0,988
4	0,765	0,889
5	0,642	0,760
6	0,560	0,698
7	0,507	0,637
8	0,468	0,590
9	0,437	0,555
10	0,412	0,527

2. **Chyby soustavné (systematické)**

mají stále stejný charakter a zkreslují výsledky vždy v určitém směru. Tím způsobují soustavně vyšší nebo nižší výsledky. Jejich příčinou je nejčastěji chybný metodický přístup, špatné nastavení nebo porucha přístroje a stále stejná chyba experimentátora. Pravidelnost těchto chyb umožňuje určit jejich velikost experimentálně nebo výpočtem. Pak můžeme například upravit pracovní postup. Ve výjimečných případech též opravit výsledky pomocí přepočítacího faktoru stanoveného dodatečně.

3. **Chyby náhodné**

způsobují nahodilé vlivy uplatňující se nepravidelně bez jakýchkoliv zákonitostí podle okamžitých podmínek. Na rozdíl od soustavných chyb je nelze ani hodnotit ani systematicky odstranit. Jejich velikost lze pouze s určitou pravděpodobností odhadnout metodami matematické statistiky.

[14] OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika* [online]. 2007 [cit. 2017-12-18]. Dostupné z: https://homen.vsb.cz/~oti73/cdpast1/

11.3 Třídění statistických dat

Třídění navazuje na etapu zjišťování dat. Účelem je rozčlenit a uspořádat data podle jednoho nebo více znaků (hledisek, vlastností, kritérií apod.). Při variačním třídění, což je třídění do skupin (intervalů, tříd) podle variant (obměn) třídícího znaku, který má kvantitativní charakter je třeba dodržet následující:

- skupiny musí být vytvořeny tak, aby bylo možné do nich začlenit všechny hodnoty souboru a pro každou hodnotu musí platit zcela jednoznačně, do které skupiny bude zařazena,
- má-li třídící znak nespojitý (diskrétní) charakter s malým počtem obměn, pak třídíme do skupin podle všech těchto obměn,
- je-li třídící znak spojitý nebo nespojitý s velkým počtem obměn, nemůžeme vzhledem k přehlednosti třídit podle všech těchto obměn. V těchto případech je třeba sloučit blízké obměny třídícího znaku do společných intervalů, čímž získáme skupinové (intervalové) rozdělení četností
- počet intervalů nejčastěji označovaný k by se měl pohybovat mezi 6 až 20,
- pokud je to možné tvoříme intervaly o stejné šířce (rozdíl mezi krajními hodnotami intervalu je vždy stejný), která má být volena tak, aby se zachovaly informace o původních datech. Proto by neměly být příliš široké,
- všechny intervaly by měly mít dostatečný počet jednotek, což však není vždy podmínkou, mohou být i prázdné intervaly,
- je vhodné volit středy intervalů jako celá (zaokrouhlená) čísla.

Vlastní technika třídění je taková, že po vymezení intervalů uděláme pro každou jednotku ze souboru čárku v příslušném intervalu pracovní tabulky (čárkovací metoda). Počet čárek v jednotlivých intervalech, tedy počet jednotek nazýváme intervalová četnost. Součet všech intervalových četností se musí rovnat rozsahu souboru, což je nejběžnější kontrola toho, že jsme roztřídili celý soubor. Četnosti pak můžeme dělit na absolutní, výše uvedené a relativní, které nám udávají procento z rozsahu celého souboru, tedy součet relativních četností ve všech intervalech musí být roven 100.

11.3.1 Zpracování údajů statistickými postupy

Spočívá ve zpracování utříděných dat, které můžeme nazvat statistická analýza. Toto zpracování dělíme na tyto části:

1. **primární zpracování dat (třídění 1. stupně)**
   - zpracování skupin dat, zjišťujeme například *absolutní a relativní četnosti, průměr*, *medián*, *směrodatné odchylky* u jednotlivých proměnných

2. **sekundární zpracování dat (třídění 2. stupně)**
   - zjištění vazeb mezi jednotlivými proměnnými, příp. jejich skupinami
   - => výpočty *korelací, regresí*, použití různých variant neparametrických výpočtů, faktorovou analýzu, trsovou analýzu atd.,
   - Testování rozdílů mezi proměnnými, skupinami apod. (Studentův t-test nebo testem chí-kvadrát) a následné zjištění, zda výsledky jsou nebo nejsou statisticky významné.

11.3.2 Statistické programy

S využitím kvantitativních metod – tedy metod zpracovávajících informace v číselné i nečíselné podobě - souvisí využití informačních technologií zahrnujících jak sběr dat, tak také jejich zpracování pomocí počítačů a příslušného software. Nejčastěji se přitom využívá statistického software a to zejména tabulkových kalkulátorů, z nichž daleko nejpopulárnější je Excel od firmy Microsoft. Ten je dnes při nákupu standardně dodáván s osobním počítačem PC spolu s operačním systémem Windows. Významnou součástí funkcí Excelu tvoří jeho statistické funkce, kde základní funkce lze nalézt přímo v seznamu statistických funkcí, standardně je však dodáván i dodatek Excelu – Analýza dat, která soubor standardních funkcí významně rozšiřuje. Pro profesionální zpracování dat, však funkce Excelu často nestačí, je zapotřebí sofistikovanějších metod. Pro oblast ekonomických věd, ale i např. sociálních věd, je velmi vhodný program SPSS (Statistical Package for Social Sciences) dodávaný firmou IBM. Zde je stručný přehled dostupných software:

- Excel (součást Microsoft Office),
- Specializovaný statistický software – SPSS, Statistica, Stata, Statgraphic, Origin aj.
  - výsledky lze snadno získat pomocí předdefinovaných operací
  - umožňují grafické znázornění výstupů,
  - po zacvičení je práce s nimi velmi jednoduchá a rychlá,
  - umožňují zkoušet různé možnosti výpočtů a vytěžit z údajů maximum.

11.4 Prezentace a interpretace dat

- *Prezentací dat* se rozumí popis třídění dat a jejich dalších analýz, komentáře tabulek, grafů a provedených operací ad.
- *Interpretace dat* je výklad zjištěných výsledků (vysvětlení, co znamenají, co z nich vyplývá, jaké závěry a jaká opatření atd.).[1]

Tyto dvě roviny práce s daty se ale v kvalitativních analýzách neodlučitelně prolínají. U kvantitativních zkoumání je pak užitečné nejprve prezentovat výsledky a teprve potom vyslovit závěry, názory, přesvědčení atd. Je potřeba dávat si pozor na to, že *vše zjištěné je jen jakási tendence, jistý náznak, možný trend* a na žádné údaje (podpořené kvantitativním či kvalitativním přístupem) nelze přísahat, tvrdit, že to tak jednoznačně je. To znamená, že si musíme dát **pozor na chyby při interpretaci** - např. statistické testování vztahů vychází z falzifikace a ze statistické indukce. K vyjádření tohoto principu můžeme využít dvou pojmů:

- **stochastická zákonitost** - vyjadřuje, že pravděpodobnost stojí v podstatě tohoto zákona, realita je pravděpodobná
- **statistická souvislost** – prokázána jen s jistou spolehlivostí [1]

Postup prezentace dat v praxi

- uspořádání a shrnutí dat, jejich transformace do grafů a tabulek
- přehledná, úsporná forma prezentování údajů,
- je třeba zdůraznit důležitá zjištění
    - ta, která např. podporují očekávané trendy nebo naopak údaje, které nebyly očekávány
- údaje lze různě přeskupovat a kombinovat,
    - lze vyrobit velké množství tabulek a grafů => vybrat jen rozumné množství,
        - ve zprávě z výzkumu uvést jen podstatné výsledky vzhledem k cíli výzkumu
        - příliš velké množství tabulek ukazuje, že se výzkumník v datech ztratil či neumí najít správnou hierarchii, a proto uvedl vše, co měl k dispozici
        - výzkumy z větším množstvím proměnných obyčejně vyžadují větší počet tabulek než jednodušší výzkumy

Pořadí tabulek a grafů

1. nejprve ty, které obsahují hlavní a souhrnné informace
    - čtenář získá globální přehled o výsledcích, pak se hlavní výsledky přeměňují na drobné

2. tematické řazení
   - dle výzkumného problému a hypotéz,
   - má-li výzkum 4 hypotézy, výsledky budou seřazeny do 4 okruhů

Kritéria dobré prezentace

- přehlednost grafů a tabulek
- srovnávání vhodných skupin v komentáři ke grafům
- komentář není převod čísel do slov, je třeba uplatnit nadhled
- vyjádřit se ke svým hypotézám (očekávám, předpokládám…)
- tematicky řadit údaje, tabulky a grafy
- rozlišit jasně samotné údaje a svou interpretaci údajů – jde o vhodné formulace
  - srovnat své závěry s údaji z předcházejících výzkumů.[15]

V této kapitole jsme objasnili základy zpracování, interpretace a prezentace statistických dat. Závěr zpracování statistických dat spočívá nejčastěji ve vypracování závěrečné zprávy – prezentace získaných poznatků, hodnocení výsledků a jejich vysvětlení. Rovněž je důležitý i podrobný popis použité metodiky. Publikované informace mají být věcné, srozumitelné a logicky strukturované. Mezi základní části zpracování dat patří: vymezení problému (cíle, předmět, hypotézy, případně výsledky předběžné analýzy), metodika (techniky, vzorek, statistické testy, procedury, výsledky předvýzkumu, sběru dat, posouzení reprezentativnosti), vlastní výsledky (interpretace, závěry s doporučením, shrnutí poznatků, vysvětlení souvislostí, doporučení, posouzení cílů výzkumů, posouzení metodiky). Důležité je rovněž věcně a metodologicky výzkum uzavřít a zajistit možnost opakování výzkumu za srovnatelných podmínek.

1. Popište etapy statistického zpracování dat.
2. Jaké existují chyby měření?
3. Popište proces třídění statistických dat.
4. Dle popsaného procesu prezentace a interpretace dat představte svoji zpracovanou případovou studii.

[15] REICHEL, J. *Kapitoly metodologie sociálních výzkumů*. 2009.

Literatura k tématu:

[1] REICHEL, J. *Kapitoly metodologie sociálních výzkumů*. Praha: Grada, 2009. Sociologie (Grada). ISBN 978-80-247-3006-6

[2] HENDL, J. *Kvalitativní výzkum: základní teorie, metody a aplikace*. Čtvrté, přepracované a rozšířené vydání. Praha: Portál, 2016. ISBN 978-80-262-0982-9.

[3] HINDLS, R. *Statistika pro ekonomy*. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

[4] KŘÁPEK, Milan. *Statistika I*. [online]. 1. vyd. Znojmo: Soukromá vysoká škola ekonomická Znojmo, 2013 [cit. 2018-01-10]. ISBN 978-80-87314-40-1. Dostupné z: http://www.svse.cz/uploads/File/statistika%20Ix.pdf

[5] ANDĚL, J. *Statistické metody*. 3. vyd. Praha: Matfyzpress, 2003. ISBN 978-80-867-3208-8.