Štatistické metódy

doc. RNDr. Štefan Peško, CSc.

Obsah

Úvod **2**

1 Štatistické metódy **3**
1.1 Základné pojmy 3
1.2 Metódy analýzy kvalitatívnych údajov 4
1.3 Modelovanie štatistických závislostí 6
1.4 Výberové metódy 8
1.5 Princíp testovania hypotéz 9
1.6 Analýza časového radu 11
1.7 Štatistické metódy riadenia kvality 13

Register **15**

Literatúra **15**

Úvod

V nasledujúcich kapitolách sa budeme zaoberať niektorými základnými stochastickými metódami, s ktorými sa v praxi stretávame pri manažérskom rozhodovaní [13]. Najskôr stručne zopakujeme niektoré pojmy a postupy z teórie pravdepodobností, ktoré sú potrebné k pochopeniu problematiky. Pre hlbšie vniknutie do tejto problematiky odporúčame učebnicu [2] a [12] a pre precvičenie príručku [8]. Naviažeme najpoužívanejšími štatistickými metódami s ukážkami ich použitia. Rozsiahlejší výklad problematiky možno nájsť v učebniciach [1, 3]. V texte sa obmedzíme len na výber takých metód, ktoré majú podporu v progamovom balíku `MS Excel`. Ako významné aplikácie stochastických metód sme vybrali modeli z teórie hromadnej obsluhy a teórie zásob z monografií [13, 6, 9]. Využitie základných poznatkov teórie markovových reťazcov demonštrujeme na riešení praktickej optimalizačnej úlohy údržby a obnovy zariadenia s podporou solvéru *Riešiteľ* v programe `MS Excel`.

Pravdepodobnostné predpovede, úsudky a modely sa už dnes stávajú, aj vďaka dostupným programovým produktom (`MS Excel, STATISTIC, R`), bežnou výbavou manažérov na všetkých stupňoch riadenia. Uplatňujú sa napr. pri vyhodnocovaní výsledkov marketingového prieskumu, kontrole akosti výrobkov, tvorbe nevyhnuných zásob tovarov, voľbe typov a režimoch obsluhy zákazníkov. Teoretickým základom takéhoto rozhodovane v podmienkach neistoty je teória pravdepodobností.

Kapitola 1

Štatistické metódy

Štatistika je odbor, ktorý má široké uplatnenie v manažérskej praxi nakoľko sa zaoberá zberom, analýzou a interpretáciou údajov ktoré boli získané pozorovaním alebo z experimentov. Umožňuje formulovať objektívne závery na základe skúmaných údajov. Súčasné využitie štatistických metód je nemysliteľné bez použitia zodpovedajúceho softvéru. Pre potreby zoznámenia sa so základnými štatistickými postupmi vystačíme aj s nástrojmi, ktoré poskytuje program `MS Excel`. Mimoriadne vydareným učebným textom s jeho podporou je učebnica *Štatistika s Excelom* [4].

V manažérskej praxi spravidla nemáme dostatok ani času ani financií aby sme sa mohli rozhodnúť na základe preskúmania všetkých údajov, ktoré sa viažu na analyzovaný problém. Napr. zistenie aký bude záujem o inovovaný výrobok je nerealizovateľný oslovením všetkých potenciálnych zákazníkov. Prieskum sa ale môže oprieť o relatívne malú časť (**vzorku**) zákazníkov (**základného súboru**). Štatistika tak používa postupy, pomocou ktorých môže s istým rizikom usudzovať o chovaní základného súboru.

Štatistický prieskum sa spravidla riadi nasledujúcimi krokmi:

Formulácia problému – je potrebné špecifikovať v čom spočíva problém.

Zber údajov – rozhodnutie o spôsobe a rozsahu zberu dát.

Analýza údajov – voľba štatistickej metódy umožňujúcej usudzovať z vlastnosti vzorky dát na vlastnosť základného súboru (**štatistická indukcia**).

Vyhodnotenie – odpovedá na otázky kladené vo formulácii problému.

1.1 Základné pojmy

Pozorované údaje, napr. maloobchodná cena výrobku, vykazujú isté náhodné kolísanie a tak je vhodné sa na zistené údaje dívať z pravdepodobnostného hľadiska ako na výsledky náhodného pokusu.

V prípade, že opakujeme n krát náhodný pokus, ktorého výsledkom je hodnota náhodnej veličiny X s distribučnou funkciou $F(x, \theta)$, kde θ je reálny parameter alebo vektor parametrov uvažovaného rozdelenia pravdepodobností, potom pozorujeme náhodný vektor $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$, ktorého zložkami sú nezávislé náhodné veličiny $\mathbf{X}_i$ s rovnakým rozdelením pravdepodobnosti. Náhodný vektor $\mathbf{X}$ sa nazýva **náhodný výber** a n je **rozsah** náhodného **výberu** .

Číselný vektor $(x_i, x_2, \ldots, x_n)$, ktorý získame realizáciou náhodného výberu $\mathbf{X}$ nazývame **štatistický súbor** a jeho prvky x_i **štatistické jednotky**. Súbor všetkých možných štatistických jednotiek – hodnôt náhodnej veličiny X, nazývame **základný súbor (populácia)**.

Na štatistických jednotkách súboru tak sledujeme nejakú vlastnosť štatistických jednotiek (životnosť výrobku, vek respondentov, dobu obsluhy zákazníka atď.), ktorú nazývame **štatistický znak**.

Pre štatistické súbory je typický vysoký rozsah, ktorý treba zohľadniť pri výbere spôsobu štatistického skúmania. Rozlišujeme tieto základné spôsoby štatistického skúmania:

Vyčerpávajúce skúmanie (consuz) všetkých štatistických jednotiek súboru. Napr. sčítanie ľudu, domov a bytov pre potreby sledovania demografických javov. Je mimoriadne nákladná no jej výhodou je presnosť zistených charakteristík s podrobnou informáciou o každom jedincovi populácie.

Výberové skúmanie. Zo základného súboru o rozsahu N vyberieme jeho časť tzv. **výberový súbor** rozsahu n a po jeho spracovaní na základe výsledkov usudzujeme na vlastnosti celej populácie. Napr. pri zisťovaní verejnej mienky, stratifikovaný výber bytov v SR; $n = 10250$ bytov je cca 6 % z celkového počtu obývaných bytov. Mimoriadne dôležitá je tu netriviálna požiadavka reprezentatívneho výberu.

Náhodná veličina X ktorej hodnoty pri jej realizácii pozorujeme môžeme popísať pomocou rôznych charakteristík, hovoríme o **parametroch základného súboru** (stredná hodnota μ, rozptyl σ^2, smerodajná odchýlka σ). Vo výberových súboroch sú analógie týchto parametrov **výberové charakteristiky** ktoré nazývame **štatistiky** a definujeme ich ako funkcie náhodného výberu $T(\mathbf{X})$. Jedná sa teda o náhodné veličiny, ktoré vieme popísať nejakým rozdelením, majú svoju strednú hodnotu, rozptyl a ďalšie charakteristiky.

Ako príklad základných štatistík uveďme **výberový priemer** a **výberový rozptyl** náhodného výberu $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$, ktoré sú definované vzťahmi

$$\overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i, \quad \mathbf{S}^2 = \frac{1}{n-1}(\mathbf{X}_i - \overline{\mathbf{X}})^2.$$

Ak máme náhodný výber $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ z toho istého rozdelenia $F(x)$ so strednou hodnotou μ_X a rozptylom σ_X^2 a naviac sú všetky dvojice náhodných veličín $\mathbf{X}_i\mathbf{X}_j$ nezávislé, potom $E(\overline{\mathbf{X}}) = \mu_X$, $D(\overline{\mathbf{X}}) = \frac{\sigma_X^2}{n}$.

Ďalej sa obmedzíme len na niektoré pomerne jednoduché štatistické metódy z [3]. Záujemcom o hlbšie poznatky z matematickej štatistky možno odporúčať z dostupnej literatúry učebnice [1],[12].

1.2 Metódy analýzy kvalitatívnych údajov

Predpokladajme, že sme v procese zisťovania získali hodnoty nominálnych (slovných, kvalitatívnych), ordinárnych (poradových znakov) alebo máme hod-

noty kategorizovaných kvalitatívnych (číselných) znakov. Úlohou je analyzovať získané údaje.

V prípade analýzy hodnoty len jedného znaku je výsledkom tabuľka s početnosťou výskytu jednotlivých hodnôt **jednostupňové triedenie**. Nech je analyzovaným kvalitatívnym znakom je znak A, ktorý môže nadobúdať m rôznych hodnôt $a_1, a_2, \ldots, a_m$. Výsledkom triedenia je zistenie absolútnych a relatívnych početností výskytov jednotlivých hodnôt v štatistickom súbore. Označme absolútne početnosti n_i a relatívne početnosti r_i potom

$$n = \sum_{i=1}^{n} n_i, \quad r_i = \frac{n_i}{n}.$$

Výsledná schéma tabuľky s výsledkami jednostupňového triedenia v tab. 1.1.

Hodnota znaku	Absolútna početnosť	Relatívna početnosť
a_1	n_1	r_1
a_2	n_2	r_2
$\ldots$	$\ldots$	$\ldots$
a_m	n_m	r_m
Spolu	n	1

Tabuľka 1.1: Schéma tabuľky s výsledkami jednostupňového triedenia

V prípade analýzy hodnôt dvoch znakov A, B je výsledkom **kontingenčná tabuľka** s početnosťou výskytu dvojíc hodnôt týchto dvoch znakov tzv. **dvojstupňové triedenie**. Kontingenčná tabuľka má nasledujúci tvar v tab. 1.2.

$A \backslash B$	b_1	b_2	$\ldots$	b_s	Spolu
a_1	n_{11}	n_{12}	$\ldots$	n_{1s}	n_{10}
a_2	n_{12}	n_{22}	$\ldots$	n_{2s}	n_{20}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
a_m	n_{1m}	n_{m2}	$\ldots$	n_{ms}	n_{m0}
Spolu	n_{01}	n_{02}	$\ldots$	n_{0s}	n_{00}

Tabuľka 1.2: Schematický tvar kontingenčnej tabuľky

V riadkoch tabuľky sú absolútne početnosti výskytu pre jednotlivé hodnoty $a_1, a_2, \ldots, a_m$ znaku A. V stĺpcoch tabuľky sú zase absolútne početnosti výskytu pre jednotlivé hodnoty $b_1, b_2, \ldots, b_s$ znaku B. Hodnota n_{ij} v i-tom riadku a j-tom stĺpci tabuľky udáva počet štatistických jednotiek pri ktorých premenná A nadobúda hodnotu a_i a súčasne premenná B nadobúda hodnotu b_j. Početnosti v riadku `Spolu` predstavujú výsledky jednostupňového triedenia podľa znaku A a početnosti v stĺpci `Spolu` predstavujú výsledky jednostupňového triedenia podľa znaku B. Konečne políčko na priesečníku riadku `Spolu` a stĺpca `Spolu` obsahuje hodnotu rozsahu súboru n_{00}.

V zložitejších verziách kontingenčných tabuliek sa v je políčkach tabuliek môžu vyskytovať aj viaceré hodnoty — okrem absolútnych početností n_{ij} to bývajú relatívne početnosti $r_{ij} = \frac{n_{ij}}{n_{00}}$, tak to uvádza väčšina vyspelých softvérových nástrojov napr. `R` s možnosťou voľby ďalších štatistík.

Pri výsledkoch dvojstupňového triedenia nás často zaujímajú aj závislosti medzi konkrétnymi dvojicami hodnôt skúmaných znakov. K meraniu **stupňa závislosti medzi hodnotami dvoch kvalitatívnych znakov** sa využíva štatistika χ^2 takto:

Najskôr sa vypočíta očakávaná početnosť

$$e_{ij} = \frac{n_{i0} \cdot n_{0j}}{n_{00}},$$

kde n_{i0} je početnosť výskytu i-tej hodnoty znaku A, n_{0j} je početnosť výskytu j-tej hodnoty znaku B, n_{00} je početnosť súboru.

Potom individuálne a celkové χ^2 hodnoty sú definované vzťahmi

$$\chi^2_{ij} = \frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \quad \chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{s} \chi^2_{ij}.$$

Štatistika slúži na testovanie hypotézy o nezávislosti kvantitatívnych znakov A a B. **Kontingenčný koeficient** je rovný

$$\sqrt{\frac{\chi^2}{\chi^2 + n_{00}}},$$

ktorý nadobúda hodnoty z intervalu $\langle 0, 1)$, pričom hodnoty blízke 0 svedčia o žiadnej alebo slabej závislosti a čím sú hodnoty väčšie a bližšie 1 tým rastie aj stupeň závislostí.

1.3 Modelovanie štatistických závislostí

Hľadajme odpoveď na otázky manažéra:

„Aká silná je závislosť medzi veľkosťou predaja výrobku Y a výdavkami na reklamu X? Aký nárast predaja možno očakávať, ak zvýšime výdavky na reklamu o 50 tisíc €?“

Pri regresnom modelovaní závislosti hodnôt premennej Y od hodnôt premennej X môžeme vyjadriť vzťahom $y_i = f(x_i) + e_i$, kde $(x_i, y_i), i = 1, 2, \ldots, n$ je n bodov, ktorých súradnice sú vyjadrené hodnotami premennej X a Y a e_i sú náhodné chyby od ktorých sa zvyčajne požaduje, aby mali normálne rozdelenie $N(0, \sigma^2)$ a boli navzájom nezávislé.

Na určenie konkrétneho tvaru modelu potrebujeme:

- Bodový graf pozorovaných hodnôt (x_i, y_i), $i = 1, 2, \ldots, n$.
- Znalosť grafického priebehu zvolenej funkcie $f(x)$ (priamka, parabola, ...).

V prípade voľby priamky $y_i = b_0 + b_1 x_i + e_i$ má **regresná priamka** tvar

$$\hat{y}_i = b_0 + b_1 x_i,$$

kde $\hat{y}_i$ je odhadom hodnoty y_i, $i = 1, 2, \ldots, n$. Náhodná chyba (reziduum) má potom tvar $e_i = y_i - \hat{y}_i$. Parametre b_0 a b_1 sa odhadujú tak, že sa minimalizuje súčet štvorcov reziduí $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ čo vedie na riešenie systému lineárnych rovníc o dvoch neznámych b_0 a b_1:

$$\sum_{i=1}^{n} y_i = b_0 n + b_1 \sum_{i-1}^{n} x_i, \quad \sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2.$$

Na ich výpočet sa používajú funkcie regresie v príslušných softvéroch. Napríklad v programe `MS Excel` je to nástroj *Regression*.

Regresný koeficient b_1 sa interpretuje v závislosti od typu výskumu. V prípade experimentu (v ktorom sa premennou X manipuluje), vyjadruje o koľko sa zvýši očakávaná hodnota premennej Y, ak sa hodnota premennej X zvýši o 1 jednotku. V prípade pozorovacej štúdie sa koeficient interpretuje ako očakávaný rozdiel hodnôt premennej Y dvoch pozorovaní, ktorých hodnota premennej X sa líši o jednu jednotku.

Za predpokladu, že údaje predstavujú náhodnú vzorku z populácie, sú vypočítané regresné koeficienty najlepšími bodovými odhadmi neznámych parametrov. Okrem toho možno testovať hypotézy (nulová hypotéza, že koeficient sa rovná nule vyjadruje, že medzi premennými v základnom súbore neexistuje vzťah) a zostrojiť ich intervalové odhady. Testy hypotéz a intervalové odhady regresných koeficientov predpokladajú, že chyby e_i sú navzájom nezávislé.

Korelačný koeficient meria silu štatistickej závislosti medzi dvoma kvantitatívnymi premennými. Korelačná analýza, na rozdiel od regresie, nevyjadruje príčinne-následný vzťah $Y = f(X)$. Premenná Y nezávisí od premennej X, ale dve náhodné premenné X a Y sa spoločne menia. Regresná analýza predpokladá, že premenná Y je náhodná a premenná X fixná. Pod pojmom korelačný koeficient sa najčastejšie myslí **Pearsonov korelačný koeficient (Pearson's product moment)** z roku 1896, ktorý je mierou lineárnej závislosti dvoch premenných. Pearsonov korelačný koeficient ρ odhadnutý z náhodnej vzorky sa zapisuje r a vypočíta sa:

$$r = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{S_X S_Y}, \tag{1.1}$$

kde S_X^2 a S_Y^2 sú výberové rozptyly premenných X a Y.

Čitateľ vo vzťahu (1.1) sa nazýva **kovariancia** a vyjadruje ako sa súčasne menia hodnoty dvoch premenných. Kladná hodnota znamená, že sa menia spoločne jedným smerom, záporná hodnota znamená že sa menia opačným smerom a nula, že sa menia nezávisle. Vydelením kovariancie štandardnými odchýlkami sa vypočíta korelačný koeficient, ktorého hodnota sa nachádza v intervale od -1 do 1. Pearsonov korelačný koeficient sa rovná -1 v prípade, že všetky pozorovania ležia na klesajúcej priamke a 1 ak pozorovania ležia na stúpajúcej priamke.

Interpretácia korelačného koeficientu závisí od kontextu. Hodnota 0,8 pri overení fyzikálneho zákona použitím presných meracích prístrojov je veľmi nízka, v sociálnych vedách je však veľmi vysoká. Cohen (1988) vytvoril jednoduchú pomôcku pre interpretáciu korelačných koeficientov v psychologickom výskume: Korelácia (v absolútnej hodnote) pod 0,1 je triviálna, 0,1–0,3 malá, 0,3–0,5 stredná a nad 0,5 veľká.

Hodnota r^2 (R-squared) sa nazýva **koeficient determinácie** a vyjadruje podiel spoločnej variability medzi dvoma premennými. Test významnosti Pearsonovho korelačného koeficientu a intervalový odhad vyžadujú nezávislé pozorovania.

Pearsonov korelačný koeficient je silne ovplyvniteľný extrémnymi hodnotami (outliers) a to v oboch smeroch. Jediný extrém vo veľkom súbore môže významne znížiť silnú závislosť, ale aj vyrobiť silnú závislosť tam, kde žiadna nie je. Touto citlivosťou na extrémne hodnoty netrpia **poradové korelačné koeficienty**. Dôležité závery sa nesmú robiť iba na základe hodnoty koeficientu. Vždy je nutné preskúmať X–Y graf. Z grafu možno zistiť aj nelineárny ale silný vzťah medzi premennými. V takom prípade treba vzťah linearizovať transformáciou premenných (napr. logaritmovaním Y), ktoré sa následne použijú na výpočet korelácie.

1.4 Výberové metódy

V matematickej štatistike je dôležité členenie na základný súbor a výberový súbor. V manažérskej praxi sú počítané štatistiky skoro nezávislé na tom, či ide o základný alebo výberový súbor nakoľko sa jedná spravidla o neúplné súbory. Štatistiky možno vypočíta na základe údajov ľubovoľného súboru. Ich interpretácia je korektnejšia ak sme neúplný súbor získali riadeným výberovým zisťovaním so známou príslušnosťou jednotiek do výberovej vzorky.

Predpokladajme, že máme n štatistických jednotiek a zistili sme hodnoty $x_1, x_2, \ldots, x_n$ skúmaného znaku. Z metodického hľadiska sa používajú dva typy odhadov populácie:

- **Bodový odhad**, kde parameter základného súboru (stred rozdelenia μ smerodajnú odchýlku σ, podiel π výskytu nejakej vlastnosti v súbore) aproximujeme jedným číslom.
- **Intervalový odhad**, kde tento parameter aproximujeme intervalom v ktorom s veľkou pravdepodobnosťou príslušný populačný parameter leží.

Majme náhodný výber $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ z nejakého rozdelenia, ktoré závisí od neznámeho parametra Θ. Bodovým odhadom T parametra Θ je výberová štatistika $T(\mathbf{X})$ ktorá nadobúda hodnoty blízke parametru Θ. Dobrý odhad má tri základné vlastnosti, je:

- **Neskreslený**, ak $E(T(\mathbf{X})) = \Theta$.
- **Výdatný**, ak má zo všetkých neskreslených odhadov najmenší rozptyl.

- **Konzistentný**, ak s rastúcim rozsahom výberu n sa zvyšuje presnosť odhadu Θ.

Príklad 1.1. *Majme náhodný výber* $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ *z normálneho rozdelenia* $N(\mu, \sigma^2)$. *Ako odhad rozptylu* σ^2 *sa často používa výberový rozptyl* $\mathbf{S}^2$. *Možno dokázať, že tento odhad je nestranný* $E(\mathbf{S}^2) = \sigma^2$ *a konzistentný*

$$\lim_{n\to\infty} E(\mathbf{S}^2) = \sigma^2 \ a \ \lim_{n\to\infty} D(\mathbf{S}^2) = 0. \qquad \square$$

V praktických úlohách často určujeme odhad príslušného intervalu pomocou intervalu. **Interval spoľahlivosti** (konfindenčný interval) pre parameter Θ so spoľahlivosťou $1-\alpha$, kde $\alpha \in \langle 0, 1\rangle$ je taká dvojica štatistík $(\mathbf{T}_D, \mathbf{T}_H)$, že

$$\mathcal{P}(\mathbf{T}_D \leq \Theta \leq \mathbf{T}_D) = 1 - \alpha.$$

Intervalový odhad parametra Θ so spoľahlivosťou $1-\alpha$, $\alpha \in \langle 0, 1\rangle$, je interval $\langle t_D, t_H\rangle$, pričom t_D, t_H sú hodnoty štatistík $\mathbf{T}_D$, $\mathbf{T}_H$ daného štatistického súboru $x_1, x_2, \ldots, x_n$. Parameter α nazývame **hladina významnosti** odhadu.

Podľa konštrukcie intervalov spoľahlivosti rozlišujeme interval spoľahlivosti:

- **Ľavostranný** – daná je dolná hranica $\mathbf{T}_D$, platí $\mathcal{P}(\Theta \geq \mathbf{T}_D) = 1 - \alpha$.
- **Pravostranný** – daná je horná hranica $\mathbf{T}_H$, platí $\mathcal{P}(\Theta \leq \mathbf{T}_H) = 1 - \alpha$.
- **Obojstranný** – dané sú obe hranice $\mathbf{T}_D, \mathbf{T}_H$, platí $\mathcal{P}(\Theta < \mathbf{T}_D) = \mathcal{P}(\Theta > \mathbf{T}_H) = \frac{\alpha}{2}$.

V aplikáciach býva často požiadavka na spoľahlivosť odhadu vopred daná. Ak chceme intervalový odhad spresniť, potom najlepšie urobíme, ak zväčšíme rozsah výberu n, pretože šírka intervalového odhadu sa zmenšuje úmerne $\sqrt{n}$. V praxi hľadáme kompromis medzi spoľahlivosťou a významnosťou odhadu.

Problematikou konštrukcie bodových odhadov ani intervalov odhadov sa nebudeme ďalej zaoberať, presahuje náplň tejto publikácie.

1.5 Princíp testovania hypotéz

Štatistická hypotéza je tvrdenie o rozdelení pozorovanej náhodnej veličiny. Ak pojednáva o parametroch rozdelenia náhodnej veličiny (stredná hodnota, medián, rozptyl, atď), hovoríme o **parametrickej hypotéze** ak o jej vlastnostiach (typ rozdelenia, nezávislosť výberu, atď) hovoríme o **neparametrickej hypotéze**.

Testovanie hypotéz je rozhodovací proces v ktorom proti sebe stoja nulová a alternatívna hypotéza. **Nulová hypotéza** H_0 (testovaná hypotéza) predstavuje tvrdenie, že sledovaný efekt je nulový, môže byť vyjadrená rovnosťou medzi testovaným parametrom θ a jeho očakávanou hodnotou θ_0

$$H_0 \colon \theta = \theta_0.$$

Alternatívna hypotéza H_A nejakým spôsobom popiera tvrdenie nulovej hypotézy H_0. K uvedenej nulovej hypotéze prichádza do úvahy jedna zo štyroch možností:

- $H_A: \theta = \theta_1$. Táto jednoduchá alternatívna hypotéza sa používa v prípade, keď sa rozhodujeme medzi dvoma hodnotami θ_0 a θ_1.
- $H_A: \theta \neq \theta_0$. Táto zložená alternatívna hypotéza popiera platnosť nulovej hypotézy nez bližšej špecifikácie, tvrdí len že hodnota parametra je iná než θ_0.
- $H_A: \theta < \theta_0$. Táto jednostranná alternatívna hypotéza popiera platnosť nulovej hypotézy tvrdiac, že hodnota parametra je menšia než θ_0.
- $H_A: \theta > \theta_0$. Táto jednostranná alternatívna hypotéza popiera platnosť nulovej hypotézy tvrdiac, že hodnota parametra je väčšia než θ_0.

Alternatívna hypotéza by mala byť v súlade s pozorovaním, ktoré sme získali z výberového súboru.

Príklad 1.2. *Overte, či sú priemerné platy absolventov manažmentu a informatiky fakulty v ich prvom zamestnaní zhodné.*

- ***Populácia 1:*** *Absolventi manažmentu, ktorí odpovedali na anketu.*
- ***Populácia 2:*** *Absolventi informatiky. ktorí odpovedali na anketu,*
- ***Sledovaný štatistický znak (náhodná veličina):*** *Mzda.*
- ***Nulová hypotéza:*** $\mu_M = \mu_I$*, kde* μ_M *a* μ_I *označuje priemernú mzdu absolventov manažmentu a informatiky.*
- ***Alternatívna hypotéza:*** $\mu_M \neq \mu_I$*, zadanie nepožaduje jednostrannú nerovnosť.* □

Testom štatistickej hypotézy rozumieme rozhodovací proces, pri ktorom na základe výberového súboru rozhodneme, ku ktorej z hypotéz sa prikláňame. Musíme ich formulovať tak aby platila práve jedna z hypotéz. Môžeme dospieť k dvom rozhodnutiam:

- Zamietame nulovú hypotézu H_0 v prospech alternatívnej hypotézy H_A.
- Nezamietame nulovú hypotézu H_0.

Jedným, žiaľ častým nepochopením princípu testovania hypotéz je domnienka, že rozhodnutím môže byť aj „Prijatie hypotézy H_0“. To však nie je možné pretože nikdy nevieme, či by údaje z iného výberového súboru neumožnili zamietnuť hypotézu H_0.

Obor hodnôt testovaného parametra sa delí na dve disjunktné množiny – **obor prijatia** H_0 a **kritický obor** W zamietnutia hypotézy H_0. Kritický obor W vieme popísať pomocou kritického oboru W^* testovanej štatistiky $T(\mathbf{X})$, ktorá sa viaže k nulovej hypotéze. Ak padne pozorovaná hodnota $T(\mathbf{X})$ do W^* zamietame H_0, v opačnom prípade nezamietame H_0.

Pri takomto spôsobe rozhodovania sa môžeme dopustiť dvoch chýb. Ak nulová hypotéza platí a my ju zamietneme, dopúšťame sa chyby označovanej ako

chyba I. druhu. Pravdepodobnosť, že nastane takáto situácia nazývame **hladina významnosti** a značíme α. Ak sme sa rozhodli s pravdepodobnosťou $1 - \alpha$ správne, platí nulová hypotéza a my sme ju nezamietli, hovoríme o **spoľahlivosti testu**. správnym rozhodnutím je aj zamietnutie nulovej hypotézy s pravdepodobnosťou $1 - \beta$ ak platí alternatívna hypotéza. S pravdepodobnosťou β nazývanou **sila testu** sa ale môžeme dopustiť **chyby II. druhu** keď nezamietneme nulovú hypotézu hoci platí alternatívna hypotéza.

O kvalite testu tak rozhodujú pravdepodobnosti α a β. Minimalizovať obe chyby súčasne nie je možné a tak sa v štatistike zvolila, ako rozhodujúci parameter, hladina významnosti α ($\alpha = 0,05$) pri ktorej sa minimalizuje β bežne na hodnotu $0,1$. Ďalšie praktické rady a ilustračné príklady možno nájsť v už zmienených učebniciach štatistiky [1, 3, 12].

1.6 Analýza časového radu

Časový rad predstavuje usporiadaná množina dvojíc čísel, z ktorých prvé číslo reprezentuje hodnota t časovej premennej T a druhé číslo hodnota y_t premennej časového radu Y. Vzhľadom na rozdielny charakter javov, ktorých vývoj zobrazujeme a na rozdiely vo frekvencii zaznamenávania, rozlišujeme spravidla tieto časové rady [11]:

Podľa charakteru zobrazovaných javov:

- **Časové rady nepretržite sa vyskytujúcich javov**, ktoré sa trvalo vyskytujú a informácie o nich možno získať v ktoromkoľvek časovom okamihu napr. zásoby firmy, ceny tovarov, mesačné výkony zamestnancov, atď.
- **Časové rady postupne vytváraných javov**, ktoré vznikajú vždy v nejakom časovom úseku napr. tržby obchodnej organizácie, produkcia firmy, spotreba obyvateľov a pod., a to viacmenej sústavne.
- **Časové rady prechodne sa vytváraných javov**, ktoré sa vyskytujú len v určitých časových intervaloch napr. ročné prémie zamestnancov, priemerné počty zákazníkov obchodno reťazci v obedňajšej špičke, atď.

Podľa postupu pri zaznamenávaní údajov:

- **Spojité**, ak sa informácie o určitom jave nepretržite vytvárajú a zaznamenávajú napr. stav zásob, stav vozidlového parku atď.
- **Diskrétne**, ak sa robia pozorovania iba v určitých okamihoch napr. k určitému termínu.

Podľa typu veličín:

- **Absolútne veličiny**, ktoré sú bezprostredným výsledkom pozorovania javov, pričom sa rozlišujú okamžité veličiny a intervalové veličiny.

- **Odvodené veličiny**, ktoré nie sú bezprostredným výsledkom pozorovania javov ale sa zisťujú výpočtom z pozorovaných absolútnych veličín napr. pomerné čísla, priemery, súčty atď.

Voľba metódy pre analýzu časových radov závisí od mnohých faktorov:

- **Účel analýzy**, ktorým sa spravidla rozpoznáva mechanizmus generovania hodnôt časového radu a predpovedanie jeho budúceho vývoja.
- **Typ časového radu**, voľba z viacerých modelov.
- **Skúsenosť štatistika**, ktorý realizuje analýzu a použitého softvéru.

Najjednoduchší spôsob analýzy časového radu vychádza z predpokladu, že jediným faktorom dymamiky ukazovateľa je čas, takže ho môžeme vyjadriť v tvare

$$y_t = f(t) + e_t,$$

kde y_t je hodnota analyzovaného ukazovateľa v čase t, $f(t)$ je funkcia času, t je časová premenná a e_t je hodnota náhodnej zložky. Pri takejto **jednorozmernej analýze** sa vychádza z dlhodobej skúsenosti, že časové rady z ekonomického prostredia majú charakteristické zložky:

- **Trendová zložka** Tr_t odráža dlhodobé zmeny v priemernom správaní sa časového radu, vzniká pôsobením síl v jednom smere. napr. pri predaji tovarov sú takýmito silami inovačné meny, zmeny kúpyschopnosti obyvateľov ale aj klimatické zmeny atď.
- **Sezónna zložka** Sz_t popisuje periodické zmeny v časovom rade napr v priebehu kalendárneho roka a každý rok sa opakujú.
- **Cyklická zložka** C_t — obtiažne verifikovateľná dĺžka medzi dvomi hornými alebo dolnými hodnotami, nebýva konštantná rovnako ako intenzita jej fáz. napr. obchodný cyklus (business cycle) je charakteristický rastom a potom poklesom ekonomickej aktivity s dĺžkou 5–7 rokov.
- **Reziduálna zložka** e_t zostáva v časovom rade po odstránení jeho trendu, sezónnej a acyklickej zložky. Je tvorení fluktuáciami, ktoré nemajú pozorovateľný systematický charakter. Často sa v modeloch predpokladá že je bielym šumom dokonca s normálnym rozdelení.

Modelu $y_t = Tr_t + Sz_t + C_t + e_t$ hovoríme **aditívny model** a modelu $y_t = Tr_t \cdot Sz_t \cdot C_t \cdot e_t$ **multiplikatívny model**. Medzi základné charakteristiky jednorozmerných časových radov patria:

- Absolútna diferencia: $\Delta y_t = y_t - y_{t-1}$, $t = 2, 3, \ldots, n$.
- Priemerný absolútny prírastok: $\overline{\Delta} = \frac{\sum_{t=2}^{n} \Delta y_t}{n-1}$.
- Relatívny prírastok: $\delta_t = \frac{\Delta y_t}{y_{t-1}}$.

- Koeficient rastu: $k_t = \frac{y_t}{y_{t-1}}$.
- Priemerný koeficient rastu: $\overline{k} = \sqrt[n-1]{k_1 \cdot k_2 \dots k_n}$.

Popri jednorozmerných modeloch sa v manažérskej praxi stretávame s modelmi, ktoré sú založené na predpoklade, že analyzovaný ukazovateľ nie je ovplyvňovaný len časovým faktorom ale aj množstvom iných. Takýto **viacrozmerný model** môžeme napísať v tvare

$$y_t = f(t, x_1, x_2, \dots, x_n, e_t),$$

kde $x_1, x_2, \dots, x_n$ sú ukazovatele ovplyvňujúce jav y_t.

Možností ako modelovať vývoj analyzovaného ukazovateľa v časovom rade je viac. Patria sem aj **autoregresné modely**

$$y_t = a_0 + a_1 y_{t-1} + \dots + a_k y_{t-k} + e_t,$$

kde $y_{t-1}, x_{t-2}, \dots, x_{t-k}$ sú analyzované hodnoty časového radu v časoch $t-1$, $t-2, \dots, t-k$ a $a_0, a_1, \dots, a_k$ sú neznáme parametre.

Iný prístup k analýze časových radov, založený na jej reziduálnej zložke berie **Box-Jenkisova metodológia** , ktorá môže byť tvorená aj korelovanými (závislými) náhodnými veličinami. Jeden z najjednoduchších modelov, konštruovaný na tomto základe, je **model kĺzavých súčtov prvého rádu**

$$y_t = e_t + 0,7 e_{t-1},$$

kde y_t je modelovaný rad a e_t je biely šum. Medzi ďalšie modely konštruované na báze Box-Jenkisovej metodológie sú tzv. **autoregresné modely AR** a **zmiešané modely ARMA**. Tieto modely sú oveľa flexibilnešie než modely konštruované na základe dekompozície.

1.7 Štatistické metódy riadenia kvality

V jednoduchom výrobnom procese sa používajú namerané hodnoty nielen na operatívne riadenie (ubrať/pridať suroviny, spĺňa/nespĺňa kritéria kvality) ale aj na zložitejšie postupy dodržiavania predpísaných parametrov.

Analyzujme výrobný parameter X, o ktorom predpokladáme, že má normálne rozdelenie $N(\mu, \sigma^2)$. Odhadom strednej hodnoty μ je aritmetický priemer $\overline{X}$ a odhadom smerodajnej odchýlky je výberová smerodajná odchýlka S. Tieto veličiny v zjednodušenej podobe charakterizujú reálny priebeh výrobného procesu, pre ktorý máme predpísané dve hodnoty:

LSL – Lower Specification Limit (dolná tolerančná medza).

USL – Upper Specification Limit (horná tolerančná medza).

Ak v i-tom pozorovaní nameraná hodnota x_i premennej X nachádza v tolerančnom páse t. j. $LSL \leq x_i \leq USL$, potom výrobný proces dodržiava predpísané parametre LSL a USL a vyrába sa v s predpísanou kvalitou. Ak je ale

hodnota x_i mimo tolerančný pás t. j. je menej ako dolná tolerančná medza LSL alebo viac ako horná tolerančná medza USL tak pozorujeme, že sa vyrába nekvalitne, predpísané hodnoty sa nedodržiavajú.

Na posúdenie kvality výrobného procesu sa používa súbor nameraných hodnôt tak, že sa z $x_1, x_2, \ldots, x_n$ vypočíta priemer $\overline{x}$ a smerodajná odchýlky s a použijú rôzne indexy spôsobilosti (capability):

- **Index spôsobilosti** $C_p = \frac{USL-LSL}{6s}$ vyjadruje schopnosť výrobného procesu dodržiavať predpísanú variabilitu porovnaním šírky tolerančného pásu so šiestimi smerodajnými odchýlkami reálneho procesu.
- **Koeficient centrovanosti** $K = \frac{\left|\overline{x} - \frac{LSL+USL}{2}\right|}{\frac{USL-LSL}{2}}$ meria do akej miery sa líšia nameraný stred $\overline{x}$ od zadaného stredu $\frac{LSL+USL}{2}$.
- **Taguchiho index spôsobilosti** $C_{PM} = \frac{USL-LSL}{6\sqrt{s^2+(\overline{x}-T)^2}}$ sa používa ak poznáme cieľovú hodnotu T, ktorú má dosahovať premenná X.

Regulačné diagramy sa používajú ak poznáme nejaký vzorový priebeh výrobného procesu a chceme sledovať jeho dodržiavanie. Prvou úlohou je určiť regulačné medze v ktorých sa má pohybovať regulovaný parameter, sú to

LCL – Lower Control Limit (dolná regulačná medza).

UCL – Upper Control Limit (horná regulačná medza).

Druhou úlohou je samotné monitorovanie výrobného procesu. Jej cieľom je reagovať na situácie keď je regulovaný parameter mimo regulačných medzí. Pri regulácii sa porovnáva počet nezhôd (chýb, odchýlok) na výrobkoch resp. na jeden výrobok pomocou testov príslušných štatistických hypotéz. Podrobnejší úvod do problematiky, postačujúci pre potreby manažmentu, možno nájsť v [3].

Literatúra

[1] ANDĚL, J.: *Statistické metody*, Vydavatelství MATFYZPRESS, Praha, (1998), ISBN 80-85863-27-8.

[2] ANDĚL, J.: *Matematika náhody*, Vydavatelství MATFYZPRESS, Praha, (2004), ISBN 80-85863-52-9.

[3] CHAJDIAK, J.: *Štatistika jednoducho*, Vydavateľstvo STATIS, Bratislava, (2010), ISBN 978-85659-60-3.

[4] JUREČKOVÁ, M., MOLNÁROVÁ, I.: *Štatistika s Excelom*, Vydavateľstvo AOS, Liptovský Mikuláš, (2015), ISBN 80-8040-257-4.

[5] LIKEŠ, J., MACHEK, J.: *Počet pravdepodobnosti*, Sešit X, Matematika pro vysoké školy technické, SNTL – Naladatelství technické literatury, Praha, (1981).

[6] LINDA, B.: *Stochastické modely operačního výskumu*, Vydavatelství STATIS, Bratislava, (2004), ISBN 80-8569-33-6.

[7] LITSCHMANNOVÁ, M.: *Vybrané kapitoly z pravděpodobnosti*, skripta, VŠB–TU, FEI, Ostrava, (2011), `http://mi21.vsb.cz/modul/vybrane-kapitoly-z-pravdepodobnosti`.

[8] HABÁK, P., KAHOUNOVÁ, J.: *Počet pravděpodobnosti v příkladech*, Vydavatelství INFORMATORIUM, Praha, (2005), ISBN 80-733-040-7.

[9] PALÚCH, S., PEŠKO, Š.: *Kvantitatívne metódy v logistike*, EDIS vydavateľstvo, Žilina, (2006), ISBN 80-8070-636-0.

[10] PEŠKOVÁ, A.: *Návrh a overenie mechanizmu optimalizácie riadenia procesu údržby a obnovy v diskrétnych výrobách* Dizertačná práca. Košice: TU SjF, (2015), 110s.

[11] PIDANY, J.: *Metódy porovnávania a sledovania dynamiky vývoja v ekonomike*, ELFA s.r.o. vydavateľstvo, Košice, (1996), ISBN 80-88786-37-1.

[12] RIEČAN, B., LAMOŠ, F., LENÁRT, C.: *Pravdepodobnosť a matematická štatistika*, ALFA vydavateľstvo, Bratislava, (1984).

[13] SAKÁL, P., JERZ, V.: *Operačná analýza v praxi manažéra*, TRIPSOFT, Edícia teória a prax manažérstva 2, Trnava, (2003), ISBN 80-968734-3-1.