INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 25, číslo 2, červen 2014

VOLBA PREZIDENTA: PROBLÉM KONTROLY PODPISŮ

PRESIDENTAL ELECTIONS: THE PROBLEM OF INSPECTION OF THE SIGNATURES

Ondřej Vencálek

Adresa: Katedra matematické analýzy a aplikací matematiky, PřF, Univerzita Palackého v Olomouci, 17. listopadu 12, 771 46 Olomouc

E-mail: ondrej.vencalek@upol.cz

Abstrakt: V článku jsou vysvětleny základy statistické přejímky, a to na poněkud netradičním příkladě kontroly podpisů na peticích předkládaných kandidáty na funkci prezidenta republiky.

Klíčová slova: kontrola, podpis, prezident, statistická přejímka srovnáváním, volby.

Abstract: In this paper we explain basics of the acceptance sampling. We use quite unusual example – inspection of signatures on petitions for support of presidential candidates.

Keywords: inspection, signature, president, acceptance sampling, elections.

1. Úvod – kontroverzní vzorec

V lednu 2013 oprávnění občané České republiky volili v přímé volbě prezidenta republiky. V prvním kole vybírali z devíti kandidátů, z nichž tři byli nominováni poslanci či senátory ČR, zatímco zbývajících šest kandidovalo na základě tzv. občanské nominace podpořené peticemi podepsanými nejméně 50 000 občany ČR staršími 18 let.

Správnost údajů na petičních arších byla u jednotlivých kandidátů kontrolována Ministerstvem vnitra ČR na dvou náhodně vybraných vzorcích, z nichž každý obsahoval 8 500 podpisů. Po provedení kontroly Ministerstvo vnitra ČR rozhodlo o registraci nebo odmítnutí kandidátních listin jednotlivých kandidátů na funkci prezidenta.

Velkou pozornost vzbudil způsob, jakým ministerstvo „vypočítalo“ počet započtených občanů podepsaných pod peticí (viz [1]). Ten vycházel z relativní chybovosti zjištěné ve dvou kontrolních vzorcích. Výpočet byl proveden podle vzorce

$$V_p = C_p - C_p \frac{C_1 + C_2}{100},$$

kde

- V_p je výsledný počet započtených občanů podepsaných na petici,
- C_p je celkový počet občanů podepsaných na petici,
- C_1 je relativní chybovost kontrolního vzorku č. 1 v %,
- C_2 je relativní chybovost kontrolního vzorku č. 2 v %.

Nesmyslnost tohoto vzorce netřeba více komentovat. Bezprostředně po zveřejnění způsobu výpočtu se médii začaly šířit vesměs pobouřené, ale i pobavené reakce. A tak tu byl Bart Simpson, který na tabuli „za trest“ vypisoval větu „Už nikdy nebudu sčítat procenta dvou kontrolních vzorků místo zprůměrování“ a bylo tu spoustu dalších podobných vtipů.

Odhalit banální chybu je v tomto případě jednoduché. Zkusme se však vrátit do okamžiku, kdy ještě nebyl systém kontroly správnosti údajů na peticích vůbec navržen. Odkud se vzalo číslo 8 500? Proč byly vzorky dva? Proč se nekontrolovaly všechny podpisy? Jak byla stanovena hranice 3 % chybových hlasů? Tyto otázky vůbec nemají jednoduché odpovědi. Cílem tohoto příspěvku je ukázat způsob, jak lze navrhnout pravidlo pro kontrolu. V závěru příspěvku ukážeme, že jde vlastně o dobře známý problém statistické přejímky srovnáváním.

2. Volba řečí paragrafů

Prezidenta České republiky volí občané přímou volbou. Ta byla zavedena ústavním zákonem č. 71/2012 Sb., který vstoupil v účinnost 1. října 2012. Ve stejný den nabyl účinnost i prováděcí zákon č. 275/2012 Sb., který stanoví podrobnosti navrhování kandidátů, vyhlašování a provádění volby, vyhlašování jejího výsledku a možnost soudního přezkumu.

V části 25 týkající prováděcího zákona nazvané Náležitosti kandidátní listiny se doslova uvádí: „Podává-li kandidátní listinu navrhující občan, připojí petici podepsanou alespoň 50 000 občany oprávněnými volit prezidenta republiky. [...] Každý občan, podporující kandidaturu kandidáta, uvede na podpisový arch své jméno, příjmení, datum narození a adresu místa trvalého pobytu a připojí vlastnoruční podpis. [...] Ministerstvo vnitra ověří správnost údajů na peticích namátkově na náhodně vybraném vzorku údajů u 8 500 občanů podepsaných na každé petici. Zjistí-li nesprávné údaje u méně než 3 % podepsaných občanů, nezapočítá Ministerstvo vnitra tyto občany do celkového počtu občanů podepsaných na petici. Zjistí-li Ministerstvo vnitra postupem podle odstavce 5 nesprávné údaje u 3 % nebo více než 3 % podepsaných občanů, provede kontrolu u dalšího vzorku stejného rozsahu (dále jen

„druhý kontrolní vzorek“). Zjistí-li Ministerstvo vnitra, že druhý kontrolní vzorek vykazuje chybovost u méně než 3 % občanů podepsaných na petici, nezapočítá Ministerstvo vnitra občany z obou kontrolních vzorků do celkového počtu občanů podepsaných na petici. Zjistí-li Ministerstvo vnitra, že druhý kontrolní vzorek vykazuje chybovost u 3 % nebo více než 3 % občanů podepsaných na petici, odečte od celkového počtu občanů podepsaných na petici počet občanů, který procentuálně odpovídá chybovosti v obou kontrolních vzorcích.“

3. Model náhodné kontroly

Na začátku je třeba vyjasnit, proč vlastně dělat kontrolu jen „namátkově“, tedy proč nekontrolovat všechny podpisy. Nejčastějším důvodem pro provedení takovéto kontroly bývá časová náročnost úplné kontroly. Náklady na kontrolu jsou dalším faktorem, který hovoří pro kontrolu jen malé části podpisů.

Rozhodneme-li se kontrolovat jen část podpisů, pak bychom to měli udělat tak, aby nikdo dopředu nemohl určit, které z podpisů budou kontrolovány. Takovouto volbou podpisů ke kontrole je volba náhodná, či alespoň „pseudonáhodná“, jak nám ji nabízejí počítačové generátory (pseudo)náhodných čísel. *Pozn. redakce: Toto však dohoda mezi Ministerstvem vnitra ČR a firmou zajišťující kontrolu podpisů jednoznačně nesplňovala; nešlo o náhodný výběr.*

Představme si kandidáta, jehož kandidatura je podložena celkem N podpisy (např. $N = 60\,000$). Část z nich je „správných“, část je „nesprávných“. Počet správných podpisů budeme značit A, počet nesprávných je tedy $N - A$. Číslo A je pro nás neznámé – nevíme, kolik z N předložených podpisů je správných a kolik nesprávných. Minimální počet správných podpisů, který je nutno shromáždit, je 50 000. Tento počet označme symbolem A_0. Pokud tedy $A \geq A_0$, kandidát splnil zákonem předepsané podmínky a jeho kandidatura by měla být registrována; v opačném případě by měla být odmítnuta.

Naším cílem je navrhnout:

1. Kolik podpisů zkontrolovat (počet kontrolovaných podpisů budeme značit písmenem n); chceme, aby počet podpisů, které budeme kontrolovat, byl co možná nejmenší.
2. Jak se na základě zjištěné chybovosti ve vzorku rozhodnout, zda kandidaturu přijmout (což bychom měli učinit, pokud $A \geq A_0$) či nikoliv (pokud $A < A_0$).

Druhý z těchto dvou problémů se zdá být relativně snadno řešitelný. Kandidaturu odmítneme, když bude počet nesprávných podpisů ve vzorku

(označme jej Z_n) příliš velký – větší než nějaká, nám zatím neznámá, hodnota c. Představme si například kandidáta, který předložil 60 000 podpisů. Z 1000 kontrolovaných podpisů jich 900 je nesprávných. Pokud by chybovost celého souboru dat byla stejná jako chybovost kontrolovaného vzorku, tj. 90 %, byl by skutečný počet správných podpisů jen 6 000. Tedy takovouto kandidaturu bychom neměli uznat. Naše rozhodnutí o přijetí či odmítnutí kandidatury bude mít tedy podobu:

$$\begin{aligned} Z_n > c &\implies \text{odmítáme kandidaturu,} \\ Z_n \leq c &\implies \text{přijímáme kandidaturu.} \end{aligned}$$

Je však třeba si uvědomit, že pokud kandidaturu neuznáme, mohli jsme se dopustit chyby. Našli jsme sice 900 nesprávných podpisů, ale je možné, že žádné další nesprávné podpisy už nejsou a jen díky náhodě bylo všech 900 nesprávných podpisů zrovna mezi 1000 kontrolovaných. V takovém případě by nebyl správný náš předpoklad, že chybovost celého datového souboru je stejná jako chybovost kontrolovaného vzorku. Ano, může se to stát. Tato možnost je však vysoce nepravděpodobná.

Z výše uvedeného příkladu vyplývá, že pokud neuděláme úplnou kontrolu všech podpisů, musíme se vyrovnat s tím, že naše rozhodnutí o zamítnutí či nezamítnutí kandidatury může být chybné. Můžeme přitom chybovat dvěma způsoby:

1. Kandidaturu kandidáta, který splňuje podmínku 50 000 správných podpisů, odmítneme (viz předchozí případ).

2. Kandidaturu kandidáta, který nesplňuje podmínku 50 000 správných podpisů, přijmeme.

Chyba je tedy možná (pokud neprovedeme úplnou kontrolu), nicméně vhodnou volbou počtu kontrolovaných podpisů n a pravidla pro přijetí kandidatury daného číslem c, můžeme zajistit, aby pravděpodobnost chybného rozhodnutí byla malá. Jak malá má tato pravděpodobnost být, je věcí dohody. Čím přísnější požadavky na pravděpodobnost možných chyb budeme mít, tím větší počet podpisů bude nutno kontrolovat. V extrémním případě, kdy bychom požadovali nulovou pravděpodobnost chyb, bychom u každého z kandidátů splňujících zákonem danou podmínku na počet podpisů museli zkontrolovat nejméně 50 000 podpisů. Navíc bychom museli předpokládat, že neuděláme chybu při kontrole.

Naším cílem je tedy určit počet kontrolovaných podpisů n a číslo c tak, aby zároveň platilo:

$$\begin{aligned} \mathrm{P}(Z_n > c | A \ \geq A_0) &\leq \alpha, \\ \mathrm{P}(Z_n \leq c | A \ < A_0) &\leq \beta, \end{aligned}$$

tedy, že pravděpodobnost odmítnutí kandidatury, když kandidát splňuje podmínku 50 000 podpisů, je nejvýše α a pravděpodobnost přijetí kandidatury kandidáta nesplňujícího podmínku 50 000 podpisů je nejvýše β. Hodnoty α, β jsou předem dohodnuté, malé; například $\alpha = \beta = 0{,}01$.

Tento cíl se však ukáže jako příliš ambiciozní. A to proto, že výše uvedený požadavek mimo jiné znamená, že pro kandidáta, který má přesně 50 000 podpisů, musí být pravděpodobnost odmítnutí jeho kandidatury malá – menší než α – zatímco u kandidáta, který má o jediný správný podpis méně, už požadujeme, aby pravděpodobnost odmítnutí jeho kandidatury byla velká – větší než $1 - \beta$. Rozdíl jediného správného podpisu přitom nejsme schopni odhalit jinak než úplnou kontrolou.

Řešení, které se nabízí, vypadá takto: místo hranice $A_0 = 50\,000$ oddělující od sebe vyhovující a nevyhovující kandidatury, rozlišujme kandidatury „bezpečně nevyhovující“, což jsou kandidatury s počtem podpisů nepřesahujícím hranici A_1 (např. $A_1 = 49\,500$), a kandidatury „bezpečně vyhovující“, což jsou kandidatury s počtem podpisů alespoň A_2 (např. $A_2 = 50\,500$). Hodnota těchto hranic je věcí dohody, stejně tak jako je věcí dohody volba hodnoty A_0 (číslo A_0 je dáno zákonem, který byl dohodnut a odsouhlasen zákonodárci). Tato čísla však musí být známa dopředu. Každý kandidát pak bude mít možnost stát se „bezpečně vyhovujícím“ kandidátem, tj. bude počítat s tím, že jestliže nasbírá méně než A_2 platných hlasů, nebude mít záruku velké pravděpodobnosti, že bude jeho kandidatura schválena. Naopak občan bude mít jistotu, že pokud nějaký kandidát nasbírá méně než A_1 platných podpisů, bude s velkou pravděpodobností jeho kandidatura odmítnuta. Naším úkolem je určit hodnoty n a c tak, aby platilo:

$$\mathrm{P}(Z_n > c | A \ \geq A_2) \leq \alpha, \tag{1}$$

$$\mathrm{P}(Z_n \leq c | A \ \leq A_1) \leq \beta. \tag{2}$$

Počet kontrolovaných podpisů n bude do značné míry ovlivněn velikostí „šedé zóny mezi bezpečně vyhovujícími a bezpečně nevyhovujícími kandidáty“, tedy hodnotou rozdílu $A_2 - A_1$ (v námi uvedeném případě $50\,500 - 49\,500 = 1\,000$). Čím bude tato zóna užší (rozdíl menší), tím více podpisů bude třeba kontrolovat (jak jsme již řekli, pro $A_2 - A_1 = 1$ bude třeba kon-

troly všech podpisů. Čím blíž jsou si A_1 a A_2, tím větší nuance musíme být schopni rozeznat, a tím víc k tomu potřebujeme kontrolovaných podpisů.

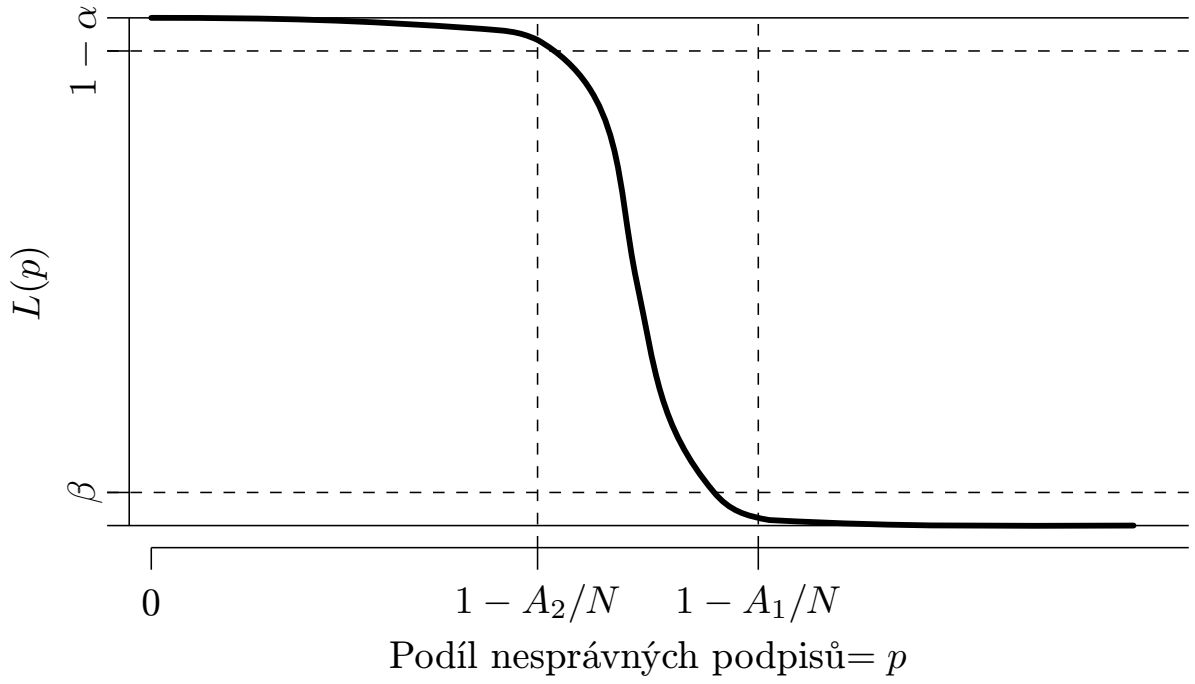
Máme-li zabezpečit splnění požadavku (1), musíme zabezpečit, aby $\mathrm{P}(Z_n > c|A = a) \leq \alpha$, pro všechna přirozená čísla a v rozmezí A_2 až N. Čím je hodnota A větší, tím spíše bude požadované nerovnosti dosaženo. Pokud by podpisový arch neobsahoval žádné nesprávné podpisy ($A = N$), nemohlo by dojít k odmítnutí kandidáta a tedy nerovnost (1) by byla triviálně splněna. Čím bude skutečný počet správných podpisů A menší, tím spíše může dojít k zamítnutí kandidatury. Bude-li však splněna nerovnost $\mathrm{P}(Z_n > c|A = A_2) \leq \alpha$, bude splněn požadavek (1). Analogickou úvahu můžeme použít i na požadavek (2). Oba tyto požadavky budou splněny, bude-li splněno:

$$\mathrm{P}(Z_n > c|A = A_2) \leq \alpha, \tag{3}$$

$$\mathrm{P}(Z_n \leq c|A = A_1) \leq \beta. \tag{4}$$

Závislost pravděpodobnosti přijetí kandidatury na podílu nesprávných podpisů znázorňuje pro pevně dané n a c tzv. operativní charakteristika (viz obrázek 1).

Na x-ové ose jsou vyznačeny podíly odpovídající situacím, kdy $A = A_1$ (vpravo) resp. $A = A_2$ (vlevo). Splnění požadavků (3) a (4) pak můžeme



Obrázek 1: Závislost pravděpodobnosti přijetí kandidatury na podílu nesprávných podpisů (operativní charakteristika).

posoudit podle y-ových hodnot v těchto bodech. Na obrázku 1 je zachycena situace, kdy jsou oba požadavky splněny. Bylo uvažováno $N = 60\,000$, $A_1 = 48\,000$, $A_2 = 52\,000$, $n = 500$, $c = 80$, $\alpha = \beta = 0{,}05$. V tomto případě je $\mathrm{P}(Z_n > c|A = A_2) = 0{,}0364 \leq 0{,}05$ a $\mathrm{P}(Z_n \leq c|A = A_1) = 0{,}0127 \leq 0{,}05$.

Pravděpodobnosti ze vztahů (3) a (4) umíme explicitně vyjádřit pomocí n a c. Platí-li rovnost $A = A_2$, jako je tomu ve vztahu (3), můžeme Z_n, tedy počet nesprávných podpisů mezi n kontrolovanými podpisy, považovat za náhodnou veličinu s hypergeometrickým rozdělením s parametry $(N,\, N - A_2,\, n)$. Podobně platí, že za podmínky $A = A_1$ má počet špatných podpisů v kontrolovaném vzorku hypergeometrické rozdělení s parametry $(N,\, N - A_1,\, n)$. Pro náhodnou veličinu s hypergeometrickým rozdělením s parametry $(N,\, N - A_i,\, n)$ a libovolné nezáporné celé číslo k nepřesahující hodnotu n platí:

$$\mathrm{P}(Z_n = k|A = A_i) = \begin{cases} \dfrac{\binom{N-A_i}{k}\binom{A_i}{n-k}}{\binom{N}{n}} & \text{pro } 0 \leq k \leq n, \\ & n - A_i \leq k \leq N - A_i, \\ 0 & \text{jinak.} \end{cases} \tag{5}$$

Podmínka $k \leq N - A_i$ zajišťuje, že počet nesprávných podpisů ve vzorku nemůže převýšit celkový počet nesprávných podpisů $N - A_i$; podmínka $k \geq n - A_i$ říká, že počet správných podpisů ve vzorku $(n - k)$ nemůže převýšit celkový počet správných podpisů A_i.

Podmínky (3) a (4) můžeme zapsat v podobě:

$$\sum_{k=c+1}^{n} \mathrm{P}(Z_n = k|A = A_2) \quad \leq \quad \alpha, \tag{6}$$

$$\sum_{k=0}^{c} \mathrm{P}(Z_n = k|A = A_1) \quad \leq \quad \beta, \tag{7}$$

kde pravděpodobnosti $\mathrm{P}(Z_n = k|A = A_i)$ spočteme pomocí vztahu (5), pro $i = 1, 2$.

Naším úkolem je tedy najít hodnoty n a c, pro které jsou splněny podmínky (6) a (7). Uvědomme si, že v těchto vztazích jsou jediné dvě neznámé hodnoty právě hodnoty n a c. Připomeňme zde význam ostatních symbolů: N je celkový počet podpisů pro kandidáta, A_1 je hranice „bezpečně nevyhovujících kandidátů“, A_2 je hranice „bezpečně vyhovujících kandidátů“, α je nejvyšší přípustná pravděpodobnost, že kandidatura bezpečně vyhovujícího kandidáta bude zamítnuta a β je nejvyšší přípustná pravděpodobnost, že kandidatura bezpečně nevyhovujícího kandidáta bude přijata. Hodnoty A_1, A_2, α a β musí být dopředu pevně dohodnuty.

Počet kontrolovaných podpisů n jistě nepřevýší celkový počet podpisů N a číslo c jistě nepřekročí hodnotu n. Možných dvojic n a c je tedy jen konečně mnoho. Není tedy problém (za pomoci počítače) ověřit, pro které dvojice n a c jsou splněny podmínky (6) a (7) a vybrat si takovou dvojici, kde n je nejmenší (naší snahou je, aby počet kontrolovaných podpisů byl co nejmenší).

Uvědomme si, že tímto způsobem nemusí být stanovený počet podpisů nějaké „hezké“ číslo jako 8 500, ale n může být stanoveno „podivně“, např. 8 327. Takové číslo by mohlo vzbudit nedůvěru širší veřejnosti. Proto můžeme množinu přípustných hodnot n redukovat např. na celočíselné násobky čísla 100 a množinu hodnot přípustných pro c na celočíselné násobky čísla 10.

Algoritmus pro nalezení optimální dvojice n a c by pak mohl vypadat takto:

1. Volme $n = 100$.
2. Pro dané n najděme co nejmenší c (násobek 10) tak, aby byla splněna podmínka (6).
3. Ověřme, zda je pro danou dvojici n a c splněna podmínka (7):
   - pokud je splněna, ukončeme hledání,
   - pokud není splněna, zvětšeme hodnotu n o 100 a opakujme krok 2.

Poznamenejme, že pokud bychom připustili, že n a c mohou být libovolná přirozená čísla, přineslo by nám to další úsporu potřebných kontrolovaných hlasů (viz diskuse na konci sekce 5).

4. Statistická přejímka

Výše popsaný problém určení počtu podpisů n, které se mají kontrolovat, a čísla c určujícího maximální počet odhalených chyb, který ještě nevede k zamítnutí kandidatury, je *problémem statistické přejímky* (anglicky Acceptance Sampling). Myslím si, že příklad volby prezidenta může studentům seznamujícím se s teorií statistické přejímky (či obecněji statistické kontroly kvality) pomoci pochopit některé nově zaváděné pojmy. Připomeňme zde tedy terminologii statistické přejímky; české pojmy lze najít např. v učebnici [2], anglické ekvivalenty pak v elektronické publikaci [3]. Statistická přejímka srovnáváním je předmětem českých technických norem řady ČSN ISO 2859 [4].

Podíl nesprávných podpisů (či jiných kontrolovaných jednotek) určující hranici „bezpečně vyhovujícího kandidáta“ $1 - A_2/N$ se označuje jako *přijatelná/přípustná úroveň jakosti* (anglicky Acceptable Quality Level či Acceptance Quality Level; AQL), zatímco podíl nesprávných podpisů určující

„bezpečně nevyhovujícího kandidáta“ $1 - A_1/N$ se označuje jako *nepřijatelná/nepřípustná úroveň jakosti* (anglicky Limited Quality = LQ). V situaci, kdy kandidát splňuje zákonné požadavky ($A \geq 50\,000$) a přesto je kandidatura odmítnuta, dochází k „poškození zájmů“ kandidáta. Maximální možná pravděpodobnost, že k takové chybě dojde u „bezpečně vyhovujícího“ kandidáta (označená ve vztahu (1) symbolem α) je tedy rizikem kandidáta – dodavatele podpisů. V teorii statistické přejímky se pro symbol α vžilo označení *riziko dodavatele* (Producer's Risk). Naopak, není-li zamítnuta kandidatura kandidáta, který ve skutečnosti nemá 50 000 podpisů, můžeme to chápat jako poškození zájmů občanů. Maximální možná pravděpodobnost, že k chybě tohoto druhu dojde u „bezpečně nevyhovujícího“ kandidáta (označená ve vztahu (2) symbolem β) je tedy rizikem občana. Pravděpodobnost skrytá pod symbolem β bývá nazývána *rizikem odběratele* (Consumer's Risk). Pro úplnost ještě dodejme, že číslu n říkáme *rozsah výběru* (Sample Size), číslu c *rozhodné/přejímací číslo* (Acceptance Number) a dvojici (n, c) *přejímací plán*.

Postup přejímky uvedený v tomto článku je postup přejímky jedním výběrem (Single Sample Plan). Alternativou k němu by mohla být přejímka dvojím výběrem podobná postupu, který byl u volby prezidenta opravdu použit, nebo sekvenční přejímka. Pomocí těchto technik by mohl být výsledný počet podpisů, které je nutno zkontrolovat, zredukován. Je totiž například „zřejmé“, že předložil-li kandidát 100 tisíc podpisů a kontrola prvního vzorku náhodně vybraných sta podpisů neodhalí žádnou chybu, je další kontrola prakticky zbytečná.

5. Volba prezidenta 2013

Nyní se podívejme, kolik podpisů by bylo nutné zkontrolovat při prezidentských volbách na přelomu let 2012 a 2013. Výpočet byl proveden pro hodnoty parametrů $\alpha = \beta = 0{,}05$ a pro dvě různé dvojice hodnot určujících bezpečně vyhovujícího/nevyhovujícího kandidáta. V prvním případě je hodnota $A_1 = A - 2000$ a $A_2 = A + 2000$, zatímco ve druhém případě jsou si hodnoty A_1, A_2 blíže: $A_1 = A - 500$ a $A_2 = A + 500$. Výsledky včetně přípustné chybovosti, tj. podílu c/n, jsou uvedeny v tabulce 1 a v tabulce 2.

V prvním případě by tedy stačilo zkontrolovat celkem pouze 9 000 podpisů. „Přísnější“ určení dvojice parametrů A_1, A_2 povede k výrazně vyššímu počtu kontrolovaných podpisů (více než 100 tisíc). Za pozornost stojí také okolnost, že kandidátům s menším počtem předložených podpisů tolerujeme pouze malé procento nesprávných podpisů, zatímco kandidát, který odevzdá řekněme dvojnásobné množství podpisů než je třeba, může mít procento ne-

Kandidát	Počet podpisů (N)	n	c	Přípustná chybovost (%)
Bobošíková	56 191	300	30	10,0
Dlouhý	59 165	400	60	15,0
Okamura	61 966	700	130	18,6
Fischerová	72 434	900	280	31,1
Roithová	81 199	1200	460	38,3
Franz	87 782	1600	690	43,1
Fischer	101 261	1800	910	50,6
Zeman	106 018	2100	1110	52,9

Tabulka 1: Přejímací plán pro jednotlivé kandidáty při hodnotách $\alpha = \beta = 0{,}05$, $A_1 = 48\,000$, $A_2 = 52\,000$.

Kandidát	Počet podpisů (N)	n	c	Přípustná chybovost (%)
Bobošíková	56 191	3 800	420	11,1
Dlouhý	59 165	5 100	790	15,5
Okamura	61 966	6 600	1 270	19,2
Fischerová	72 434	11 200	3 470	31,0
Roithová	81 199	14 500	5 570	38,4
Franz	87 782	16 800	7 230	43,0
Fischer	101 261	22 300	11 290	50,6
Zeman	106 018	24 000	12 680	52,8

Tabulka 2: Přejímací plán pro jednotlivé kandidáty při hodnotách $\alpha = \beta = 0{,}05$, $A_1 = 49\,500$, $A_2 = 50\,500$.

správných podpisů téměř 50 %, což je chybovost, při které už by byl kandidát s menším počtem odevzdaných podpisů vyřazen. Obhájit „férovost“ tohoto postupu před laickou veřejností je úkolem nás statistiků. A není to úkol zrovna snadný.

Uveďme zde ještě pro úplnost, jaké počty podpisů by bylo nutno kontrolovat, kdybychom se neomezovali jen na násobky 100 (resp. 10 pro c). Tyto počty jsou uvedeny v tabulce 3. Při použití hodnot $A_1 = 48\,000$, $A_2 = 52\,000$

Kandidát	$A_2 - A_1 = 4\,000$		$A_2 - A_1 = 1\,000$	
	n	c	n	c
Bobošíková	211	22	3 170	348
Dlouhý	305	46	4 585	709
Okamura	405	77	5 868	1 132
Fischerová	752	232	10 409	3 223
Roithová	1 043	400	13 991	5 375
Franz	1 263	543	16 595	7 142
Fischer	1 727	874	21 819	11 045
Zeman	1 865	985	23 586	12 462

Tabulka 3: Přejímací plány pro jednotlivé kandidáty – bez použití zaokrouhlování hodnot n a c.

by stačilo místo 9 000 podpisů kontrolovat jen 7 571, při použití hodnot $A_1 = 49\,500, A_2 = 50\,500$ bychom místo 104 300 vystačili s 100 023 podpisy.

6. Závěr

Výše uvedený příklad volby prezidenta jsem použil při výuce předmětu Statistická kontrola kvality, který je určen studentům 2. ročníku bakalářského studijního programu Aplikovaná statistika na Univerzitě Palackého v Olomouci. Věřím, že takovýto (poněkud neobyklý) příklad může pomoci k pochopení problematiky statistické přejímky srovnáváním.

Věnováno památce Ing. Josefa Machka, který mě seznamoval se základy statistické přejímky.

Literatura

[1] `http://www.mvcr.cz/clanek/rozhodnuti.aspx`

[2] Piskáček, B., Kašová, V., Zmatlík, J.: *Řízení jakosti.* Praha: Vydavatelství ČVUT, 2001.

[3] *NIST/SEMATECH e-Handbook of Statistical Methods,* `http://www.itl.nist.gov/div898/handbook/`

[4] *ČSN ISO 2859-10 Statistické přejímky srovnáváním* – Část 10: Úvod do norem ISO řady 2859 statistických přejímek pro kontrolu srovnáváním. Praha: Český normalizační institut, 2007.

ŘÍZENÍ KVALITY V PROGRAMU XLSTATISTICS
QUALITY CONTROL IN XLSTATISTICS

Petr Klímek

Adresa: Univerzita Tomáše Bati ve Zlíně, Fakulta managementu a ekonomiky, náměstí T. G. Masaryka 5555, 760 01 Zlín

E-mail: klimek@fame.utb.cz

Abstrakt: Tento článek se zabývá regulačními diagramy, které je možné znázornit v programu XLStatistics. Program je souborem sešitů pro Microsoft Excel, který může sloužit pro výpočty mnoha typů statistických úloh. Po jeho stručném představení následuje část věnovaná regulačním diagramům. Pozornost je věnována jak regulaci měřením, tak regulaci srovnáváním. Po teoretickém představení regulačního diagramu následuje vždy jeho praktická ukázka v programu XLStatistics. V závěru je uvedeno celkové zhodnocení. programu.

Klíčová slova: XLStatistics, Microsoft Excel, řízení kvality, regulační diagramy, Paretova analýza.

Abstract: This paper deals with the control charts that can be displayed in XLStatistics software tool. The program is a set of Microsoft Excel workbooks that can be used for processing of many types of statistical tasks. A section devoted to the control charts diagrams follows after its brief introduction. Attention is given both to the control charts for numerical variable and to the control charts for attributes. The theoretical performance of each control chart is always followed by practical demonstration in XLStatistics. Final program evaluation is given at the end of the paper.

Keywords: XLStatistics, Microsoft Excel, quality control, control charts, Pareto analysis.

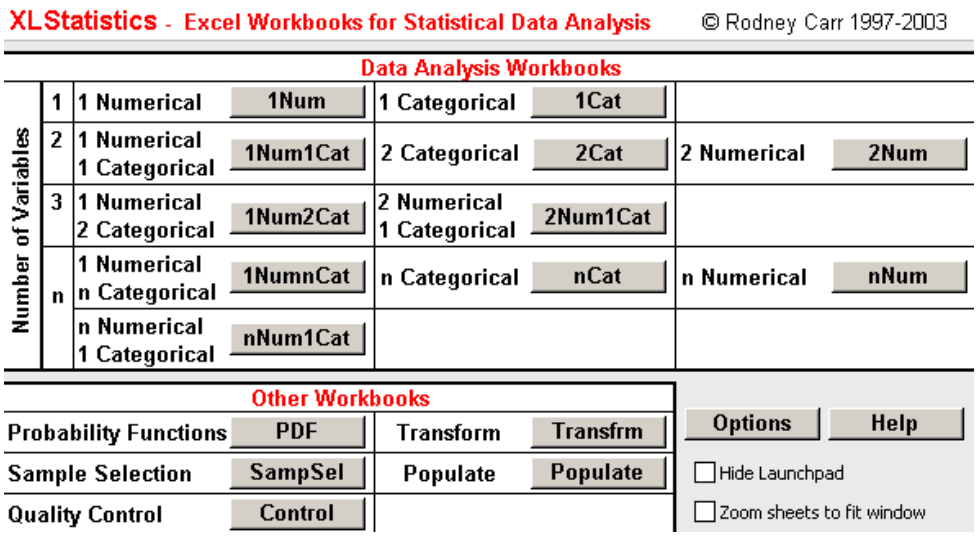
1. Úvod

XLStatistics je soubor sešitů pro Microsoft Excel (pro verze 97 a vyšší), který může sloužit pro výpočty mnoha typů statistických úloh. Jeho praktické využití bylo diskutováno v [3].

Jeho autorem je Dr. Rodney Carr z Deakin University Warrnambool (Austrálie). Pro zkušební účely a pro výuku je program zdarma, jinak licence stojí 30 AUD nebo 20 USD. S rostoucím počtem licencí cena za licenci ještě dále klesá. Vidíme, že v porovnání s komerčními statistickými softwarovými

produkty je tato cena víceméně symbolická. Je rovněž autorem programu XLMathematics. [1]

XLStatistics obsahuje 11 základních sešitů (Data Analysis Workbooks) pro analýzy jak numerických (Numerical), tak i kategoriálních (Categorial) proměnných. Dále zde můžeme nalézt dalších 5 sešitů (Other Workbooks), pomocí kterých lze provádět další možné analýzy a výpočty. Jejich kompletní seznam je na obrázku 1.



Obrázek 1: Hlavní sešit XLStatistics

XLStatistics spolupracuje s Microsoft Excelem a Wordem ve verzích 97 a vyšších. Požadavky na hardware jsou tedy shodné s těmito produkty. Po stažení z adresy `http://www.deakin.edu.au/~rodneyc/XLStatistics/` získáváme archiv `XLS6.zip` (cca 6 MB). Tento archiv rozbalíme do předem připravené nové složky. V něm je obsaženo celkem 90 souborů včetně manuálů ve Wordu. Nyní můžeme postupovat v následujících krocích:

1. Otevřeme soubor `XLStatistics.xlam` a zároveň vytvoříme na ploše jeho zástupce pro snadnější spuštění.
2. Skrytí hlavního souboru (volitelně). Jestliže na obrázku 1 zaškrtneme políčko Hide Launchpad můžeme pracovat s XLStatistics z hlavního menu na liště MS Excelu. Pro některé uživatele to může být pohodlnější.
3. Naše data umístíme do vlastního sešitu.
4. Označíme naše data, která chceme analyzovat.
5. Vybereme si vhodnou analýzu ze sešitů XLStatistics.
6. Data, která jsou v XLStatistics označena modře, můžeme nahradit našimi daty, ostatní buňky nelze editovat. Nápověda je skryta pod buňkami, které jsou označeny červeně.

7. Prohlédneme si výsledky a provedeme další úpravy, jestliže jsou potřebné.
8. Uložíme naše výsledky. Nelze ovšem ukládat přímo v sešitech XLStatistics, ale zvlášť v jiném sešitu pomocí kopírovat – vložit, protože sešity XLStatistics jsou navzájem provázány a naše úpravy by mohly způsobit chyby v dalších výpočtech. [2]

V tomto článku se budeme zabývat základními nástroji řízení kvality, které tento program umožňuje použít. Jsou obsaženy v sešitu Quality Control. Ten spustíme zmáčknutím tlačítka **Control** v tabulce **Other Workbooks** (obrázek 1). Tato nabídka je od verze Excelu 2007 realizovaná vlastním pruhem nástrojů v nabídce (ribbon menu).

2. Regulační diagramy pro regulaci měřením

Základním nástrojem SPC (Statistical Process Control; statistické řízení procesů) je regulační diagram. Je to grafický prostředek zobrazení vývoje variability procesu v čase využívající principů testování statistických hypotéz. Rozhodnutí o statistické zvládnutosti procesu umožňují 3 základní čáry: CL – střední přímka; odpovídá tzv. referenční (požadované) hodnotě použité znázorňované charakteristiky. Z hlediska účinnosti regulačního diagramu a základního rozhodnutí o statistické zvládnutosti procesu je rozhodující stanovení horní a dolní regulační meze:

- UCL je horní regulační mez (Upper Control Limit),
- LCL je dolní regulační mez (Lower Control Limit).

Těmto regulačním mezím se také říká akční meze. Vymezují pásmo působení pouze náhodných příčin variability a jsou základním rozhodovacím kritériem, zda učinit regulační zásah do procesu či nikoliv. V některých aplikacích se zakreslují do regulačního diagramu další meze nazývané výstražné meze: UWL (Upper Warning Limit – horní výstražná mez) a LWL (Lower Warning Limit – dolní výstražná mez). Pásmo, které vymezují tyto meze, je vždy užší než pásmo mezi akčními mezemi, nejčastěji $\pm 2\sigma$ od CL.

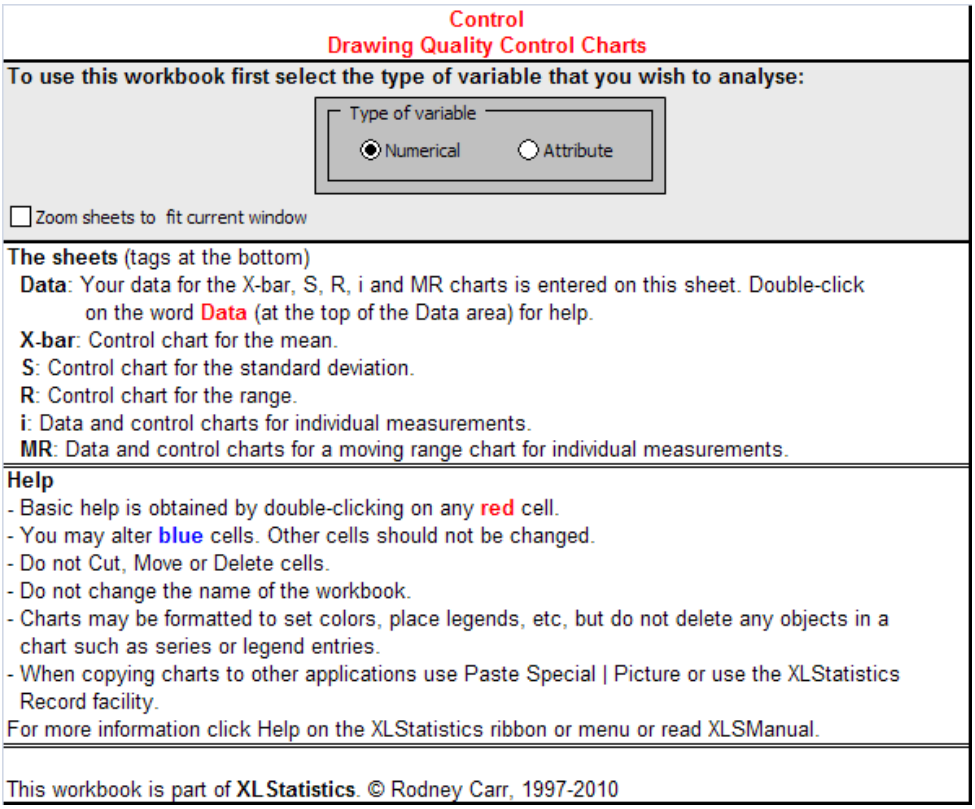
Regulační diagramy pro regulaci měřením (Control Charts for Variables) se používají pro měřitelné znaky jakosti či technologické parametry. [8]

Základní předpoklady podle [9] pro Shewhartův regulační diagram měřením jsou:

- normalita rozdělení dat, symetrie,
- konstantní střední hodnota procesu,
- konstantní rozptyl (směrodatná odchylka) dat,

- nezávislost dat,
- nepřítomnost vybočujících hodnot. [9]

Výše uvedené předpoklady se musí testovat před vlastní konstrukcí regulačního diagramu. Program XLStatistics umožňuje výpočty jednoduchých regulačních diagramů pro číselný nebo atributový typ proměnné. Problematika vícerozměrných diagramů je diskutována v publikaci [6]. Pokročilé výpočty stochastických diferenciálních rovnic pro detekci bodu změny ve volatilitě časových řad se zabývají publikace [5] a [7]. Po otevření sešitu **Quality Control** se otevře úvodní list na obrázku 2, na kterém zvolíme typ proměnné (Type of variable). Nejprve se zaměříme na regulaci měřením (Numerical).



Obrázek 2: Úvodní list sešitu Quality Control

2.1. Regulační diagram pro výběrové průměry ($\overline{x}$)

Testovým kritériem, jehož hodnoty se zakreslují do regulačního diagramu ($\overline{x}$), je výběrový průměr $\overline{x}$ z výběru o konstantním rozsahu n. Hodnota výběrového průměru v j-tém výběru $\overline{x}$ se vypočte dle vztahu:

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \tag{1}$$

kde x_{ij} je i-tá naměřená hodnota regulované veličiny v j-tém výběru. Jestliže zvolíme riziko zbytečného signálu $\alpha = 0{,}0027$ a neznáme cílové hodnoty μ_0 a σ_0, určíme CL následovně:

$$CL = \widehat{\mu}_0 = \overline{\overline{x}} = \frac{1}{k} \sum_{j=1}^{k} \overline{x}_j \tag{2}$$

Protože ve dvojici regulačních diagramů $(\overline{x}, R)$ se odhaduje variabilita procesu pomocí výběrového rozpětí R, použijeme pro stanovení odhadu směrodatné odchylky procesu $\widehat{\sigma}_0$ vztah:

$$\widehat{\sigma}_0 = \frac{\overline{R}}{d_2} \tag{3}$$

kde $\overline{R}$ je průměrné výběrové rozpětí ve výběrech, d_2 je Hartleyova konstanta závislá na rozsahu výběru n a odvozená za předpokladu regulované veličiny pocházející z normálního rozdělení. $\overline{R}$ se vypočte ze vztahu:

$$\overline{R} = \frac{\sum_{j=1}^{k} R_j}{k}, \tag{4}$$

kde k je počet výběrů použitých k výpočtu $\overline{R}$ (alespoň 20), R_j je výběrové rozpětí v j-tém výběru a stanoví se ze vztahu:

$$R_j = x_{\max,j} - x_{\min,j} \tag{5}$$

kde $x_{max,j}$ je největší naměřená hodnota v j-tém výběru, $x_{min,j}$ je nejmenší naměřená hodnota v j-tém výběru. Nyní dostaneme pro výpočet akčních regulačních mezí v diagramu $(\overline{x})$ tyto vztahy:

$$UCL = \overline{\overline{x}} + \frac{3}{\sqrt{n}} \cdot \frac{\overline{R}}{d_2} = \overline{\overline{x}} + A_2 \cdot \overline{R} \tag{6a}$$

$$LCL = \overline{\overline{x}} - \frac{3}{\sqrt{n}} \cdot \frac{\overline{R}}{d_2} = \overline{\overline{x}} - A_2 \cdot \overline{R} \tag{6b}$$

Hodnoty A_2 a d_2 pro n od 2 až do 25 jednotek nalezneme v normě ČSN ISO 8258. [4]

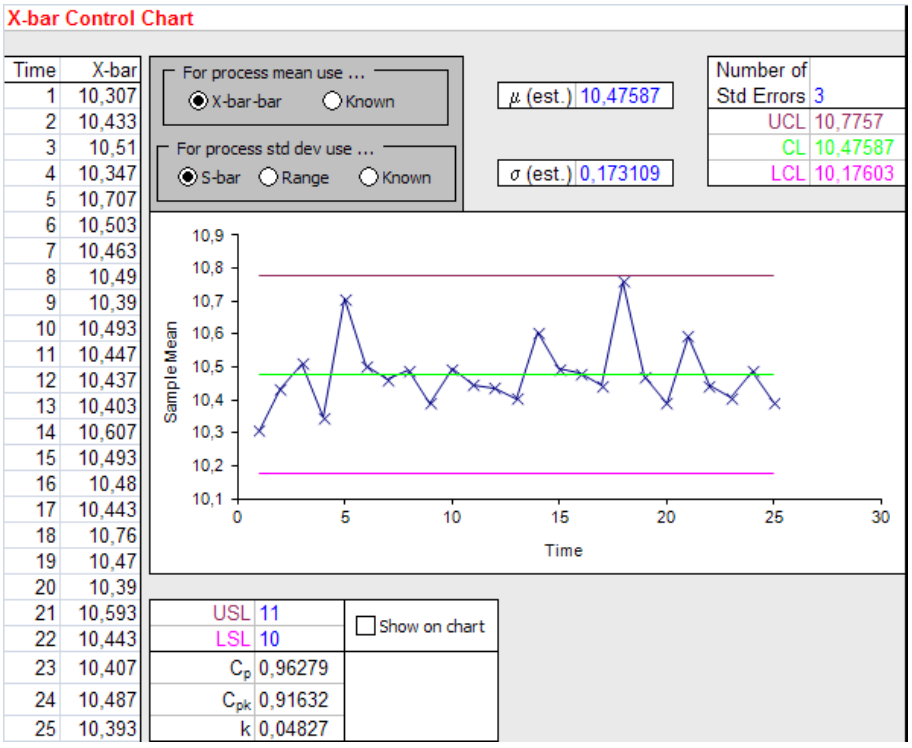
Nyní tento typ diagramu vykreslíme v programu XLStatistics na konkrétním příkladu. Na obrázku 3 máme vstupní údaje číselného typu.

Na následujícím listu sešitu obdržíme regulační diagram pro výběrové průměry $\overline{x}$ (obrázek 4).

Data

xX	Time	Obs1	Obs2	Obs3
	1	10,37	10,19	10,36
	2	10,48	10,24	10,58
	3	10,77	10,22	10,54
	4	10,47	10,26	10,31
	5	10,84	10,75	10,53
	6	10,48	10,53	10,5
	7	10,41	10,52	10,46
	8	10,4	10,38	10,69
	9	10,33	10,35	10,49
	10	10,73	10,45	10,3
	11	10,41	10,68	10,25
	12	10	10,6	10,71
	13	10,37	10,5	10,34
	14	10,47	10,6	10,75
	15	10,46	10,46	10,56
	16	10,44	10,68	10,32
	17	10,65	10,42	10,26
	18	10,73	10,72	10,83
	19	10,39	10,75	10,27
	20	10,59	10,23	10,35
	21	10,47	10,67	10,64
	22	10,4	10,55	10,38
	23	10,24	10,71	10,27
	24	10,37	10,69	10,4
	25	10,46	10,35	10,37

Obrázek 3: Data procesu



Obrázek 4: Regulační diagram pro výběrové průměry v XLStatistics

2.2. Regulační diagram (s)

Testovým kritériem v regulačním diagramu (s) je výběrová směrodatná odchylka s_j. Při $\alpha = 0{,}0027$ a neznámých cílových hodnotách μ_0 a σ_0 se stanoví CL pro tento regulační diagram dle vztahu $CL = \overline{s}$, kde $\overline{s}$ vypočteme podle vztahu

$$\overline{s} = \frac{\sum_{j=1}^{k} s_j}{k} \tag{7}$$

Při odvození vztahů pro stanovení akčních mezí v diagramu (s) vyjdeme ze vztahu pro odhad výběrové směrodatné odchylky $\widehat{\sigma}_s$:

$$\widehat{\sigma}_s = \frac{\overline{s}}{C_4} \cdot \sqrt{1 - {C_4}^2}. \tag{8}$$

Pro regulační meze pak dále platí:

$$LCL = \overline{s} \sqrt{\frac{\chi^2_{0,00135}(n-1)}{n-1}} \tag{9a}$$

$$UCL = \overline{s} \sqrt{\frac{\chi^2_{0,99865}(n-1)}{n-1}} \tag{9b}$$

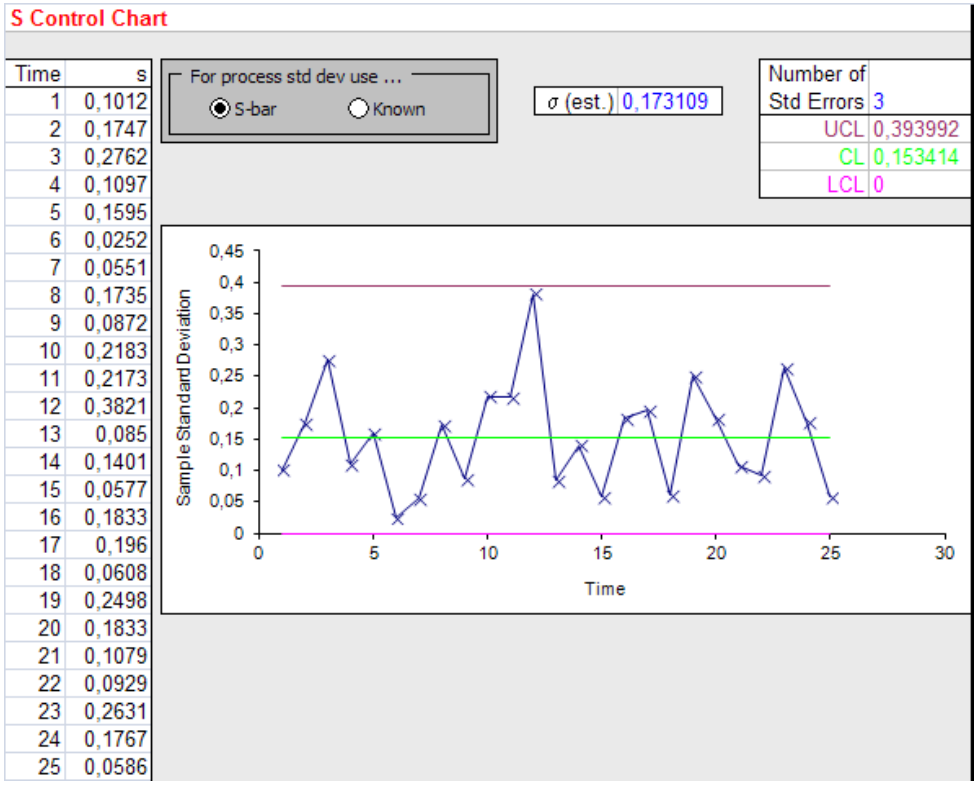
Při výpočtu regulačních mezí pro směrodatnou odchylku jsme využili kvantilů, které odpovídají pravidlu 3σ. Symbol $\chi^2_\alpha(v)$ označuje α-kvantil rozdělení chí-kvadrát s ν stupni volnosti.

Nyní opět tento typ diagramu vykreslíme v programu XLStatistics na konkrétním příkladu. Z dat předchozího příkladu vytvoříme na obrázku 5 pomocí programu XLStatistics regulační diagram (s) (S Control Chart).

2.3. Regulační diagram (R)

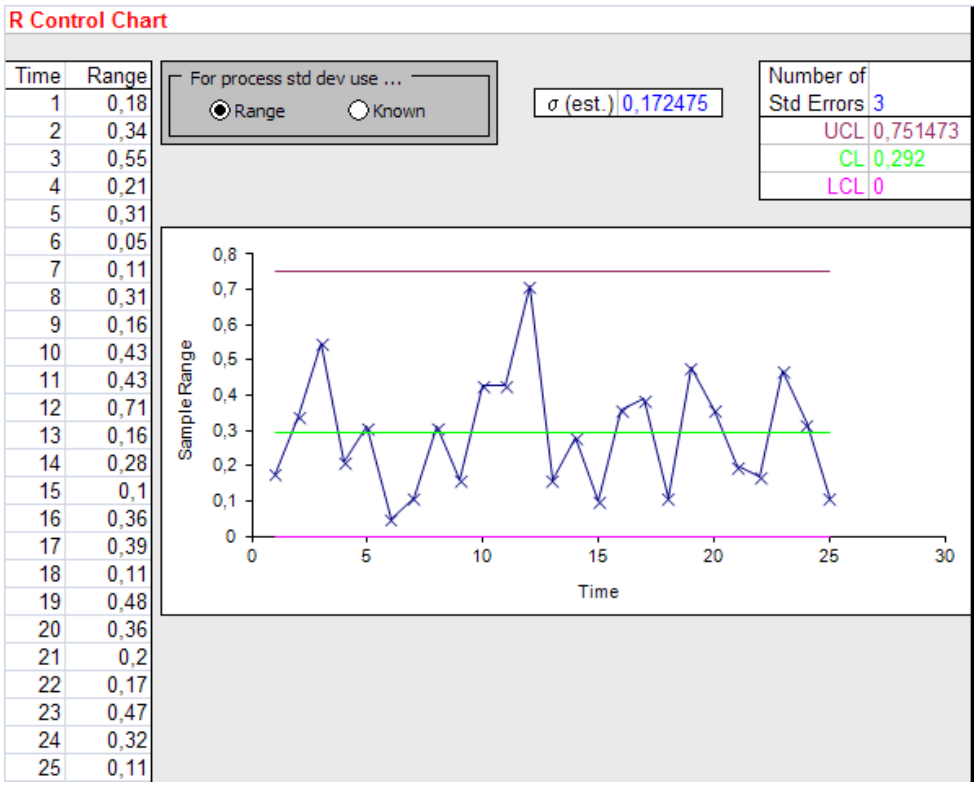
Testovým kritériem v regulačním diagramu (R) je výběrové rozpětí R_j. Jestliže $\alpha = 0{,}0027$ a nejsou neznámy cílové hodnoty μ_0 a σ_0, stanoví se CL pro tento regulační diagram ze vztahu $CL = \overline{R}$. Při odvození vztahů pro stanovení akčních regulačních mezí v diagramu (R) vyjdeme ze vztahu pro odhad směrodatné odchylky výběrového rozpětí $\widehat{\sigma}_R$:

$$\widehat{\sigma}_R = d_3 \cdot \frac{\overline{R}}{d_2} \tag{10}$$



Time	s
1	0,1012
2	0,1747
3	0,2762
4	0,1097
5	0,1595
6	0,0252
7	0,0551
8	0,1735
9	0,0872
10	0,2183
11	0,2173
12	0,3821
13	0,085
14	0,1401
15	0,0577
16	0,1833
17	0,196
18	0,0608
19	0,2498
20	0,1833
21	0,1079
22	0,0929
23	0,2631
24	0,1767
25	0,0586

Obrázek 5: Regulační diagram (s) v programu XLStatistics



Time	Range
1	0,18
2	0,34
3	0,55
4	0,21
5	0,31
6	0,05
7	0,11
8	0,31
9	0,16
10	0,43
11	0,43
12	0,71
13	0,16
14	0,28
15	0,1
16	0,36
17	0,39
18	0,11
19	0,48
20	0,36
21	0,2
22	0,17
23	0,47
24	0,32
25	0,11

Obrázek 6: Regulační diagram (R) v programu XLStatistics

kde d_3 je konstanta pro stanovení odhadu směrodatné odchylky výběrového rozpětí; její hodnota závisí na rozsahu výběru n a byla odvozena pro regulovanou veličinu pocházející z normálního rozdělení. Pro regulační meze pak dále podle [10] platí:

$$UCL = CL + u_{0,99865} \cdot \widehat{\sigma}_R = \overline{R} + 3 \cdot d_3 \cdot \frac{\overline{R}}{d_2} = \left(1 + \frac{3 \cdot d_3}{d_2}\right) \cdot \overline{R} \qquad (11a)$$

$$LCL = CL - u_{0,99865} \cdot \widehat{\sigma}_R = \overline{R} - 3 \cdot d_3 \cdot \frac{\overline{R}}{d_2} = \left(1 - \frac{3 \cdot d_3}{d_2}\right) \cdot \overline{R} \qquad (11b)$$

Obrázek 6 znázorňuje regulační diagram (R) v programu XLStatistics, který vychází z údajů na obrázku 3.

2.4. Regulační diagram pro jednotlivé hodnoty (i)

V případech, kdy z nějakého důvodu není účelné stanovování podskupin, lze použít Shewhartův diagram pro jednotlivé hodnoty, x-individual. Místo průměrů podskupin se pracuje přímo s naměřenými hodnotami x_i. Jako příslušný diagram pro variabilitu se používá diagram R. Místo rozpětí podskupiny se však použijí rozpětí mezi po sobě následujícími hodnotami. Tato hodnota se nazývá klouzavé rozpětí a označuje se MR (moving range), $MR_i = |x_i - x_{i-1}|$. První hodnota se nedefinuje. Pro základní linii a regulační meze diagramu x_i se používají následující vztahy:

$$UCL = \overline{x} + 3 \cdot \frac{\overline{MR}}{d_2}, \qquad (12)$$

$$CL = \overline{x}, \qquad (13)$$

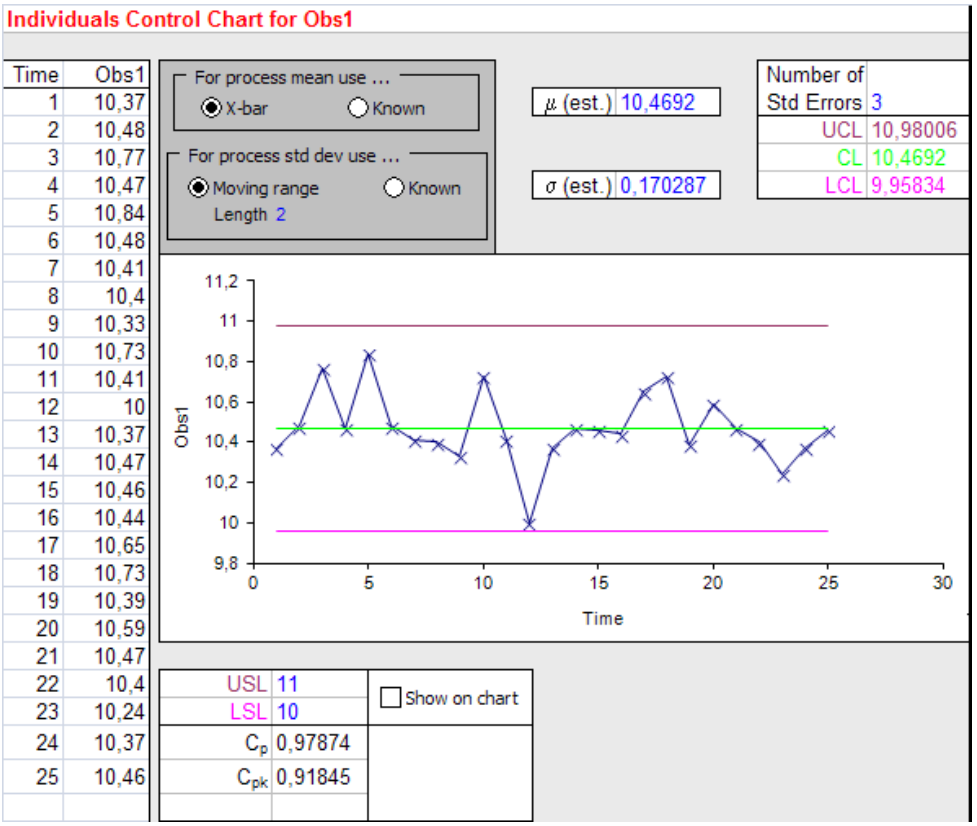
$$LCL = \overline{x} - 3 \cdot \frac{\overline{MR}}{d_2}. \qquad (14)$$

Statistické vlastnosti klouzavého rozpětí jsou stejné jako u rozpětí podskupiny pro $n = 2$. Koeficient d_2 má hodnotu 1,128. [4], [10]

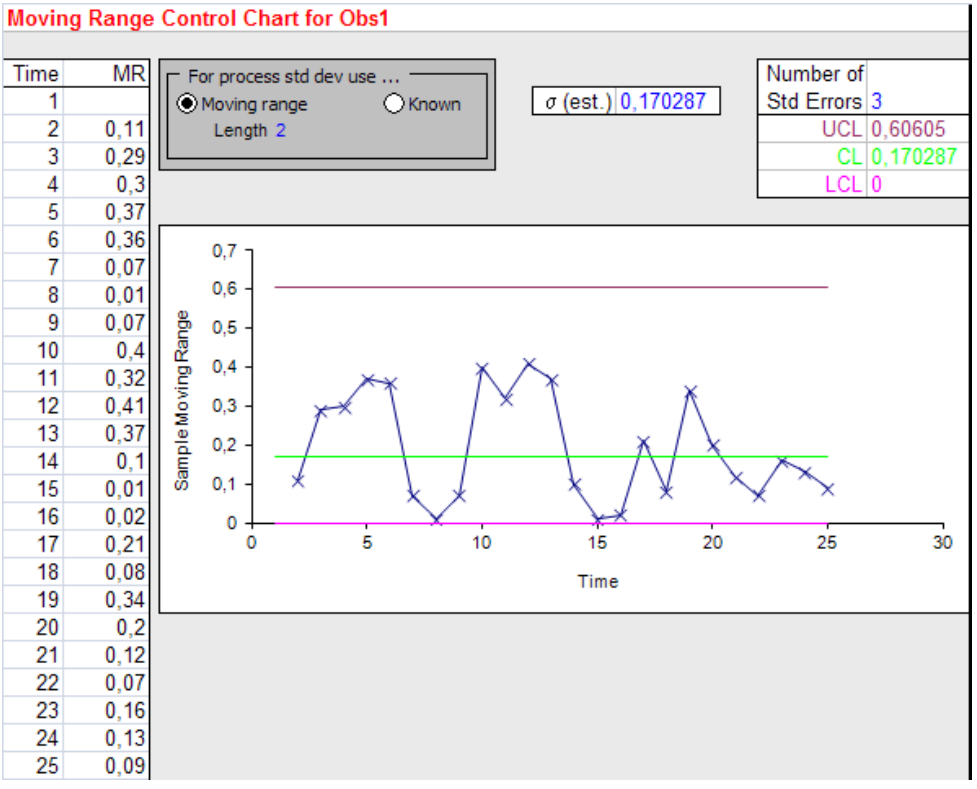
Na obrázku 7 vidíme regulační diagram pro individuální hodnoty (i) v programu XLStatistics, který vychází stejně jako předchozí z údajů na obrázku 3 pro sloupec Obs1.

2.5. Regulační diagram pro klouzavé rozpětí (MR)

V tomto diagramu se zakreslují hodnoty klouzavého rozpětí (MR). Vyjdemeli ze stejných předpokladů jako u předchozího diagramu, stanovíme CL, LCL,



Obrázek 7: Regulační diagram pro individuální hodnoty v XLStatistics



Obrázek 8: Regulační diagram pro klouzavé rozpětí MR v XLStatistics

UCL následovně:

$$UCL = D_4 \cdot \overline{R}_{kl}, \tag{15}$$

$$CL = \overline{R}_{kl}, \tag{16}$$

$$LCL = D_3 \cdot \overline{R}_{kl}. \tag{17}$$

Konstanty D_4 a D_3 jsou stejné součinitele jako pro stanovení regulačních mezí pro výběrové rozpětí ve dvojici $(\overline{x}, R)$. Jsou zde ale stanoveny vzhledem k rozsahu výběru $n = 2$ (viz ČSN ISO 8258). [4]

Nyní následuje praktický příklad. Na obrázku 8 je vypočítán regulační diagram pro klouzavé rozpětí v programu XLStatistics, který vychází stejně jako předchozí z údajů na obrázku 3 pro sloupec Obs1.

3. Regulační diagramy pro atributy

Je-li sledovaným parametrem diskrétní veličina (atribut) jako počty vad, používají se regulační diagramy pro atributy, nazývané také regulační diagramy srovnáváním. Protože však rozdělení počtu není normální, používají se pro výpočet regulačních mezí jiné vztahy, odpovídající příslušným kvantilům binomického nebo Poissonova rozdělení. Binomické rozdělení mají např. počty vadných součástek, jejichž počet je omezen celkovým počtem. V takovém případě se používají diagramy np a p. Poissonovo rozdělení mají počty, které nejsou omezené pevnou hodnotou, např. počet škrábanců na lakovaném povrchu. Zde se používají diagramy c a u. [9]

3.1. Regulační diagram np

Diagram np je vhodný pro sledování počtu vadných výrobků (jednotek) z nějakých dávek, který má binomické rozdělení. Je to tedy diagram pro diskrétní, celočíselné hodnoty neboli diagram srovnáváním. Šíře kontrolních mezí závisí na velikosti dávky.

Jednotlivé dávky představují podskupinu, počet vadných (neshodných, nevyhovujících) výrobků np z dané dávky je hodnota, která se vynáší do diagramu. [10]

3.2. Regulační diagram p

Diagram p je vhodný pro sledování podílu vadných výrobků (jednotek) z nějakých dávek. Šíře kontrolních mezí závisí na velikosti dávky.

Zvolíme-li oboustrannou regulaci, $\alpha = 0{,}0027$ a hodnotu p_0 musíme odhadnout, pak lze určit CL, LCL a UCL v diagramu (p) pomocí běžně používaných vztahů uváděných také v normě ČSN ISO 8258 [4]:

$$\overline{n} = \frac{\sum_{j=1}^{k} n_j}{k}, \tag{18}$$

kde k je počet výběrů nebo kontrolovaných objektů.

$$CL = \widehat{p}_0 = \overline{p} = \frac{\sum_{j=1}^{k} x_j}{\sum_{j=1}^{k} n_j} \tag{19}$$

kde x_j je počet neshodných jednotek v j-tém výběru, n_j je rozsah j-tého výběru, k je počet výběrů. Pak

$$UCL = \overline{p} + 3\sqrt{\overline{p}(1-\overline{p})/\overline{n}}, \tag{20a}$$

$$LCL = \overline{p} - 3\sqrt{\overline{p}(1-\overline{p})/\overline{n}}. \tag{20b}$$

Podle předchozích vztahů se určují regulační meze v diagramu (p) v případě, že pro n_j platí:

$$n_j \in \langle \overline{n} - 0{,}25\overline{n};\ \overline{n} + 0{,}25\overline{n} \rangle \tag{21}$$

Tento vztah však program XLStatistics nepoužívá.

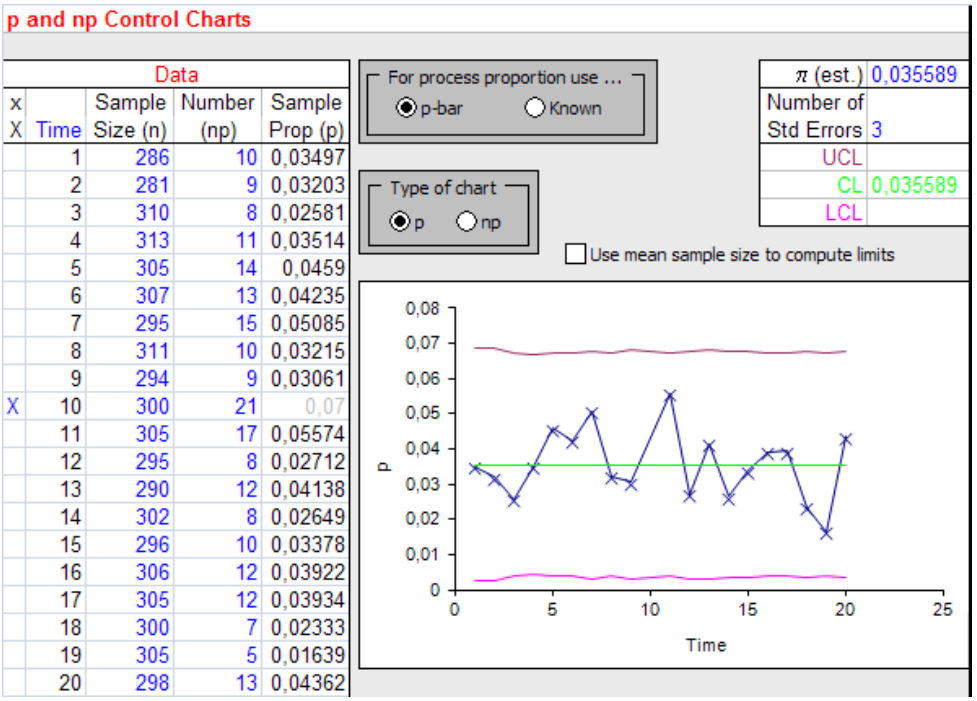
Pokud pro některý výběr tento vztah neplatí, určují se takzvané individuální meze pro j-tý výběr dle vztahů:

$$UCL_j = \overline{p} + 3\sqrt{\overline{p}(1-\overline{p})/n_j}, \tag{22a}$$

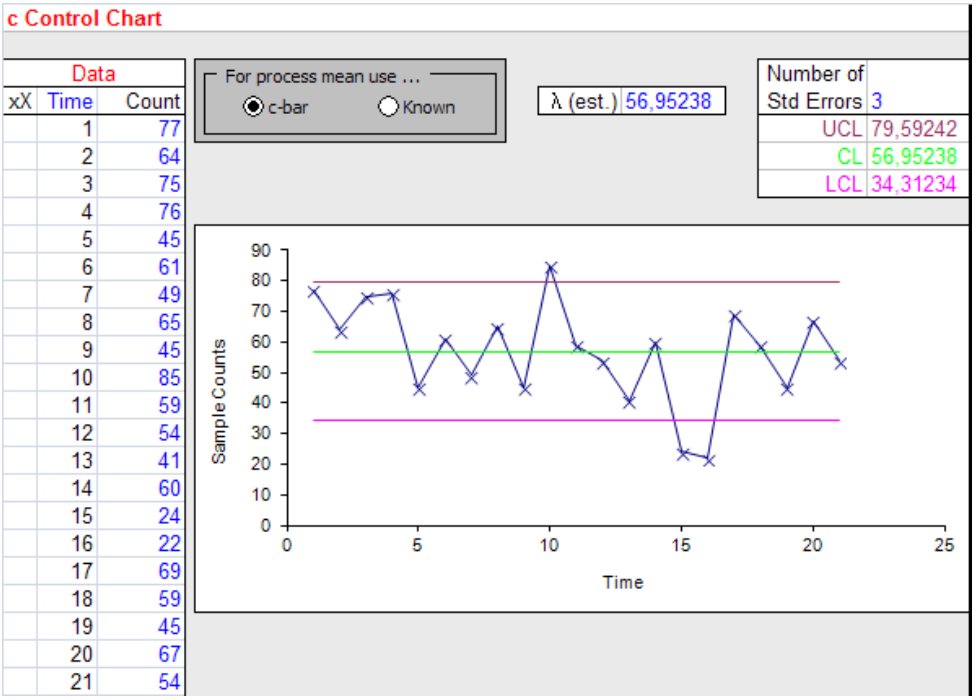
$$LCL_j = \overline{p} - 3\sqrt{\overline{p}(1-\overline{p})/n_j}. \tag{22b}$$

Jednotlivé dávky představují podskupinu, podíl vadných (neshodných, nevyhovujících) výrobků p z dané dávky je hodnota, která se vynáší do diagramu. [10]

Praktický příklad tohoto regulačního diagramu následuje na obrázku 9. V jeho levé části jsou zadána vstupní data a v pravé je vykreslen regulační diagram typu p, který jde změnit na typ np pomocí přepínače Type of chart nad regulačním diagramem. Dle volby uživatele lze pomocí symbolu X vyřadit některé skupiny pozorování (zde byl zvolen řádek 10).



Obrázek 9: Regulační diagram p v programu XLStatistics



Obrázek 10: Regulační diagram pro počet neshod c v programu XLStatistics

3.3. Regulační diagram pro počet neshod c

Tento regulační diagram, který se stručně nazývá regulační diagram c, se používá v těchto případech:

- Kontroluje se počet neshod na vybraných logických podskupinách o stejném počtu n produktů, kde $n \geq 1$ (např. počet bublin na tabuli skla).
- Regulovanou veličinou, označenou c_i, je počet neshod v i-té skupině, kde $i = 1, 2, \ldots, k$.

Počet neshod je tedy diskrétní náhodnou veličinou, a pokud jsou splněny následující předpoklady:

- počet neshod na produktu může být teoreticky neohraničený,
- pravděpodobnost, že v určitém místě na produktu bude více, než jedna neshoda je zanedbatelná,
- střední počet neshod na jednotce produktu je roven číslu λ_0,

lze předpokládat, že tato náhodná veličina má Poissonovo rozdělení se střední hodnotou $n\lambda_0$. Pro sestrojení diagramu pro počet neshod c se předpokládá, že Poissonovo rozdělení počtu neshod lze aproximovat normálním rozdělením, pokud $n\lambda_0 \geq 5$. Pak při zvoleném riziku zbytečného signálu $\alpha = 0{,}0027$ se určí střední přímka a akční meze tohoto regulačního diagramu, označené $CL(c)$, $UCL(c)$ a $LCL(c)$, pomocí vzorců

$$\begin{aligned} UCL\,(c) &\approx n\lambda_0 + u_{0,99865}\sqrt{n\lambda_0}, \\ CL\,(c) &= n\lambda_0, \\ LCL\,(c) &\approx n\lambda_0 - u_{0,99865}\sqrt{n\lambda_0}, \end{aligned} \tag{23}$$

kde $u_{0,99865}$ je 99,865 % kvantil normovaného normálního rozdělení. Číslo $n\lambda_0$ se odhadne pomocí průměrného počtu neshod, označeného $\overline{c}$, podle vzorce

$$\overline{c} = \frac{1}{k}\sum_{i=1}^{k} c_i \tag{24}$$

kde jednotlivé symboly ve vzorci značí:

c_i je počet neshod v i-té logické podskupině,

k je počet logických podskupin, přičemž se doporučuje, aby počet logických podskupin byl dvacet až dvacet pět.

Dosadíme-li získaný odhad pro $n\lambda_0$ do předchozích vztahů, je střední přímka pro počet neshod c dána vzorcem

$$CL\,(c) = \overline{c}. \tag{25}$$

Akční meze regulačního diagramu pro počet neshod c jsou dány vzorci

$$UCL\,(c) = \overline{c} + 3\sqrt{\overline{c}}; \tag{26a}$$

$$LCL\,(c) = \overline{c} - 3\sqrt{\overline{c}}. \tag{26b}$$

Pokud je ale $5 \leq n\lambda_0 \leq 9$, vychází dolní akční mez $LCL(c)$ záporná, což nemá reálný smysl. V tomto případě se buď pokládá tato mez rovna nule, nebo se zvýší rozsah výběru n tak, aby $n\lambda_0$ bylo větší než 9. V následujícím příkladu ukážeme sestrojení regulačního diagramu c. [12]

Na obrázku 10 vidíme regulační diagram pro počet neshod c v programu XLStatistics. Vstupní data se zadávají v levé části (Data).

3.4. Regulační diagram u

Regulační diagram pro počet neshod na jednotku, označený jako regulační diagram u, je odvozen od regulačního diagramu pro počet neshod c, který jsme popsali v předchozím oddílu. Používá se v těchto případech:

- Rozsah logických podskupin není konstantní, takže se sleduje průměrný počet neshod na jeden produkt z výběru.
- Sleduje se počet neshod na jednotlivých produktech o nestejné velikosti, přičemž se určuje počet neshod na jednotku rozměru produktu.

Testovým kritériem v tomto diagramu je průměrný počet neshod na jednotku produktu, který se pro i-tý výběr označuje u_i, kde $i = 1, 2, \ldots, k$.

Pro regulační diagram u, za předpokladu, že je splněna podmínka pro aproximaci Poissonova rozdělení normálním rozdělením, tj. když $n\lambda_0 \geq 5$, se při zvoleném riziku $\alpha = 0{,}0027$ určí střední přímka $CL(u)$ pomocí vzorce

$$CL\,(u) = \overline{u} = \frac{\sum_{i=1}^{k} c_i}{\sum_{i=1}^{k} n_i}, \tag{27}$$

kde jednotlivé symboly značí:

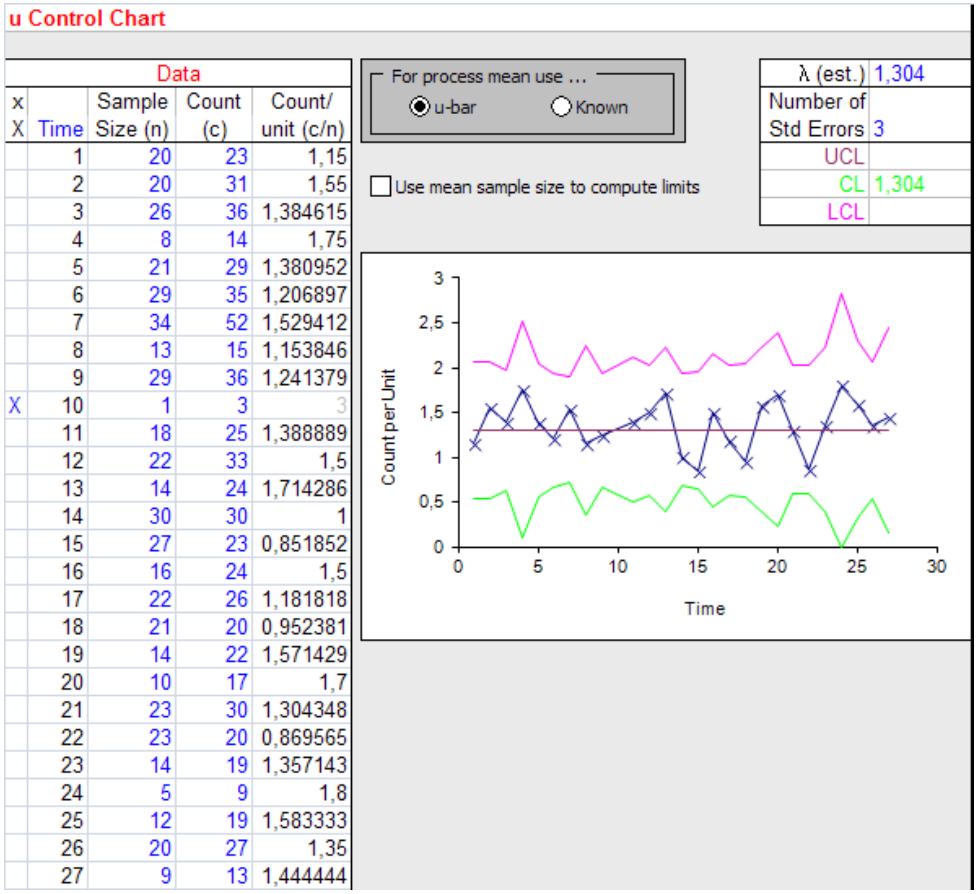
c_i je počet neshod v i-té logické podskupině, resp. na i-tém produktu,

n_i je rozsah i-tého výběru, resp. i-tého produktu.

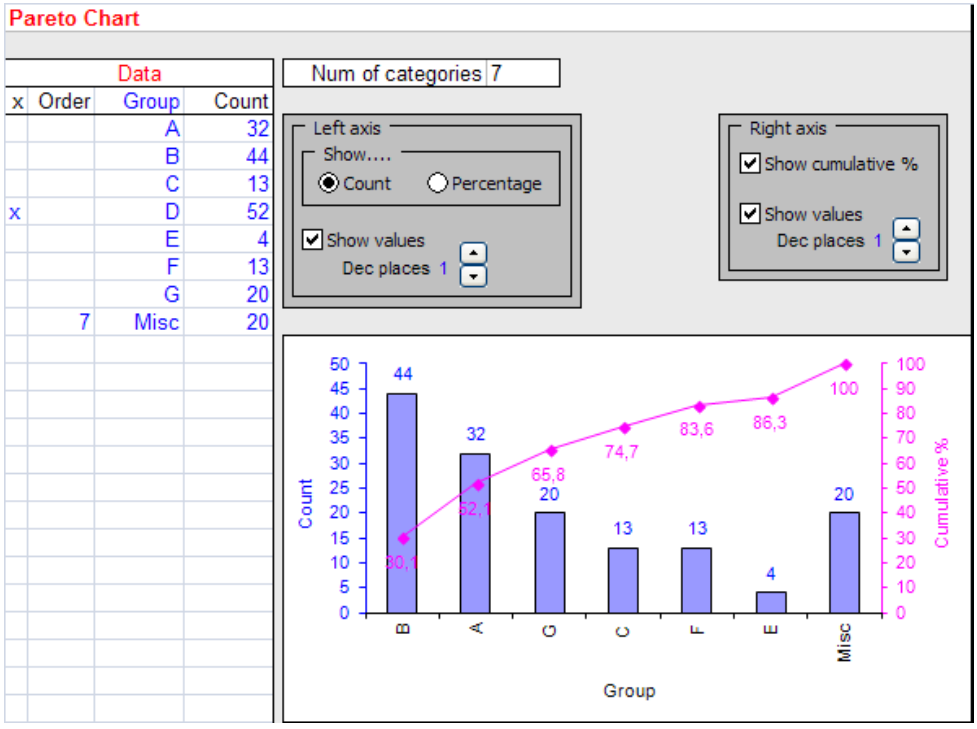
Akční meze tohoto diagramu, označené $UCL(u)$ a $LCL(u)$ jsou rovny

$$UCL\,(u) = \overline{u} + 3 \cdot \sqrt{\overline{u}\,/\overline{n}}; \tag{28a}$$

$$LCL\,(u) = \overline{u} - 3 \cdot \sqrt{\overline{u}\,/\overline{n}}, \tag{28b}$$



Obrázek 11: Regulační diagram u v programu XLStatistics



Obrázek 12: Paretova analýza v programu XLStatistics

kde $\overline{n}$ je průměrný rozsah logické podskupiny, resp. průměrný počet měrných jednotek na jeden produkt. Hodnotu $\overline{n}$ určíme pomocí vzorce

$$\overline{n} = \frac{1}{k} \sum_{i=1}^{k} n_i. \tag{29}$$

Zde k je počet kontrolovaných objektů. Akční meze se nazývají průměrnými mezemi. Používají se pro jednotlivé logické podskupiny v těch případech, pokud se počty n_i prvků v nich neliší od hodnoty $\overline{n}$ více než 25 %, tj. když pro n_i platí

$$n_i \in \langle 0{,}75\overline{n};\ 1{,}25\overline{n} \rangle\,. \tag{30}$$

Tento vztah však program XLStatistics nepoužívá.

Pokud pro i-tou logickou podskupinu, resp. i-tý produkt tento vztah neplatí, určují se pro tyto případy tzv. individuální meze podle předpisu

$$UCL(u)_i = \overline{u} + 3 \cdot \sqrt{\overline{u}\,/n_i} \tag{31a}$$

$$LCL(u)_i = \overline{u} - 3 \cdot \sqrt{\overline{u}\,/n_i} \tag{31b}$$

Regulační meze pro diagram u nemusí být, na rozdíl od diagramu pro c, konstantní. [11]

Opět následuje praktická ukázka tohoto regulačního diagramu v programu XLStatistics. Obrázek 11 nám ukazuje regulační diagram u spolu se vstupními daty v jeho levé části. Dle volby uživatele lze pomocí symbolu X vyřadit některé skupiny pozorování (zde řádek 10).

Kromě výše uvedených regulačních diagramů pro regulaci měřením a srovnáváním lze v programu XLStatistics provádět také Paretovu analýzu na posledním listu **Pareto,** viz obrázek 12 (Pareto Chart). Hlavním nástrojem této analýzy je Paretův diagram. Je základním nástrojem statistického řízení kvality. Je typem grafu, který je kombinací sloupcového a spojnicového grafu. Sloupce znázorňující četnost pro jednotlivé kategorie jsou seřazeny podle velikosti (nejvyšší sloupec vlevo, nejnižší vpravo) a spojnice představuje kumulativní četnost. To znamená, že spojnice začíná na prvním sloupci a každý další její bod je zvýšen oproti předchozí hodnotě o hodnotu odpovídající kategorii. Tak spojnice ukazuje kumulaci hodnot aktuální kategorie a kategorií, které jsou vlevo od ní. Kumulativní četnost bývá vyjádřena v procentech. Hodnoty procent jsou potom druhou stupnicí na vertikální ose grafu. XLStatistics umožňuje vyřazení kategorie (skupiny) dle volby uživatele pomocí symbolu X. V našem příkladu se jedná o skupinu D.

Závěr

XLStatistics je uživateli oblíben pro svoji jednoduchost instalace, přehlednost a snadné ovládání. Program je poskytován pro zkušební a výukové účely zdarma. Pro uživatele je nespornou výhodou, že program pracuje v prostředí Microsoft Excelu, které je všeobecně známé a dostupné téměř na všech počítačích. Pro potřeby řízení kvality program poskytuje nabídku všech základních klasických regulačních diagramů jak pro regulaci měřením, tak pro regulaci srovnáváním. V nabídce je navíc i Paretova analýza. Samozřejmě pro potřebu pokročilých nástrojů řízení kvality (například diagramy CUSUM, EWMA, Hotellingovy diagramy a další) je možné použít například komerční softwarový nástroj (například Minitab nebo český QC Expert) nebo statistické výpočetní prostředí R.

Práce s programem XLStatistics je velice snadná a intuitivní. Hodnoty v buňkách, které jsou modře, lze dle potřeby uživatele měnit, dvojím poklepáním na červené buňky pak získáme příslušnou nápovědu. I rozsah dalších nabízených statistických procedur programu (viz obrázek 1) bohatě pokrývá potřebu statistických analýz většiny běžných uživatelů a také postačuje jako software pro výuku standardních kurzů statistiky na vysokých školách.

Literatura

[1] Carr R.: *XLMathematics.* Excel workbooks for Mathematical Analysis. Version 2. `http://www.deakin.edu.au/~rodneyc/XLMathematics/`

[2] Carr R.: *XLStatistics.* Excel workbooks for Data Analysis. Version 6. Manuál programu.
`http://www.deakin.edu.au/~rodneyc/XLStatistics/`

[3] Klímek P.: XLStatistics pro výuku statistiky na FaME, UTB ve Zlíně. *Informační bulletin České statistické společnosti,* 18(3): 15–27, 2007. ISSN 1210-8022. doi: 10.5300/IB

[4] ČSN ISO 8258: *Shewhartovy regulační diagramy.* Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 1994.

[5] Kovářík M.: Volatility Change Point Detection Using Stochastic Differential Equations and Time Series Control Charts. *International journal of mathematical models and methods in applied sciences,* 2(7): 121–132, 2013. `http://www.naun.org/main/NAUN/ijmmas/16-661.pdf`

[6] Kovářík M.: Vícerozměrné statistické řízení procesů. *Informační bulletin České statistické společnosti,* 23(3): 31–50, 2012. ISSN 1210-8022. doi: 10.5300/IB

[7] Kovářík M.: *Využití regulačních diagramů a stochastických diferenciálních rovnic pro detekci bodu změny ve volatilitě časových řad.* Žilina: Georg, 2012. ISBN 978-80-89401-61-1.

[8] Kovářík M., Klímek P.: *Využití matematicko-statistických metod v řízení kvality.* Žilina: GEORG, 2011. ISBN 978-80-89401-54-3.

[9] Kupka K.: *Statistické řízení jakosti.* Pardubice: TriloByte, 2001. ISBN 80-238-1818-X.

[10] Tošenovský J., Noskievičová D.: *Statistické metody pro zlepšování jakosti.* Ostrava: Montanex, a. s., 2000. ISBN 80-7225-040-X.

[11] Václavek J.: *Statistická regulace výrobních procesů.* České Budějovice: Bartoň QSV, 1996. 174 s. ISBN 80-902236-0-5.

[12] Wheeler J.: *Advanced Topics in Statistical Process Control: The Power of Shewhart's Charts.* 2nd edition. USA: SPC Press, Inc., 2004. ISBN 978-0-945-32063-0.

K PĚTASEDMDESÁTINÁM PROFESORA JIŘÍHO ANDĚLA

CONGRATULATIONS ON PROFESSOR JIŘÍ ANDĚL'S 75TH BIRTHDAY

Redakce časopisu

Dlouholetý vedoucí katedry pravděpodobnosti a matematické statistiky a donedávna proděkan Matematicko-fyzikální fakulty a jedna z nejvýraznějších postav české matematické statistiky prof. RNDr. Jiří Anděl, DrSc., oslavil své pětasedmdesátiny (7. 3. 1939).

Vzhledem k zájmu o matematiku se přihlásil ke studiu na MFF UK, kde studoval v letech 1956–1961. Již během studia si ho prof. Janko vybral jako asistenta na katedře matematické statistiky. Vědeckou přípravu na téže katedře absolvoval pod vedením prof. Hájka a řadí se tak ke známým žákům a pokračovatelům v Hájkově díle. Kandidátskou disertační práci *Lokální asymptotická mohutnost testů typu Kolmogorova–Smirnova* obhájil v roce 1965 (výsledky této práce byly v roce 1967 publikovány v prestižních Annals of Mathematical Statistics). Na docenta MFF UK se habilitoval v roce 1972 na základě habilitační práce *Mnohorozměrné autoregresní posloupnosti*. Jmenován docentem byl v roce 1977. Od téhož roku byl pověřeným vedoucím a od roku 1981 pak řádným vedoucím katedry pravděpodobnosti a matematické statistiky (za velmi úspěšné působení v této funkci mu MFF UK udělila v roce 1978 medaili 2. stupně a v roce 1982 medaili 1. stupně). Po vypracování a obhájení doktorské disertace na téma *Některé míry závislosti v časových řadách* mu byla udělena v roce 1981 vědecká hodnost DrSc. V roce 1986 byl jmenován vysokoškolským profesorem. V letech 1993–1996 působil prof. Anděl jako proděkan pro matematiku a od roku 1996 až do září 2012 jako pedagogický proděkan.

Profesor Anděl odborně pracuje především v oblasti matematické statistiky a časových řad. O rozsahu a úspěšnosti jeho odborné činnosti svědčí mimo jiné dosud publikovaných 92 vědeckých prací (často ve velmi prestižních odborných časopisech), 6 knih, 4 skripta, 28 aplikačních prací, 55 popularizačních prací, 26 výzkumných zpráv a velké množství zahraničních citací. Z jeho

citovaných výsledků lze uvést práce týkající se různých typů časových řad: autoregresních, mnohorozměrných, nelineárních, nezáporných, inverzních, s náhodnými či periodickými parametry, s dlouhou pamětí aj. V oblasti časových řad jsou dále citovány práce prof. Anděla věnované interpolování a extrapolování (predikcím), závislosti mezi časovými řadami, řadám s daným marginálním rozdělením či danými momenty, speciálním (např. bayesovským) odhadovým procedurám, spektrálním vlastnostem a další problematice. Za soubor prací *Statistické modely časových řad a jejich simulace* mu byla v roce 1990 udělena Národní cena ČR.

Vedle teoretického výzkumu se prof. Anděl také významně věnoval činnosti aplikační (27 aplikačních prací), která byla mimo jiné motivována spoluprací s některými praktickými institucemi z oblasti zdravotnictví (např. Institut hygieny a epidemiologie v Praze), průmyslu (např. Škoda Plzeň) nebo hydrologie (Vodohospodářský ústav). Podílel se tak na řešení praktických problémů typu periodicity v průtocích vodních toků či analýzy biosignálů EEG a dalších. Praktické výsledky, které pracovníkům z praxe při řešení konkrétních problémů předkládal, jsou velmi úspěšnými a přesvědčivými argumenty o užitečnosti matematické statistiky pro praxi.

Zvláštní pozornost si zasluhují knihy, které prof. Anděl napsal. Monografie *Statistická analýza časových řad* (SNTL 1976) je dodnes používána jako referenční materiál v pracích věnovaných teorii či aplikacím časových řad (to samé platí pro její německý překlad z roku 1984 v německy mluvících zemích). Jeho nejznámější knihou je ovšem *Matematická statistika* (SNTL/ALFA 1978, druhé vydání 1985): pokud měl u nás kdokoli co do činění s (matematickou) statistikou, určitě se setkal s touto vynikající publikací nebo dostal radu, aby se podíval do „modré knihy“. Také díky ní a knihám, které na tuto základní publikaci později navázaly, patří u nás prof. Anděl k nejznámějším statistikům a jeho názor má velkou váhu (např. ve společensky závažných či méně závažných situacích, kdy se uplatňuje pravděpodobnost včetně různých televizních pořadů).

Profesor Anděl byl v roce 1990 zakládajícím předsedou České statistické společnosti, v jejímž čele stál do roku 1993.

Hlavní smysl své práce prof. Anděl vždy spatřoval v činnosti pedagogické. Nejenže se na své přednášky pečlivě připravuje, ale vlastní dar vyložit i velmi složité partie názornou a snadno pochopitelnou formou. I v nejstresovějších situacích má u něj výuka vždy přednost. V anketním hodnocení studentů získává tradičně vysoký počet bodů a říká se, že právě díky jeho úvodním přednáškám se relativně velký počet posluchačů hlásí na statistický obor.

Redakce časopisu

ROZHOVOR S JIŘÍM ANDĚLEM
INTERVIEW WITH PROFESSOR JIŘÍ ANDĚL

Profesores anonymi

• *Kteří lidé nejvíce ovlivnili váš pohled na statistiku a pravděpodobnost?*

První dva roky studia na MFF byly společné nejen pro matematické obory, ale i pro matematiku a fyziku. Jednotlivé specializace tedy začínaly ve třetím ročníku. Matematickou statistiku přednášel MgMat. Marcel Josífko. Výklad založil převážně na Cramérově učebnici, což byl tehdy moderní počin. Navíc pan magistr Josífko měl hodně zkušeností s praktickými aplikacemi a tím dokázal výrazně oživit přednášenou látku. Dalším výrazným učitelem byl Ing. Josef Machek. U něho jsem ostatně psal i svou diplomovou práci. Pan inženýr měl hluboké znalosti statistických metod a některé komplikované výpočty dokázal provést rychleji než osobní počítače. To už je ale zkušenost z pozdější doby. Teorii odhadu a testování hypotéz přednášel prof. Josef Bílý, mimořádně vzdělaný a zdvořilý člověk. Můj pohled na statistiku a pravděpodobnost samozřejmě nejvíc ovlivnil prof. Jaroslav Hájek. Měl jsem to štěstí, že jsem byl jeho prvním aspirantem. Znalosti, které jsem pod jeho vedením získával, byly neocenitelným přínosem pro mou další pedagogickou i vědeckou práci.

• *Často býváte spojován s legendárním prof. Hájkem, byl jste jeho asistent, když vyučoval v té době velmi moderní kurz matematické statistiky.*

MFF jsem absolvoval v roce 1961. Tehdy byl vedoucím katedry statistiky prof. Janko a tajemníkem katedry Ing. Machek. V rámci tzv. vědecké přípravy, což byla jedna z forem aspirantury, se mě ujal jako školitel prof. Dupač. Rovnou prohlásil, že mě bude školit jen dva roky, a to během přípravy na odbornou kandidátskou zkoušku. To se už počítalo s tím, že na katedru přejde z Matematického ústavu ČSAV prof. Hájek, který měl prof. Dupače ve funkci mého školitele nahradit. Prof. Hájek se také měl stát vedoucím katedry po prof. Janko, což se také realizovalo. Prof. Hájek ihned převzal základní přednášku z matematické statistiky a ze stacionárních procesů. Přednášku ze statistiky založil na teorii pseudoinverzních matic, což byl tehdy velice moderní postup. Já a Karel Zvára jsme vedli cvičení k této přednášce. Samozřejmě jsme byli nejpilnějšími návštěvníky přednášek, abychom věděli, co vlastně máme cvičit. Ještě bych měl uvést jednu historku z doby, kdy prof. Hájek přebíral funkci vedoucího katedry. Ing. Machek využil příležitosti a po mnoha letech tajemníkování se této nepopulární funkce vzdal. Nově

nastupujícímu šéfovi bylo sděleno, že tajemníkem katedry bude Anděl. Při první schůzi katedry prof. Hájek hned na začátku velmi rozčileně prohlásil, že je zdvořilé a jedině možné, aby si vedoucí vybral svého tajemníka sám. On že s dosavadním postupem naprosto nesouhlasí. Viděl jsem, že to není dobrý začátek mé spolupráce s prof. Hájkem, ale on stále zvýšeným hlasem pokračoval. Sdělil, že se rozhodl, že tajemníkem bude Anděl, ale bude to z jeho rozhodnutí a ne z rozhodnutí předchozího vedení katedry. Přednášku ze statistiky pak každým rokem podstatně obměňoval. Snad jen my asistenti jsme věděli, že se vykládá stále stejná látka. Pokud nějaký student propadl a musel tuto přednášku navštěvovat znovu, musel mít pocit, že chodí na úplně jiný předmět. Ještě náročnější bylo jeho pojetí předmětu Stacionární procesy. Velkou část látky si připravoval sám na základě svých vlastních výsledků. Jednou se stalo, že byl zrovna v nemocnici a já jsem ho měl v přednášení zastoupit. Navštívil jsem ho a on mi řekl: „Myslím, že platí zhruba následující tvrzení. Tak to ověř a zítra to studentům i s důkazem předneseš.“

• *Jak přistupoval k výuce?*

K výuce přistupoval velmi zodpovědně. O tom ostatně svědčí i různá skripta, která napsal se svými asistenty a kolegy. Občas studentům během přednášky uváděl vědecké problémy. Slíbil, že kdo mu první přinese správné řešení, dostane u zkoušky o jednu otázku méně a místo ní mu bude vyřešení problému klasifikováno známkou výborně. Zadávané problémy bývaly dost obtížné, a tak jen málokdy se stalo, že je některý posluchač vyřešil.

• *Založil vědeckou školu a vedl disertaci řadě současných významných odbornic a odborníků. Jaký byl jeho způsob vedení aspirantů? Využívají se některé jeho postupy i dosud? Které byste zavedl do současné výuky na MFF UK a výchovy doktorandů?*

Prof. Hájek své aspiranty školil v malých skupinkách. Všichni řešili cvičení uvedená v právě probírané knize. Podmínkou bylo, že každý musí pracovat zcela samostatně a řešení nesmí hledat ani v literatuře. Jednou mi prof. Hájek vyprávěl, jak pracoval ve své aspirantuře on. Když studoval nějakou knihu, ve které cvičení nebyla, četl text jen do místa, kde se objevila nějaká matematická věta. Její důkaz si zakryl papírem a vypracovával ho sám. Je jasné, že taková příprava k vědecké práci je velmi účinná, ale nedá se masově použít. Rovněž disertaci musel každý jeho aspirant vypracovat zcela samostatně. Prof. Hájek zadal téma a pak jen velmi kriticky sledoval, zda je postup řešení správný. Když my, jeho aspiranti, jsme se sami stali školiteli, snažili jsme se podobným způsobem školit své vlastní aspiranty. Po roce 1989 se však podstatně změnil vysokoškolský zákon a změnila se i výuka doktorandů. Byly

zavedeny předměty vyučované v doktorském studiu a to se tak poněkud přiblížilo magisterskému studiu. Řekl bych, že dnes školitelé více pomáhají svým doktorandům při psaní disertací. Je to dáno i podmínkami, které platí pro doktorské obhajoby. Má-li mít doktorand své výsledky v době obhajoby již publikované, nebo aspoň k publikaci přijaté, musí na disertaci začít pracovat co nejdříve. A to se bez účinné pomoci školitele neobejde.

• *Vzpomněl byste také na další kolegy a kolegyně?*

Někteří z nás, kteří jsme na katedře pravděpodobnosti MFF UK, jsme přímo aspiranty prof. Hájka, mladší kolegové jsou zas našimi doktorandy. Tak uvedu alespoň několik jmen v abecedním pořadí. Prof. Dupačová, prof. Hušková, prof. Jurečková a prof. Štěpán (který bohužel nedávno zemřel) byli přímými aspiranty prof. Hájka. Myslím, že jejich mezinárodní reputace je dobře známa a není třeba ji uvádět.

• *Jak se podle vás změnila statistika v posledních letech, a jak se měnila statistika během vašeho života?*

Krátce bych se vyjádřil jen k vývoji matematické statistiky. Podle mého názoru jsme svědky velkého rozvoje zejména abstraktních partií. Stačí porovnat obsah časopisu The Annals of Mathematical Statistics s jeho nástupci The Annals of Statistics a The Annals of Probability. Takových příkladů by se dalo uvést víc. Články popisující konkrétní statistické metody se spíše najdou v časopisech pro techniky, pro lékaře a podobně. Velké změny pochopitelně přinesl nástup počítačů.

• *Jak váš výzkum a statistiku vůbec ovlivnily počítače?*

Přínos počítačů by se dal rozdělit do dvou kategorií. Jednak jde o obecné využití výpočetní techniky počínaje elektronickou poštou, textovými editory, literaturou v elektronické podobě atd., jednak o využití v matematice a statistice. Uvedl bych malý příklad. V předpočítačové době na naší katedře pracovala jednak paní sekretářka, jednak paní výpočtářka. S kolegou Karlem Zvárou jsme dělali průzkum, jak souvisí úspěšnost studia na výsledcích přijímacích zkoušek. Použili jsme na to diskriminační analýzu, napsali vzorečky a požádali paní výpočtářku o provedení výpočtů. Pracovala asi tři dny, pak přinesla výsledky. Kontrola ukázala, že je někde numerická chyba. Tak paní výpočtářka začala znovu. Dospěla k jinému, ale také chybnému výsledku. Ani třetí pokus nepřinesl správný výsledek. Bylo zřejmé, že během dlouhého výpočtu výpočtář s velkou pravděpodobností někde udělá chybu. Když nastoupily počítače, vyměňovali si statistici informace o tom, který program

má jaký nedostatek. Dnes už, alespoň pokud jde o základní výpočty, jsou programy dost spolehlivě odladěné. Pokud jde o mne, k usnadnění výpočtů a k jejich kontrole používám programy Mathematica a Maxima, k numerickým statistickým výpočtům pak program R.

• *Jaký máte postoj k mnohdy neuváženému používání „hotového“ statistického softwaru bez dohledu statistika?*

Posluchačům tento problém demonstruji na následujícím příkladě. Provedu simulaci dat pro ilustraci lineární regrese. Vypočtou se odhady parametrů přímky a intervaly spolehlivosti. Všechno funguje tak, jak má. Odhady jsou rozumně blízko skutečným parametrům a intervaly spolehlivosti překrývají dané parametry. Pak se místo hodnot nezávisle proměnné vezmou jejich „pokažené“ hodnoty, jako když i ta nezávisle proměnná je měřena s nezanedbatelnou chybou. Znovu se provedou odhady metodou nejmenších čtverců. Nějaké výsledky vyjdou. Ale odhady jsou daleko od parametrů a intervaly spolehlivosti nemají nic společného s hodnotami parametrů. Teprve použití správné teorie místo metody nejmenších čtverců vede k očekávaným výsledkům. Počítač tedy většinou nějaký výsledek vydá. Ale pokud nejsou splněny základní předpoklady použité metody, je zpravidla výsledek chybný a matoucí.

• *V České republice jste považován za jednoho z největších odborníků na analýzu časových řad, dlouhá léta jste v této oblasti spolupracoval především s hydrology. Jak se z této pozice díváte na často diskutovaný fenomén „změna klimatu“? Mnozí publicisté a politici přitom velmi rádi odkazují právě na analýzu (nezřídka velmi krátkých) environmentálních časových řad, často bez naprostého porozumění meritu věci.*

Podle mého názoru se použité analýzy týkají observačních dat, nikoli dat experimentálních. To ostatně vyplývá ze samé podstaty problému. Proto je třeba na výsledky výpočtů nahlížet jen jako na určité statistické charakteristiky, které nemohou vypovídat nic o kauzální závislosti sledovaných jevů. V oblasti analýzy hydrologických dat byla k dispozici tvrdá kontrola. Pomocí vypočteného modelu byla provedena předpověď průtoků řek a pak se ukázalo, zda tato předpověď byla úspěšná či nikoli. A právě takováto kontrola závěrů, ke kterým dospívají zastánci změny klimatu, mi zatím schází.

• *Je o vás známo, že se již delší dobu zabýváte průzkumy veřejného mínění. Co vás k tomu přivedlo? Co si myslíte o profesionalitě těchto výzkumů u nás? Prakticky po každých volbách se většinou nestačíme divit, že to vlastně dopadlo úplně jinak, než jak bylo předpovězeno.*

Průzkumy veřejného mínění jsou zajímavé samy o sobě. Dají se také použít jako hezká ilustrace při výuce. Navíc to máme takříkajíc v rodině. Můj syn, který také vystudoval matematickou statistiku, pracuje u takové instituce. Jsem přesvědčen, že se používají metody na dobré úrovni. Ostatně společnosti, které takové průzkumy zadávají, by rychle od takové praxe ustoupily, kdyby jim to nepřinášelo správné informace. Záležitost předvolebních průzkumů je trochu komplikovanější. Je třeba si uvědomit, že mnozí voliči mění své názory na poslední chvíli, tedy v době, kdy už je na prezentaci průzkumů vyhlášeno moratorium. Bylo by třeba vzít v úvahu nejen výsledky dotazování, ale i analýzu trendů. A to by veřejnost obtížně přijímala. Přechod od úplného zjišťování k výběrovým šetřením je patrně nevyhnutelný. Došlo k tomu i u vyspělých států. Spíše stojíme před jiným problémem. Naše zákony v tomto směru nejsou dostatečně účinné a jejich vymáhání je mírně řečeno problematické. Stojí za to podívat se do historie, jakou důležitost statistickému zjišťování připisovali vládci. Tak třeba ve starém Římě bylo podání požadované informace povinné a muži, který by se ke sčítání nedostavil, hrozilo, že ještě ten den může být uvržen do otroctví. Nedočetl jsem se, zda se tato hrozba opravdu uskutečňovala a v jakém rozsahu, ale patrně byla dostatečně účinná.

• *Co si jako statistik myslíte o trendu nahrazovat výběrová zjišťování odhady využívajíc administrativní data a registry? Není to vlastně návrat k vyčerpávajícím statistických zjišťováním, kdy místo statistického šetření se data od statistické jednotky (respondenta) získají pomocí povinného hlášení, administrativního úkonu?*

Pokud jsou registry spolehlivé, není důvod proti takovému trendu něco namítat. Na druhé straně registry zachytí jen některá statistická data, takže výběrová zjišťování nepůjde ani v budoucnu zcela nahradit.

• *Jaká je role aplikací pro rozvoj matematické statistiky? Mohl byste zmínit nějakou netradiční aplikaci v oblasti, která nebývá nebo ve své době nebyla považována za vhodnou pro aplikaci statistiky (např. archeologie)? Jakou aplikaci, na které jste spolupracoval, považujete za nejpřínosnější?*

Aplikace jsou jedním (ale ne jediným) zdrojem podnětů pro další rozvoj matematické statistiky. Třeba prof. C. R. Rao uvádí ve své knize aplikace v archeologii, které ho nepochybně inspirovaly k dalším statistickým výzkumům. Nebo McNemarův test, který byl publikován v časopisu Psychometrika. Jsem přesvědčen, že jeho autor byl inspirován podněty z lékařského výzkumu. Pokud jde o mé vlastní aplikace, tak záleží na tom, co se rozumí slovem nejpřínosnější. Třeba výše zmíněné aplikace v hydrologii lze označit za přínosné

třeba proto, že některé byly publikovány v předních vědeckých časopisech. Na druhé straně spolupráce s podnikem Škoda v Plzni byla úspěšná, protože zadavateli ušetřila nemalé finanční prostředky. Přitom šlo o standardní aplikaci, o které se, pokud vím, nic nepublikovalo.

• *Na jedné straně se statistika každý rok rozrůstá o několik desítek metrů časopisů a knih; když se však člověk problémem trochu důkladněji zabývá, exaktní řešení s poctivým vyřešením všech praktických problémů nedohledá. Máte tu samou zkušenost?*

Rozhodně není snadné najít zrovna tu metodu, kterou bychom zrovna potřebovali. To platí pro teoretické výsledky zrovna tak jako pro ty aplikační. Nejrychlejší, nejpohodlnější a asi i nejspolehlivější je obrátit se na někoho, kdo to ví. Já jsem měl to štěstí, že jsem se mohl obracet na prof. Dupače, Ing. Machka a další. Tito moji učitelé a později kolegové velkoryse přerušili svou práci a snažili se mi pomoci.

• *Byl jste prvním předsedou České statistické společnosti. Po prvotním nadšení ze znovuobjeveného spolkového života zažíváme jeho pomalý úpadek. Jakou roli přisuzujete vědeckým společnostem v současné době „volného přístupu" k informacím, například přes internet.*

Role vědeckých společností je nemalá. Nejde jen o organizaci vědeckého života, ale i o rozvoj mezilidských vztahů a kontakty mezi kolegy. Podle mého názoru založením České statistické společnosti se podařilo vybudovat dobré vztahy mezi Českým statistickým úřadem, VŠE a MFF. I kdyby žádná další spolupráce neexistovala, již tento fakt stojí za všechnu tu nezbytnou organizační práci. Ostatně třeba takoví lékaři, právníci, lékárníci a další jsou hrdi na své stavovské a odborné společnosti a snaží se zvyšovat jejich prestiž.

• *Více než 50 let učíte a vaše hodiny jsou pověstné vysokou úrovní a přesností výkladu. Jsou dnešní studenti přicházející na vysoké školy připraveni z matematiky o tolik hůře než dříve, jak se často uvádí ve sdělovacích prostředcích?*

Myslím, že matematická připravenost uchazečů přijatých na MFF se v poslední době nijak podstatně nemění. Máme možnost to posoudit na základě testu ze středoškolské matematiky, který studenti 1. ročníku absolvují na úvodním soustředění na Albeři. Někdy se zdá, že výsledky jsou horší a horší, ale pak zas přijde obrat. Rozdíl však může být v pracovní morálce. My jsme tvrdě studovali od první přednášky. Dnešní studenti většinou začnou intenzivně pracovat až na své bakalářské nebo diplomové práci. Téma je obvykle zaujme a vidí, že jsou schopni odvodit nové a přitom zajímavé výsledky.

• *Dlouhá léta se zabýváte popularizací statistiky, napsal jste řadu čtivých článků v různých časopisech, dlouholetá popularizační práce vyvrcholila knihou Matematika náhody, která vyšla též anglicky. Jaké máte zkušenosti s prezentováním aplikací statistiky a pravděpodobnosti na zajímavé praktické problémy (volby, rekordy) laikům a zejména novinářům, rozhlasovým a televizním redaktorům? Můžete uvést nějakou zajímavou zkušenost?*

Zmíním se o jedné zkušenosti. Je to už řada let, kdy v televizi probíhala nějaká soutěž, jejíž vítěz si mohl vybrat jednu ze tří cihliček. Ačkoli vypadaly všechny stejně, jedna byla zlatá, druhá stříbrná a třetí bronzová. Zastavil se u mne v Karlíně pan Ulm. Sdělil mi, že už po nějakou dobu si nikdo nevybral zrovna zlatou cihličku a že jim rozhořčení diváci píší, že tam snad ani tu zlatou nedávají. Byl jsem pozván do vysílání, vysvětlil jsem výpočet příslušných pravděpodobností a jejich interpretaci. Všechno bylo v pořádku i proto, že při další hře už zlatá cihlička byla zas vytažena. Já jsem se ale pana Ulma zeptal, proč se obrátil s tím problémem zrovna na mě. Čekal jsem, že mi odpoví, že o mně ví z nějakých propagačních akcí, ale odpověď byla nečekaná. Jednou jel tramvají přes Malostranské náměstí a uviděl nápis Matematicko-fyzikální fakulta. Vystoupil a pana vrátného se zeptal, kdo by mu mohl něco říci o pravděpodobnosti. Pan vrátný se podíval do telefonního seznamu, vyhledal tam seznam osob na katedře statistiky a řekl mu, ať se obrátí na profesora Anděla. A bylo to. Věc má však ještě jeden aspekt. Zjistil jsem, že se nespokojení diváci začali obracet na Československou televizi ve chvíli, kdy poprvé pravděpodobnost nevybrání zlaté cihličky klesla pod naše magické číslo 0,05. Od té doby to občas uvádím ve výuce, když se studenti ptají, proč se používá u testů většinou hladina 0,05.

• *A co studenti? Již po tisíce let se říká, že jsou stále horší a horší, snaží se proplout za cenu nejmenšího odporu a nemají přirozenou úctu ke svým pedagogům a jsou stále drzejší. Souhlasíte s těmito škarohlídy? Jako dlouholetý proděkan pro studijní činnost jste si s nimi jistě „užil“ dost a dost.*

Není pochyb o tom, že dnešní lidé nejsou tak zdvořilí jako bývali dřív. Jako proděkan jsem jednal s mnoha studenty. Často naše setkání začínalo tím, že jsem je poučoval, že mají napřed pozdravit, pak se představit a pak učitele zdvořile oslovit. Většinou to pro ně byla úplně nová informace. Na druhé straně musím konstatovat, že učitelé zdvořilost od studentů ani nevyžadují, jen si mi pak stěžují, že se studenti vůči nim nechovají tak, jak by měli. Většinou však jsem se studenty vycházel přátelsky. Snad jen jeden student z tisíce se snažil zneužít benevolence studijních předpisů a systematicky doslova otravovat ovzduší na fakultě. Obecně mohu říci, že posluchači MFF se chovají lépe než jejich ostatní vrstevníci.

• *Co byste popřál statistice do nejbližších let? Jaký učební text nebo dosud dostatečně nevyučovanou oblast statistiky v českém jazykovém prostředí postrádáte? V které oblasti nejvíce postrádáte původní učebnici v českém jazyce? Kdybyste měl dostatek času, jakou další učebnici byste napsal? Co byste sdělil nastupující generaci vysokoškolských pedagogů a aplikovaných statistiků?*

Statistice bych přál větší počet posluchačů, kteří ji budou studovat. Dosavadní vývoj jejich počtu není příliš optimistický. Z hlediska výuky jsou podle mého názoru všechny hlavní oblasti více či méně pokryty. Zato však porovnání seznamu přednášek a seznamu dostupné české odborné literatury ukazuje, že učebnic je žalostně málo. Není divu, protože napsání učebnice na rozdíl od vědeckého článku není nijak oceňováno. Připadá mi, že psaní učebnice je pokládáno za koníčka, kterému se věnuje jen nějaký málo vytížený učitel. Studenti jsou od prvního ročníku zvyklí, že si všechny přednášky podrobně zapisují a jen z těchto zápisků se ke zkouškám připravují. Pokud náhodou k nějakému předmětu existuje učebnice, stojí nad ní v rozpacích a netuší, že by její užití mohlo výrazně zefektivnit výukový proces.
V časopise The American Statistician byl v roce 2007 uveřejněn rozbor toho, jaké statistické metody jsou v lékařských článcích používány. V 91 článcích časopisu The New England Journal of Medicine bylo toto pořadí: intervaly spolehlivosti (61), kontingenční tabulky (48), analýza přežívání (39) a pak následovaly další metody.
Kdybych se k tomu někdy přinutil, napsal bych učebnici analýzy kategoriálních dat, která by samozřejmě zahrnovala i kontingenční tabulky.
Další učebnice v českém jazyce, která schází, by měla být věnována analýze přežívání. Tu by měl napsat někdo jiný, někdo, kdo se touto tématikou zabýval a zabývá. Možná, že k tomu dojde v rámci nového předmětu, který je věnován analýze cenzorovaných dat.
Nastupujícím pedagogům bych přál příjemné prostředí na katedrách, hodně úspěchů v pedagogické i vědecké práci a dobré rodinné zázemí. Jako měl pan profesor Hájek.

Tázání prováděli Profesores anonymi, Praha

Obsah

Vědecké a odborné statě

Zprávy a informace

Informační bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Vydavatelem je Česká statistická společnost, IČ 00550795, adresa společnosti je Na padesátém 81, 100 82 Praha 10. Evidenční číslo registrace vedené Ministerstvem kultury ČR dle zákona č. 46/2000 Sb. je E 21214.

The Information Bulletin of the Czech Statistical Society is published quarterly.
The contributions in bulletin are published in English, Czech and Slovak languages.

DOI: 10.5300/IB, http://dx.doi.org/10.5300/IB
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)

Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.