Jemný úvod do statistických metod v netržním oceňování

Ing. Jan Brůha PhD.
Karlova univerzita

Struktura prezentací

- První prezentace
  - Cíle, možnosti a omezení
  - Nástroje: metodologie a software
  - CVM (open ended)
- Druhá prezentace
  - TCM (single site)
  - Závěr

Cíle statistických metod

- Ekonomická teorie predikuje jistá tvrzení ohledně netržního oceňování (viz předchozí přednášky)
  - Testovat predikce teorie
  - Numericky kvatifikovat teoretické předpovědi

- Možnosti
  - Využití nástrojů statistické analýzy **společně** s ekonomickou teorií k empirické práci

Cíle prezentací

- Ukázat posluchačům možnosti a omezení statistické analýzy v oblasti metod netržního oceňování
  - Co lze udělat snadno a co těžko
  - Jak interpretovat výsledky
- Nemá suplovat statistické učebnice
  - Použité pojmy budou buď vysvětleny nebo jsou základní
  - Interaktivní učebnice statistiky

  http://badame.vse.cz/iastat

Omezení empirických metod

- Empirické metody mají svou sílu, ale nejsou samospasitelné
  - Neexistují data „sama o sobě“ – vždy nutno interpretovat v určitém paradigmatu
  - Sláva a bída matematické statistiky (testování hypotéz)
  - Nezodpovědné předpoklady mnohdy znehodnocují statistickou analýzu

Metodologie (\1)

- Cíl statistické analýzy
  - Na základě teorie ověřit nebo kvantifikovat hypotézy
    - Je nebezpečné „těžit data“ – data mining
- Klíčové předpoklady:
  - Důležité myšlenky
  - Dobrá data
  - Správné statistické metody

Software (\1)

- Dnes důležité: mít vhodný software
  - Více kritérií pro výběr
    - Cena
    - Snadnost užívání
    - Výběr metod
    - Možnosti implementace speciálních metod
    - Ověřitelnost
  - Dále rozlišíme 3 typy

Software (\2)

- Spreadsheetové nástroje
  - levné,
  - v poslední době vybaveni celkem užitečnými nástroji statistické analýzy,
  - snadné uživatelské ovládání,
  - těžké programování speciálních postupů,
  - mnoho je před uživateli skryto

Software (\3)

- „Profesionální“ programy
  - SPSS, SAS, Statgraphics, TSP, Stata, …
  - už ne tak levné,
  - větší množství procedur než u spreadsheetů,
  - relativně snadné uživatelské ovládání,
  - těžší programování speciálních postupů,
  - stále mnoho je před uživateli skryto.

Software (\4)

- Numerické balíky
  - MATLAB, GAUSS (případně programovací jazyky – Fortran)
  - Drahé,
    - ale existují i lacinější varianty: (Octave, OX, …)
  - Možnost naprogramovat prakticky cokoliv
    - nutnost něco se naučit,
    - ale uživatel má vše pod kontrolou.

CVM (\1)

- Cíle tohoto pod-bloku
  - Jak přistoupit ke zpracování dat
  - Jaké otázky (a za jakých okolností) lze zodpovědět
  - Jak interpretovat výstupy analýz
  - Odkazy

  - Bude ukázáno na Open-Ended CVM

CVM (\2) – příklad

WTP	Pohlaví	Vzdělání	Příjem	Věk	Choroby
8,57	0	2	12365	39	0
7,5731	1	4	28850	43	0
4,8602	0	1	22207	55	1
7,4897	1	2	28393	38	0
10,8755	0	2	33828	42	1
7,0437	0	3	27842	52	1
10,7701	1	4	33544	65	1
10,8019	1	4	34103	39	0
3,5299	0	1	17891	56	1
6,7426	1	1	26648	42	1
10,3151	0	4	33057	46	0
2,2461	0	5	11443	43	1
12,971	0	4	36208	62	1
4,9923	0	2	22778	58	1
8,7723	1	2	30717	28	0

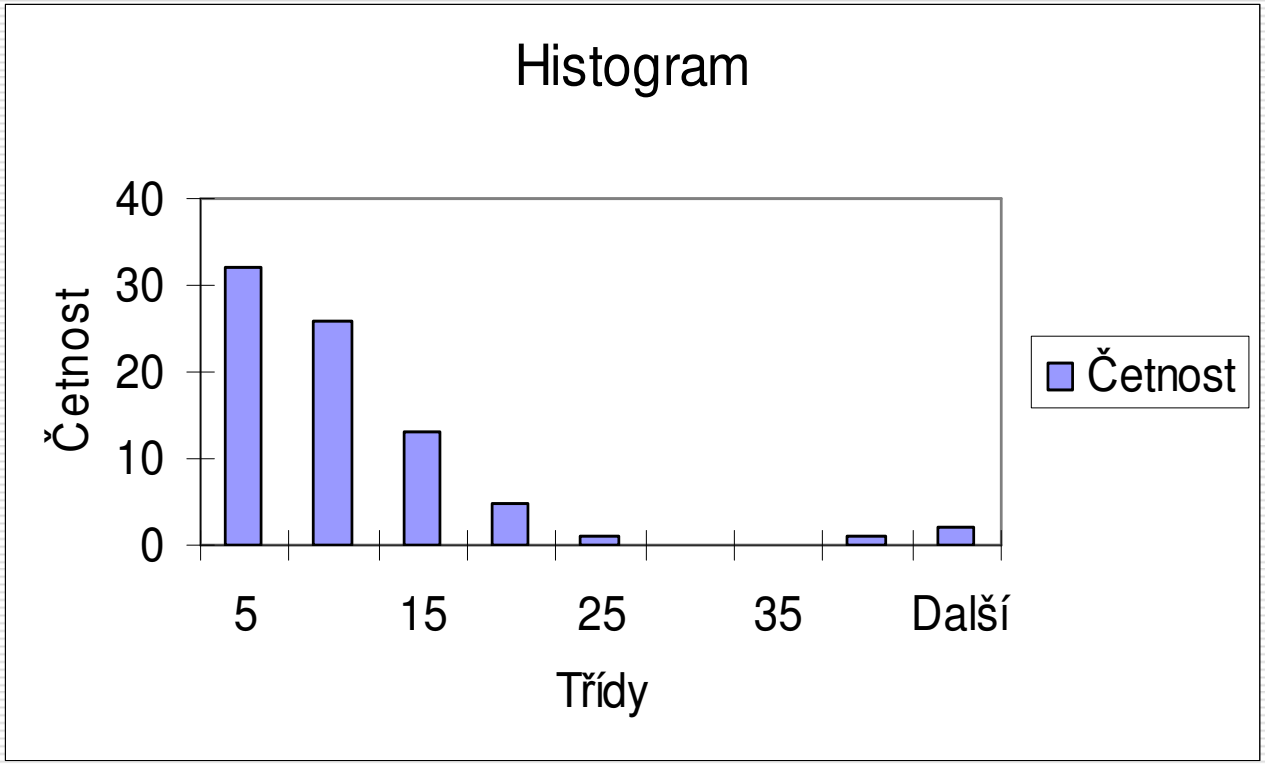
Výběrové charakteristiky (\1)

- Charakteristiky polohy
  - Průměr
  - Medián
- Charakteristiky variability
  - Výběrové rozpětí
  - Rozptyl
  - Směrodatná odchylka

Charakteristika dat (\2a)

Histogram

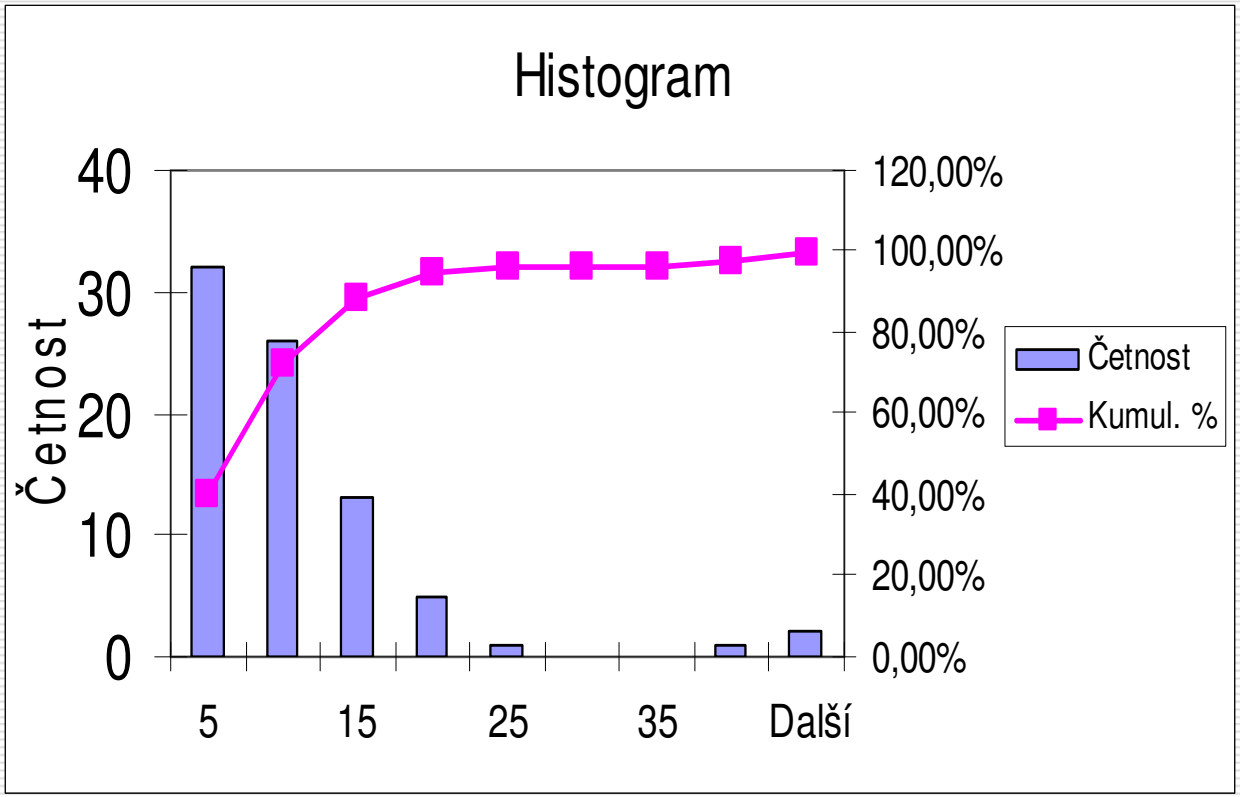
Třídy	*Četnost*
5	32
10	26
15	13
20	5
25	1
30	0
35	0
40	1
Další	2



Charakteristika dat (\2b)

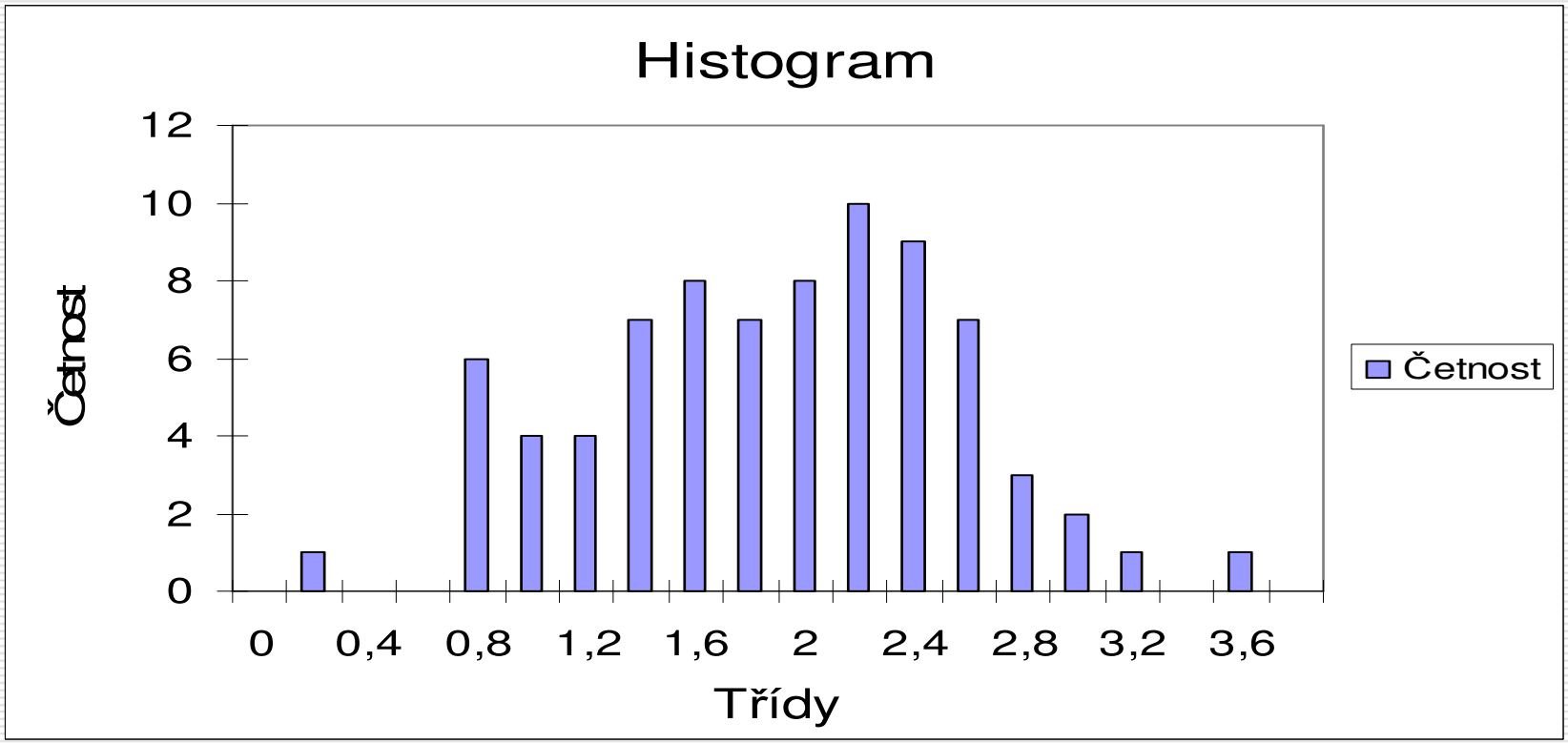
Kumulativní Histogram

Třídy	*Četnost*	*Kumul. %*
5	32	40,00%
10	26	72,50%
15	13	88,75%
20	5	95,00%
25	1	96,25%
30	0	96,25%
35	0	96,25%
40	1	97,50%
Další	2	100,00%



Charakteristika dat (\3)

- Jak vypadá WTP po logaritmování?



Jemnější nástroje (\1)

- Analýza rozptylu
  - Umožňuje testovat závislost veličin (jejich průměru) na diskrétních znacích
- Regresní analýza
  - Mnohorozměrná závislost

$$Y = f(X,\beta)$$

Analýza rozptylu

- Jednorozměrný model

Střední hodnota $x_i = m + t_i$

  - Vliv diskrétního znaku na střední hodnotu zkoumané veličiny
  - Umožňuje testování $t_i > t_k$
    - Mají muži systematicky vyšší WTP než ženy?

- Zobecnění
  - Dvourozměrný model

$x_{ik} = m + t_i + s_k + u_{ik}$

  - Vícerozměrné modely

ANOVA – příklad

závisí WTP na pohlaví?

Anova: jeden faktor

Faktor

Výběr	*Počet*	*Součet*	*Průměr*	*Rozptyl*
Sloupec 1	43	76,71018	1,783958	0,555829
Sloupec 2	37	71,14525	1,922845	0,558111

ANOVA

Zdroj variability	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*
Mezi výběry	0,383621	1	0,383621	0,688872	0,409081	3,963472
Všechny výběry	43,43685	78	0,556883			
Celkem	43,82047	79				

Omezení analýzy rozptylu

- Lze snadno použít pouze při vysvětlování vlivu proměnných, jež nabývají několika málo hodnot

- Většina testů (skrytě) předpokládá normalitu rozložení
  - Nemusí být vždy splněno,
  - a pak chybné výsledky

  - Alternativa: neparametrické a robustní testy
    - Kruskal-Wallisův test (ANOVA1)
    - Friedmanův test (ANOVA2)

Regresní analýza (\1)

- Obecný model

  $Y = f(X,\beta,\varepsilon)$

  - Y vysvětlované proměnné (WTP, …)
  - X vysvětlující proměnné (věk, pohlaví, příjem, vzdělání, …)
  - β vektor koeficientů udávající vliv jednotlivých proměnných
  - ε náhodné chyby

  Přítomnost náhodných chyb neznamená celkovou bezzákonitost, pouze nepřítomnost jednoduchého, zřejmého, deterministického vztahu!!

Regresní analýza (\2a)

- Typy regresních rovnic

  - Lineární

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_N X_N + \varepsilon$$

    - Koeficienty udávají přímý vliv $\beta_i = dY/dX_i$

  - Semi Log-lineární

$$Log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_N X_N + \varepsilon$$

    - Oblíbená ve Open-ended CVM
    - Koeficienty udávají semi-elasticitu $\beta_i = \Delta Y/dX_i$

Regresní analýza (\2b)

- Log-lineární

$Log(Y) = \beta_1 Log(X_1) ++ \ldots \beta_N Log(X_N) + \varepsilon$

  - Koeficienty udávají elasticity $\beta_i = \Delta Y / \Delta X_i$

- Cox-Boxova specifikace
  - Možnost testování funkčního tvaru modelu

- Obecná nelineární specifikace
  - Nutno koeficienty interpretovat případ od případu

Regresní analýza (\3a)

- Metody odhadu
  - Nejčastěji Metoda nejmenších čtverců

$$\sum(Y- \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_N X_N)^2 \rightarrow \min$$

- Výhody
  - Snadno se počítá
  - Známé statistické vlastnosti
- Nevýhody
  - Citlivá k odlehlým pozorováním

Regresní analýza (\3b)

- Alternativy
  - Metoda maximální věrohodnosti
    - Nutné silné předpoklady o rozložení náhodných chyb, ale dobré statistické vlastnosti
    - Může být relativně složitá
    - Více u TCM
  - Robustní metody
    - Např. nejmenší absolutní chyba (LAD)
    - Robustní vůči odlehlým pozorováním
    - Složitější výpočty, statistické vlastnosti nejsou mnohdy dobře probádány

Regresní analýza (\4)

- Typický výstup - interpretace

Regresní statistika	
Násobné R	0,99393323
Pozorování	79

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	5	308,6889097	61,73778	1208,67022	0,0000
Rezidua	74	3,779853084	0,051079		
Celkem	79	312,4687628			

	Koeficienty	*t stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Pohlaví	-0,0208124	-0,405215529	0,686488	-0,1231522	0,081527354
Vzdělání	0,03931011	1,611797899	0,111263	-0,009286	0,087906211
Příjem	0,0000597	26,53850595	0,0000	0,00000546	0,00000635
Věk	0,00420947	3,066970111	0,003019	0,00147467	0,006944277
Choroby	-0,0541953	-1,060692334	0,292279	-0,1560029	0,047612252

Regresní analýza (\5)

- Problémy s regresní analýzou
  - Nezávislost náhodné složky ε na vysvětlujících proměnných X
    - Pokud porušeno, pak jsou odhady vychýlené
    - To se může stát typicky pokud není v modelu proměnná, jež tam má být
      - Např. příjem může „tahat“ vliv vzdělání
  - Výběr modelu a výběr proměnných
    - Hodně složité věci

Aplikace na TCM (\1a)

- V této přednášce se zaměříme na single-site modely
- Obecný model

  $r = f(p_r, p_s, y, X)$

  - r Počet cest
  - p_r cena cesty, p_s cena subsitutů
  - y důchod
  - X ostatní relevantní charakteristiky

Aplikace na TCM (\1b)

- Problémy s měřením ceny cest
  - Náklady cesty (např. pohonné hmoty)
  - Cena vybavení
    - Závisí na účelu cesty
  - Vstupné
    - V některých případech může být těžko pozorovatelné
    - Ceteris paribus představuje vhodný „experiment“
  - Cena času
    - Velice obtížné, ušlá mzda je ne vždy vhodná pomocná proměnná

Aplikace na TCM (\2)

- Odhad spotřebitelského přebytku

  $CS = \int_{p}^{\infty} f(p_r, p_s, y, X) \mathrm{d}p_r$

- Nejčastější problémy
  - On-site sampling
    - Vychýlený soubor
      - Nejsou ti, co tam necestují + větší pravděpodobnost pro návštěvníky, co cestují více
  - Off-site sampling
    - Nákladné, nicméně občas se to děje
      - Hausman, Leonard, McFadden (1995) J Pub Econ

Aplikace na TCM (\3)

- Off-site sampling
  - Ještě významnější problém správné funkcionální specifikace problému
    - ANOVA je dost nevhodná
  - Možné přístupy
    - Lineární regrese
      - Pozor není možné logaritmovat závislou proměnou!
    - Metoda maximální věrohodnosti

Aplikace na TCM (\4)

- Metoda maximální věrohodnosti

$\Pr(R = i) = \pi_i(p_c, p_s, y, X)$

$i = 0,1,2, \ldots$

- Statistické restrikce

  $\pi_i(p_c, p_s, y, X) \geq 0$

  $\Sigma_i \pi_i = 1$

- Možná formulace

$\pi_i(p_c, p_s, y, X) = \exp\{b_{1i} * p_c + b_{2i} * p_s + b_{3i} * y + \ldots\}/M$

Aplikace na TCM (\5)

- Metoda maximální věrohodnosti

  Odhad = argmax Σ_{jj} log Pr(R_j=i)$I_{(Rj=i)}$

  - Nutnost využít optimalizačních postupů
    - Lze to udělat v spreadsheetu, ale lépe je použít jiných programů
      - Protože např. odhad směrodatných chyb odhadů je složitý
  - Problémy s přístupem naznačeným výše
    - Mnoho parametrů
      - Interpretace
      - Problémy s odhady

Aplikace na TCM (\6)

- Redukce parametrů
  - Poissonův model
    - Počet návštěv má Poissonovo rozložení

$$\Pr(R_j = i) = \exp\{-\lambda_j\} * \lambda_j^i / i!$$

$$\lambda_j = \exp\{b_1 * p_{cj} + b_2 * p_{sj} + b_3 * y_j + b_4 * X_j\}$$

    - Spotřebitelský přebytek

$$CS_j = \lambda_j / (-b_1)$$

Aplikace na TCM (\7)

- Obvyklé regresory *X*
  - Velikost rodiny
  - Věk, pohlaví, vzdělání
  - Bydliště (venkov, město)
  - Povolání
  - Členství v klubech, vlastnictví vybavení, zkušenost s aktivitou

Aplikace na TCM (\7)

- Možné problémy s Poissonovým modelem
  1. Příliš mnoho nul
  2. Příliš velká variace

  - 1) může znehodnotit statistickou analýzu (nekonzistentní odhady)
  - 2) vychýlení v odhadu chyb

Aplikace na TCM (\8)

- Alternativy k Poissonovu modelu
  - Explicitní modelování návštěvy

    $\Pr(R_j=0) = \exp\{-\mu_j\}$

    $\Pr(R_j=i) = (1-\exp\{-\mu_j\})^* \exp\{-\lambda_j\}^*\lambda_j^i / i!/(1-\exp\{-\lambda_j\})$

    pro $i > 0$

  - Složitější modely
    - Negativně-binomické rozložení
    - Neparametrické / semiparametrické modely

Aplikace na TCM (\9)

- On-site sampling
  - Je nezbytná korekce počtu návštěv
  - Příklad: Poissonův model

$$\Pr(R_j = i \mid i > 0) = \exp\{-\lambda_j\} * \lambda_j^{i-1} / (i-1)!$$

$$\lambda_j = \exp\{b_1 * p_{cj} + b_2 * p_{sj} + b_3 * y_j + b_4 * X_j\}$$

  - Odhad spotřebitelského přebytku je nezměněn

Aplikace na TCM (\10)

- Odhad CS
  - Off-site sampling

    $CS_j = \lambda_j / (-b_1)$

    $AS = (POP_{off}/N) * \Sigma_j CS_j$
  - On-site sampling

    $M = \Sigma_j (n_j / r_j)$

    $AS = (POP_{on}/M) * \Sigma_j (CS_j / r_j)$

- Pozor: jedná se o náhodné veličiny!!