LEKCE 5
STATISTICKÁ INFERENCE ANEB ZOBECŇOVÁNÍ VÝSLEDKŮ Z VÝBĚROVÉHO NA ZÁKLADNÍ SOUBOR

Ve většině případů pracujeme s výběrovým souborem a výběrové výsledky zobecňujeme na základní soubor. Smysluplné je to ale jen:

- Jde-li skutečně o VÝBĚROVÝ SOUBOR (při vyčerpávajícím šetření to nemá smysl).
- Jde-li o NÁHODNÝ VÝBĚR kdy každá jednotka dané populace má stejnou pravděpodobnost, že bude vybrána.
- Jde-li o NEZÁVISLÝ VÝBĚR (výběr žádné jednotky nezvyšuje ani nesnižuje pravděpodobnost výběru jiných jednotek).

Příklad: Opisují-li studenti v testu, jejich výsledky nejsou nezávislé.

Tak jako se při měření musíme vyrovnat s měřícími chybami, musíme se při inferenci vyrovnat s výběrovou chybou (výběr je jen částí základního souboru).

ZÁKLADNÍ OTÁZKA: JAK JSOU POZOROVANÉ VÝSLEDKY PRAVDĚPODOBNÉ?

Pozorovaný výsledek představuje

- STATISTIKU (jak je to v našem výběrovém souboru),

z níž usuzujeme na

- PARAMETR (jak je to v populaci, z níž byl soubor vybrán).

PARAMETR

Neznámá (pokud nemáme vyčerpávající šetření) vlastnost základního souboru

- μ = průměr základního souboru
- σ = standardní odchylka základního souboru
- σ^2 = variance základního souboru

PARAMETRY ZÁKLADNÍHO SOUBORU obvykle neznáme, ale můžeme je odhadovat z VÝBĚROVÝCH STATISTIK.

STATISTIKA

Známá vlastnost výběrového souboru

- $\bar{X}$ = průměr výběrového souboru
- s = standardní odchylka výběrového souboru
- s^2 = variance výběrového souboru

Smysl otázky "JAK JSOU POZOROVANÉ VÝSLEDKY PRAVDĚPODOBNÉ?":

- Lze, se zvolenou pravděpodobností předpokládat, že STATISTIKA jakožto pozorovaný výsledek reprezentuje nepozorovatelný PARAMETR?
- Není STATISTIKA v důsledku výběrové chyby přece jen příliš vzdálená PARAMETRU?
- V jakém intervalu kolem STATISTIKY můžeme s danou pravděpodobností očekávat výskyt PARAMETRU?

INFERENCE ZE STATISTIKY NA PARAMETR

- BODOVÝ ODHAD jako číslo, jehož hodnota je v nějakém (teoreticky) stanoveném smyslu optimálně určena.
- INTERVALOVÝ ODHAD, kdy hledáme interval (spolehlivosti), v kterém s určitou, předem zvolenou pravděpodobností neznámý populační parametr leží.

NA PŘEDCHOZÍ OTÁZKY LZE ODPOVĚDĚT DÍKY VLASTNOSTEM NORMÁLNÍHO RESPEKTIVE STANDARDIZOVANÉHO NORMÁLNÍHO ROZLOŽENÍ.

STANDARDNÍ ODCHYLKA VE VÝBĚROVÉM SOUBORU

$$\sigma = \sqrt{\frac{\sum (\bar{x}_i - x)^2}{N}}$$

STANDARDNÍ ODCHYLKA V ZÁKLADNÍM SOUBORU

$$\sigma = \sqrt{\frac{\sum (\bar{x}_i - \mu)^2}{M}}$$

STANDARDNÍ CHYBA PRŮMĚRU

$$\sigma = \sqrt{\frac{\sum (\bar{x}_i - \mu)^2}{n_s}}$$

μ ← populační průměr

n_s ← počet provedených výběrů

$\bar{x}_i$ ← průměr z provedených výběrů

Příklad různých náhodných výběrů

VÝBĚROVÉ SOUBORY	průměr	std. odchylka
1. výběr (N = 892)	56,5	13,35
2. výběr (N = 892)	56,8	13,52
3. výběr (N = 892)	56,5	13,34
4. výběr (N = 892)	56,5	13,26
5. výběr (N = 892)	56,7	13,33
PRŮMĚR	56,6	13,36
ZÁKLADNÍ SOUBOR (N=1191)	56,4	13,33
ROZDÍL (při 5 výběrech)	0,2	0,03

PROČ JE STANDARDNÍ/SMĚRODATNÁ CHYBA PRŮMĚRU DŮLEŽITÁ?

S 95% pravděpodobností (5% riziko chyby) můžeme tvrdit, že:

průměr základního souboru (parametr)
=
průměr výběrového souboru (statistika)
± 1,96 směrodatná chyby
(často se zaokrouhluje na dvojnásobek)

S 99% pravděpodobností (1% riziko chyby) můžeme tvrdit, že:

průměr základního souboru (parametr)
=
průměr výběrového souboru (statistika)
± 2,96 směrodatná chyby
(často se zaokrouhluje na trojnásobek)

DOSTÁVÁME SE K POJMU INTERVAL SPOLEHLIVOSTI

Protože pracujeme s výběrovými soubory, můžeme vypočítat statistiky, ale nevíme, jak tyto statistiky korespondují s parametry. Víme ovšem, že se - se zvolenou pravděpodobností - pohybují v intervalu (spolehlivosti), jehož obecný vzorec je:

$$C.I = \bar{X} \pm z . \sigma_X$$

$\bar{X}$ = vypočítaný výběrový průměr (statistika)
z = z-skóre korespondující s požadovanou úrovní pravděpodobnosti (hladinou významnosti). Pro HV=95% je to 1,96.
σ_X = standardní/směrodatná chyba distribuce výběrových průměrů

Interval spolehlivosti pro 95% HV znamená:
Jestliže bychom z populace opakovaně činili výběry stejné velikosti, v 95% z nich výběrů by se populační průměr nacházel uvnitř intervalu spolehlivosti (s 95% pravděpodobnost interval spolehlivosti tento populační průměr zahrnuje).

INTERVAL SPOLEHLIVOSTI (pro průměr na HV = 95%)

$$C.I._{95\%} = \bar{X} \pm 1,96 . s/\sqrt{N}$$

s/ = standardní/směrodatná chyba

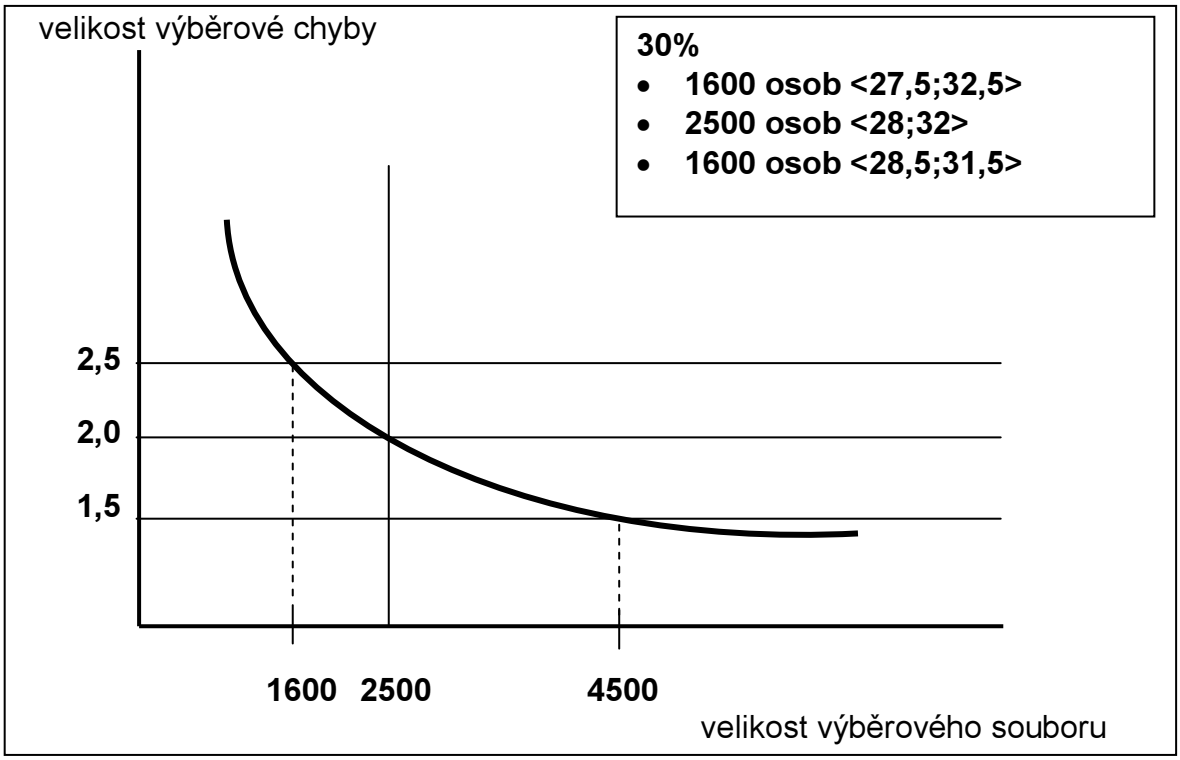
INTERVAL SPOLEHLIVOSTI (pro procento výskytu na HV = 95%)

$$C.I._{95\%} = p \pm 1,96 . \sqrt{p.(1-p) / N}$$

- p = pozorovaný podíl, kolem něhož je interval spolehlivosti konstruován
- N = velikost výběrového souboru

VELIKOST VÝBĚROVÉHO SOUBORU A VELIKOST VÝBĚROVÉ CHYBY

□ VÝBĚROVÁ CHYBA sice s velikostí vzorku klesá, ale po dosažení určité velikosti souboru je její zmenšování s dalším zvětšováním výběru nepodstatné. Proto není další růst početnosti výběrového souboru ekonomický.



□ Stanovíme-li si přípustnou výběrovou chybu (nakolik se, s jistou zvolenou pravděpodobností, mohou výsledky zjištěné ve výběrovém souboru odchylovat od skutečnosti v základním souboru), můžeme určit potřebnou velikost výběrového souboru. A to s přihlédnutím k homogenitě základního souboru z hlediska vlastností, které nás zajímají (usuzujeme na ni z velikosti rozptylu).

VELIKOST VÝBĚROVÉHO SOUBORU A VÝBĚROVÁ CHYBA (NA HV=95%)

výběrová chyba v %	interval spolehlivosti	velikost výběru (sample size)	výběrová chyba v %	interval spolehlivosti	velikost výběru (sample size)
1,0	±1,0	10000	6,0	±6,0	277
1,5	±1,5	4500	6,5	±6,5	237
2,0	±2,0	2500	7,0	±7,0	204
2,5	±2,5	1600	7,5	±7,5	178
3,0	±3,0	1100	8,0	±8,0	156
3,5	±3,5	816	8,5	±8,5	138
4,0	±4,0	625	9,0	±9,0	123
4,5	±4,5	494	9,5	±9,5	110
5,0	±5,0	400	10,0	±10,0	100
5,5	±5,5	330			

- Výběrová chyba (Sampling Error). Pro 95% hladinu významnosti (Confidence Level) de facto dvě standardní chyby
- Interval spolehlivosti (Confidence Interval) vymezují Confidence Limits) - jeho hraniční hodnoty.
- V případě tabelovaných hodnot se předpokládá heterogenní soubor (50%:50%) - platí pro alternativní neboli binomické proměnné. Vezmeme-li v úvahu heterogenitu souboru, je výpočet intervalu spolehlivosti složitější

VÝBĚROVÁ CHYBA V ZÁVISLOSTI NA HOMOGENITĚ VÝBĚROVÉHO SOUBORU

- V každém sloupci výběrové chyby pro příslušná procenta sledované vlastnosti ve výběrovém souboru
- V každém řádku výběrové chyby pro danou velikost výběrového souboru

velikost výběru	1% nebo 99%	5% nebo 95%	10% nebo 90%	15% nebo 85%	20% nebo 80%	25% nebo 75%	30% nebo 70%	35% nebo 65%	40% nebo 60%	45% nebo 55%	50%
25	4,0	8,7	12,0	14,3	16,0	17,3	18,3	19,1	19,6	19,8	20,0
50	2,8	6,2	8,5	10,1	11,4	12,3	13,0	13,5	13,9	14,1	14,2
75	2,3	5,0	6,9	8,2	9,2	10,0	10,5	11,0	11,3	11,4	11,5
100	2,0	4,4	6,0	7,1	8,0	8,7	9,2	9,5	9,8	9,9	10,0
150	1,6	3,6	4,9	5,9	6,6	7,1	7,5	7,8	8,0	8,1	8,2
200	1,4	3,1	4,3	5,1	5,7	6,1	6,5	6,8	7,0	7,0	7,1
250	1,2	2,7	3,8	4,5	5,0	5,5	5,8	6,0	6,2	6,2	6,3
300	1,1	2,5	3,5	4,1	4,6	5,0	5,3	5,5	5,7	5,8	5,8
400	0,99	2,2	3,0	3,6	4,0	4,3	4,6	4,8	4,9	5,0	5,0
500	0,89	2,0	2,7	3,2	3,6	3,9	4,1	4,3	4,4	4,5	4,5
600	0,81	1,8	2,5	2,9	3,3	3,6	3,8	3,9	4,0	4,1	4,1
800	0,69	1,5	2,1	2,5	2,8	3,0	3,2	3,3	3,4	3,5	3,5
1000	0,63	1,4	1,9	2,3	2,6	2,8	2,9	3,1	3,1	3,2	3,2
2000	0,44	0,96	1,3	1,6	1,8	1,9	2,0	2,1	2,2	2,2	2,2
5000	0,28	0,62	0,85	1,0	1,1	1,2	1,3	1,4	1,4	1,4	1,4

Pramen: A Broadcast Research Primer. National Association of Broadcasters, Washington, DC 1976, p. 19.