Inferenční statistika

Nejdůležitější sdělení

- Statistická inference ukazuje, jak spolehlivé jsou výsledky některých statistických analýz

Důležité pojmy

- Populace
- Vzorek
- Standardní chyba
- Interval spolehlivosti
- Hladina významnosti

Populace

- Množina všech prvků, kterých se týká náš výzkum
  - Základní soubor
- Politické strany, kandidáti, plakáty v kampaních, země světa
- Voliči/lidé

Vzorek

- Případy z populace, které jsou zahrnuty v našem výzkumu
- Někdy může vzorek zahrnovat celou populaci
  - Nejlepší možná situace
  - Výsledky analýz platí pro celou populaci
  - Není nutné dělat výběry, jen abychom mohli mít inferenční statistiku
- Pokud ne, tak máme výběrový soubor
  - Vzorek je zatížen řadou chyb
    - Výběrovou chybu můžeme spočítat
  - Otázka, zda naše výsledky platí pro celou populaci
    - => inferenční statistika

Jak může vzorek vzniknout?

- Vezme co je po ruce
  - Dotazníky distribuované po sociálních sítích
  - „Sněhová koule“
  - Zvířata v zoo

- Náhodný (pravděpodobnostní) výběr
  - Každý člen populace má stejnou pravděpodobnost, že se dostane do vzorku
  - reprezentativita

Problémy s populací

- Známe populaci?
  - Bez znalosti populace je těžké vytvořit výběr
- Je dostupný seznam všech členů populace?
  - Např. registr obyvatelstva (v čr existuje, ale není dostupný)
- Je možné všechny členy populace „kontaktovat“/zjistit o nich údaje?
  - Exit-poll při možnosti poštovní volby

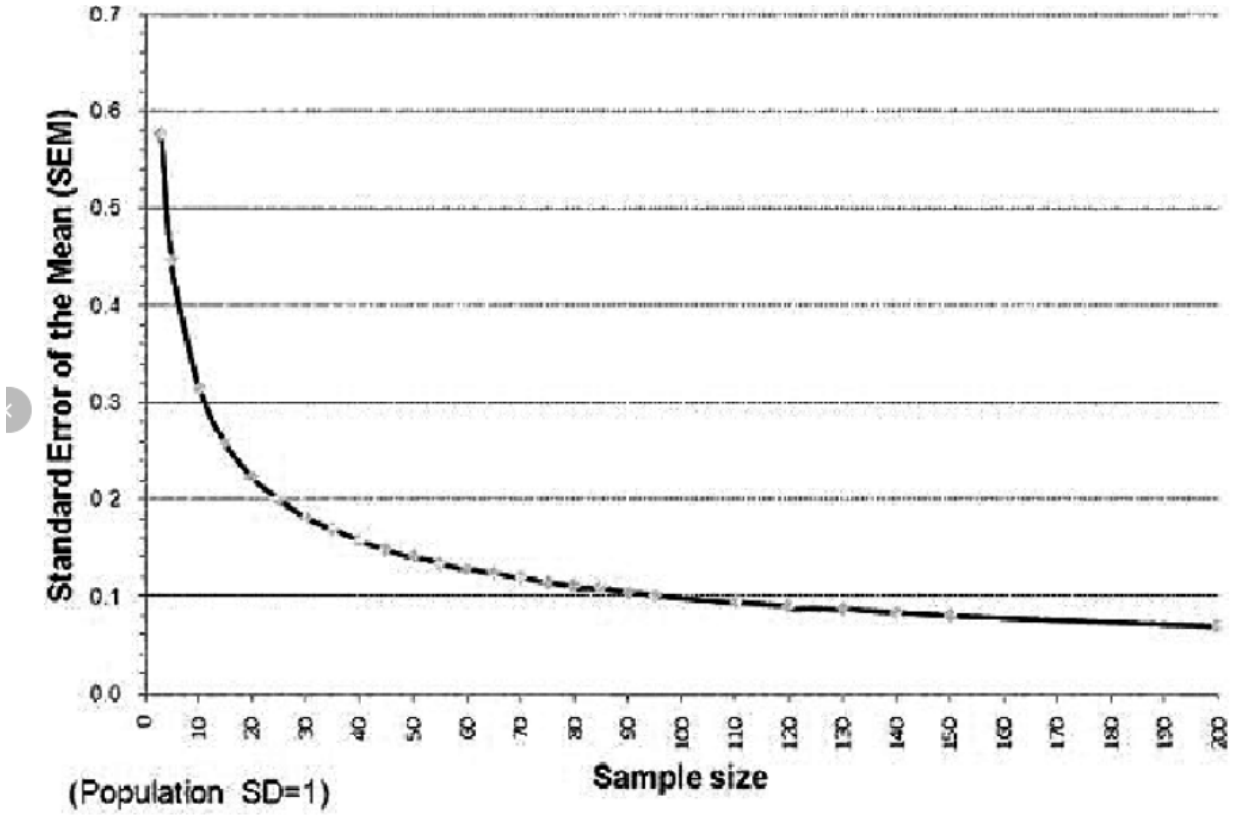
Jak velký by měl být vzorek

- Jaké máme k dispozici prostředky?
  - Větší vzorek stojí více peněz a/nebo času
- Jak velkou přesnost požadujeme?
  - Velký vzorek umožňuje mít spolehlivější výsledky
- Jaké bude chtít dělat analýzy?
  - Různé typy analýz potřebují různě velké vzorky
- Jak běžný/raritní je jev, který chceme zkoumat?
  - Pokud chceme zkoumat něco, co se děje vzácně, potřebujeme hodně velký vzorek

- Dostatečný počet může být 30 nebo také 2000

Proč je velikost vzorku důležitá

- Čím větší vzorek, tím jistější výsledek
- Od velikosti cca 2000 už jistota prakticky neroste
- Velký skok mezi 2 - 30



Inferenční statistika

- Můžeme to, co jsme zjistili na vzorku, zobecnit na celou populaci?
- Má to, co jsme zjistili, nějaký smysl?
- Jak moc si můžeme být jisti výsledkem?
- Jen v případě **pravděpodobnostních** výběrových souborů
- Spolehlivost, signifikance

Znáte z předchozích hodin

- Co je to průměr?
- Co je to směrodatná odchylka?
- Co je to normální rozdělení?
- Co je to n?

Odhady

- Jaká je průměrná ideologická pozice voličů?
  - Pokud pracujeme se vzorkem, tak získaný průměr je jen odhad (bodový odhad)
  - Lepší je poskytovat intervalový odhad
    - V jakém rozmezí se „pravděpodobně“ pohybuje skutečný průměr v populaci
      - Interval spolehlivosti
      - Různě široký v závislosti na hladině významnosti
    - Skutečný průměr obvykle neznáme

Výpočet odhadu průměru

- 95% interval spolehlivosti průměru
- Průměr ± 1.96xSE (SE – Standard error, česky směrodatná chyba)
- SE= směrodatná odchylka/√počet případů

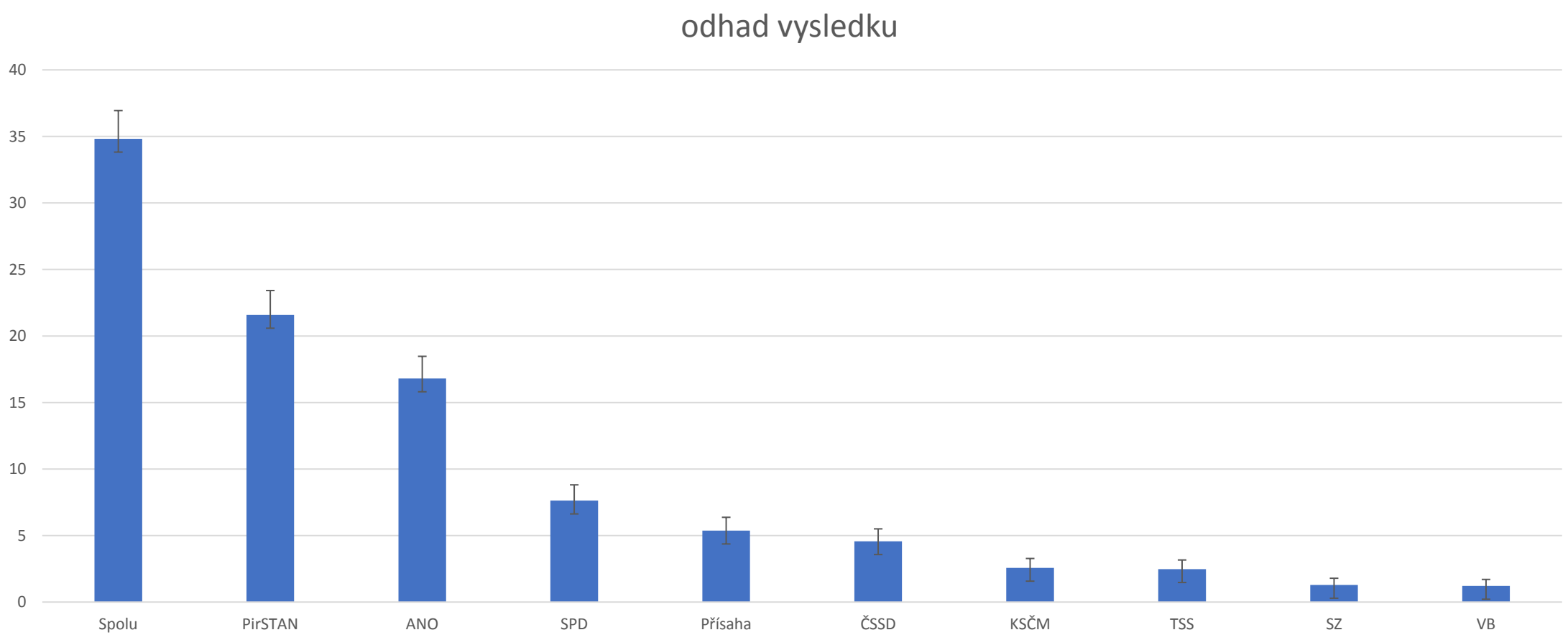
- Pokud je vypočtený průměr 100, sm.odch. 10 a N (počet případů) 100
- 100 ± 1.96x(10/√100)
- =98 – 102

Výpočet odhadu četnosti (procent)

- C.I. 95% = $p \pm 1{,}96 \times \sqrt{p(1-p) / n}$
- P je relativní četnost před vynásobením 100

Příklad exitpoll

- Chceme zjistit, jak dopadnou volby, ještě než budou sečtené výsledky
- Ptát se všech voličů by bylo velice nákladné a trvalo by to
- Ptáme se jen některých voličů – náhodný výběr
  - Každý 5 odcházející od voleb
- Z vytvořeného vzorku můžeme **odhadnout** podporu stran nebo průměrnou ideologickou pozici


odhad vysledku
40
35
30
25
20
15
10
5
0
Spolu
PirSTAN
ANO
SPD
Přísaha
ČSSD
KSČM
TSS
SZ
VB

strana	realný výsledek	odhad	dolní hranice	horní hranice	odhad - realita
Spolu	34.5	34.8	32.7	36.9	0.4
PirSTAN	18.6	21.6	19.7	23.4	3.0
ANO	21.5	16.8	15.1	18.5	-4.7
SPD	7.4	7.6	6.4	8.8	0.2
Přísaha	4.9	5.4	4.4	6.4	0.5
ČSSD	4.1	4.6	3.6	5.5	0.5
KSČM	2.4	2.6	1.9	3.3	0.2
TSS	2.7	2.5	1.8	3.2	-0.2
SZ	1.3	1.3	0.8	1.8	0.0
VB	1.1	1.2	0.7	1.7	0.1

Nejdůležitější sdělení

- Statistická inference ukazuje, jak spolehlivé jsou výsledky statistických analýz provedených na „dobrých“ vzorcích

Důležité pojmy

- Populace - všechny případy
- Vzorek – některé připady
- Náhodný výběr – všechny případy mají stejnou pravděpodobnost dostat se do vzorku
- Odhad – jakákoli statistika spočítaná na vzorku
- Standardní chyba – jak velká je chyba odhadu
- Interval spolehlivosti – v jakém rozpětí se s jistou spolehlivostí pohybuje pravá hodnota v populaci
- Hladina významnosti – určuje jakou chceme spolehlivost