7_Rozložení výběrových statistik

Distribuce výběrových proporcí

Distribuce výběrových průměrů

Inferenční statistika

- Inference používá statistiky (průměr,proporce) ze vzorku za účelem rozhodování o hodnotě parametrů v populaci
- Jak pravděpodobnost a normální rozložení poskytuje základnu pro statistickou inferenci

Výběrová distribuce

- Druh pravděpodobnostního rozložení
- Umožňuje zjistit jak daleko od populačního parametru pravděpodobně statistika vzorku leží

Příklad

- Povolební průzkum (parametr dosud neznámý) na vzorku 3889 voličů ukazuje proporci pro zelené 4,5 procenta (0,045)
- Jak víme že tento odhad je dobrým odhadem (blízko populační proporci)? Sestavením distribuce výběrových proporcí

Distribuce výběrových proporcí

- Hodnoty náhodné proměnné (0 = nezelení a 1 = zelení) a jejich četnost (0,955 a 0,045) z jednoho průzkumu formují distribuci dat pro jeden vzorek (individuální data, část populace, mění se vzorek od vzorku, tedy i proporce je proměnlivá)

- Celkový výsledek voleb dopadl pro zelené 3,19 procenta = populační proporce v době průzkumu neznámá. Hodnoty náhodné proměnné (0=nezelení a 1=zelení) a jejich četnost (0,0319 a 0,9681) v populaci = populační distribuce. (individuální data, distribuce z které bereme vzorek, parametr je fixní ale neznámý)

- Měly by jiné průzkumy na jiných vzorcích tendenci být blíže nebo dále skutečné populační proporci? Klíč se nachází v distribuci výběrových proporcí. (sdružuje hodnoty statistik vzorků, poskytuje pravděpodobnosti všech možných hodnot konkrétní statistiky, hypotetická distribuce neboť ve skutečnosti pozorujeme pouze data jednoho vzorku – distribuci dat)

Distribuce výběrových statistik a její konstrukce

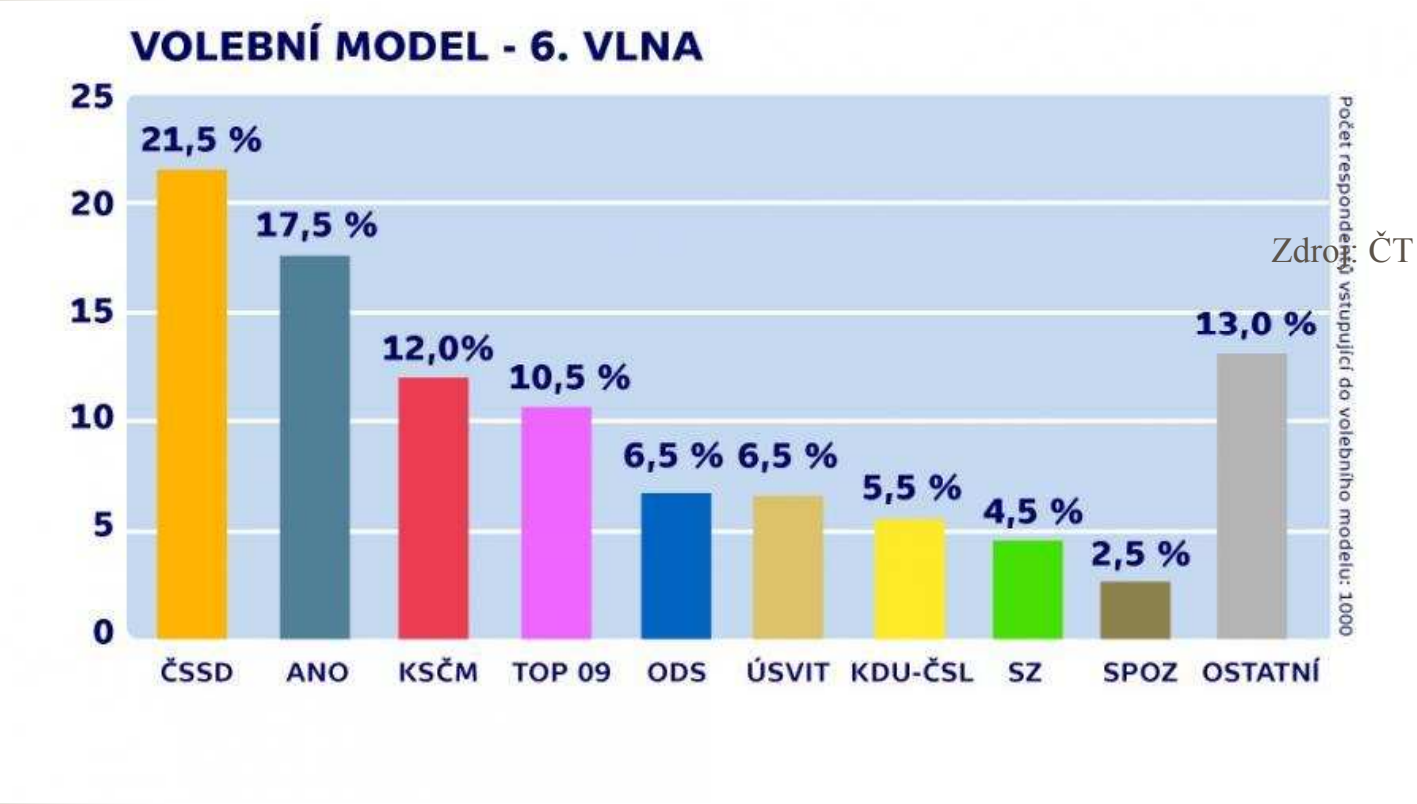
- = distribuce pravděpodobností všech možných výsledků konkrétní statistiky (př. proporce, průměr)
- Jak často konkrétní hodnota statistiky je očekávána při náhodném výběru
- Opakovaně vybírám vzorek a hodnoty statistik všech vzorků nanáším na novou distribuci = distribuce výběrových statistik

Průměr, odchylka a tvar distribuce výběrových proporcí

- Průměr a odchylka závisí na velikosti vzorku a populační proporci
- Průměr = p = populační proporce (pokud neznáme používáme proporci ve vzorku jako nejlepší odhad proporce v populaci)
- Směrodatná odchylka = $\sqrt{\frac{p(1-p)}{n}}$
- Pokud je velikost vzorku dostatečně velká takže očekávaný počet výskytu v kategorii zájmu (počet hlasů pro zelené) a očekávaný počet výskytu v ostatní kategorii (počet hlasů pro jiné strany) jsou větší než 15, pak má distribuce tvar normálního rozložení

Příklad: parlamentní volby ČR 2013

- Poslední předvolební průzkum ČR pro ČT naznačuje 4,5 procenta hlasů pro zelené
  - Ot. 1. Jak blízko je statistika proporce blízko skutečné proporci v populaci?
  - Ot. 2. Jaké jsou pravděpodobné hodnoty skutečného populační proporce?



Zdroj: ČT

- Ot. 1: směrodatná odchylka = √ (p(1-p) / n) = √ (0,045*0,955 / 1000) = 0,0065
- Protože je splněna podmínka pro normální rozložení (1000*0,0045 a 1000*0,955 > 15) leží 99,8% plochy pod křivkou v rozmezí +- 3 směrodatné odchylky od průměru

  tj. 0,045+(3*0,0065) a 0,045-(3*0,0065) = 0,045 +- 0,02 = 0,043 až 0,047
- S pravděpodobností téměř 100% leží populační proporce někde v intervalu 0,043 až 0,047 tedy 4,3 až 4,7 procent
- Zelení ve skutečnosti získali pouze 3,19 procenta, proč?
  - 1. preference voličů se mezi posledním průzkumem a dnem voleb změnily – populační proporce se změnila
  - 2. výběr nebyl proveden náhodně

Distribuce výběrových průměrů

- opakovaně vybírám vzorek a jeho průměry nanáším na novou distribuci
- vzniká nová distribuce s těmito charakteristikami:
  - 1. Průměr distribuce = průměr výběrových průměrů = populační průměr (zákon velkých čísel)
  - 2. Odchylka = chybu průměru = $\sigma_{m(\bar{x})} = \sigma / \sqrt{n}$
  - 3. Čím větší velikost vzorku, tím víc se distribuce blíže normální distribuci, bez ohledu na tvar populační distribuce (centrální limitní věta)
    - distribuce se blíží normálnímu rozdělení když populační distribuce je normálně rozdělena (na velikosti vzorku nezáleží) nebo když populace není normálně rozdělena a velikost výběru je větší než 30

Příklad: distribuce výběrových průměrů

- Př. Výsledky IQ testu jsou aproximovány (blíží se) normálním rozložením o průměru $\mu = 100$ a $\sigma=16$. Když vytáhneme z této populace vzorek o velikosti 36 dětí, jak je pravděpodobné, že dosáhne průměru 105 bodů a více?
- vypočítám chybu průměru = $\sigma_{m(\bar{x})} = \sigma / \sqrt{n} = 16 / \sqrt{36} = 2.67$
- Očekáváme že pokud je populační předpoklad pravdivý (vzorek je tažen z populace s danými parametry) pak výběrová distribuce bude v rozsahu 100 +- 8 = 92 až 108
- Vypočítám z-skór pro vzorek = $(\bar{x} - \mu) / \sigma_{m(\bar{x})}$ $= (105 – 100) / 2.67 = 1.87$
- A příslušnou pravděpodobnost z tabulky pro Z = 1.87
- Výsledek: $P(\bar{x} >= 105) = 0.03$
- Pravděpodobnost je velmi nízká