Výberové skúmanie

Štatistická indukcia

Osnova prednášky

- Cieľ výberového skúmania
- Druhy výberov
- 2 úlohy výberového skúmania
- Intervaly spoľahlivosti
  - pre strednú hodnotu základného súboru
  - pre relatívnu početnosť základného súboru

Typy súborov

- **základný súbor** (rozsah N prvkov – štatistických jednotiek alebo nekonečný)
- **výberový súbor** (rozsah n prvkov – štatistických jednotiek)
  - je vytvorený výberom zo základného súboru,
  - reprezentatívnosť! – každá n-tica náhodných jednotiek má rovnakú pravdepodobnosť, že bude vybraná

Cieľ výberového zisťovania

- na základe informácií, získaných vo výberovom súbore urobiť úsudky o stave a správaní sa javu (javov)

  v celom základnom súbore.

Druhy výberov

technika výberu závisí od cieľa zisťovania

náhodný výber

- jednoduchý (prostý) - losovanie
- oblastný – do skupín, náhodný výber v skupinách
- skupinový – náhodný výber skupín, všetky prvky
- systematický – náhodne prvá jednotka, krok
- viacstupňový – náhodne skupiny, náhodne jednotky

nenáhodný výber

- zámerný (expertný) – názor odborníka
- konvenčný – podľa zoznamu, prvých napr. 10
- kvótny - kvóty
- samovýber – ankety v časopisoch
- náhodilý – náhodné stretnutie

Náhodný výber

s rozsahom n je n-tica nezávislých náhodných premenných $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ s rovnakým rozdelením.

Zistené (pozorované) údaje $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$, hodnoty náhodných premenných, sú **realizáciou náhodného výberu.**

Z výberových dát sa počítajú **výberové charakteristiky** (tzv. **štatistiky**), ktoré sú tiež náhodnými premennými: výberový priemer, výberová smerodajná odchýlka, výberová relatívna početnosť a iné.

Niektoré výberové štatistiky a parametre základného súboru

odhad parametra – náhodná premenná $\hat{\theta}(\boldsymbol{X})$	parameter základného súboru θ
$\overline{X}$ - výberový priemer	μ - stredná hodnota základného súboru
S_1^2 - výberový rozptyl	σ^2 - rozptyl základného súboru
S_1 - výberová smerodajná odchýlka	σ - smerodajná odchýlka základného súboru
P - výberový podiel	π - relatívna početnosť základného súboru
b - výberový regresný koeficient	β - regresný koeficient základného súboru
r - výberový koeficient korelácie	ρ – koeficient korelácie základného súboru

Získané údaje z výberového súboru je možné využiť

1. pre **odhad parametra základného súboru** a to dvoma spôsobmi:
   - bodový odhad,
   - intervalový odhad (prostredníctvom konštrukcie intervalu spoľahlivosti),
2. na **testovanie štatistických hypotéz**, teda predpokladov o hodnotách parametrov základného súboru, prípadne viacerých základných súborov alebo o ich rozdelení.

Vlastnosti odhadov

- konzistentnosť:
$$\lim_{n\to\infty} \mathrm{P}\left(\left|\hat{\theta}_n(\boldsymbol{X})-\theta\right|<\varepsilon\right)=1 \text{ pre } \varepsilon>0$$
- neskreslenosť (nevychýlenosť):
$$\mathrm{E}\left[\hat{\theta}_n(\boldsymbol{X})\right]=\theta$$
- asymptotická neskreslenosť:
$$\lim_{n\to\infty}\left(\mathrm{E}\left[\hat{\theta}_n(\boldsymbol{X})\right]-\theta\right)=0$$
- výdatnosť, efektívnosť – zo všetkých odhadov najmenší rozptyl

Intervaly spoľahlivosti pre parametre základného súboru

Interval spoľahlivosti pre **strednú hodnotu** (Confidence Interval For Mean)

- $\mathrm{E}(\overline{\boldsymbol{X}}) = \mu$
- $\mathrm{D}(\overline{\boldsymbol{X}}) = \sigma^2/n$

$$U=\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N[0;1]$$

$$P(-u_{1-\frac{\alpha}{2}} \le U \le u_{1-\frac{\alpha}{2}})=1-\alpha$$

$$P\left(\overline{X}-u_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}+u_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

$$\boxed{\left(\overline{X}-u_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}};\ \overline{X}+u_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\right)}$$

$$\Delta = u_{1-\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$$

je **prípustná chyba odhadu**, z ktorej vyplýva, že rozsah výberového súboru
$$n=\frac{u_{1-\alpha/2}^2\cdot\sigma^2}{\Delta^2}$$

1-α je **spoľahlivosť odhadu** t.j. pravdepodobnosť, že interval pokrýva skutočnú hodnotu parametra základného súboru. t. z. (1-α) – percent intervalov skonštruovaných na základe náhodných výberov s rovnakým rozsahom, pokrýva skutočnú hodnotu parametra

α je riziko odhadu, t.j. pravdepodobnosť, že interval spoľahlivosti nepokrýva skutočnú hodnotu parametra základného súboru

- ak σ ZS nepoznáme

$$T = \frac{\bar{X} - \mu}{\frac{S_1}{\sqrt{n}}} \sim t(n-1)$$

$$P(-t_{1-\frac{\alpha}{2};n-1} \leq T \leq t_{1-\frac{\alpha}{2};n-1}) = 1-\alpha$$

$$P\left(\bar{X} - t_{1-\alpha/2;n-1} \cdot \frac{S_1}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\alpha/2;n-1} \cdot \frac{S_1}{\sqrt{n}}\right) = 1-\alpha$$

$$\boxed{\left(\bar{X} - t_{1-\alpha/2;n-1} \cdot \frac{S_1}{\sqrt{n}}; \bar{X} + t_{1-\alpha/2;n-1} \cdot \frac{S_1}{\sqrt{n}}\right)}$$

Jednostranné intervaly spoľahlivosti pre **strednú hodnotu**:

- ľavostranné

$$\left(\bar{X} - u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}; +\infty\right)$$

$$\left(\bar{X} - t_{1-\alpha;n-1} \cdot \frac{S_1}{\sqrt{n}}; +\infty\right)$$

- pravostranné

$$\left(-\infty; \bar{X} + u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

$$\left(-\infty; \bar{X} + t_{1-\alpha;n-1} \cdot \frac{S_1}{\sqrt{n}}\right)$$

Pozn.: prevodník na hľadanie kritickej hodnoty t – rozdelenia:

$t_{1-\frac{\alpha}{2},\nu}$ sa v Exceli hľadá ako $t_{\alpha,\nu}$

$t_{1-\alpha,\nu}$ sa v Exceli hľadá ako $t_{2\alpha,\nu}$

$t_{\alpha,\nu}$

stĺpec tabuľky

riadok tabuľky

Interval spoľahlivosti pre relatívnu početnosť

Ak pravdepodobnosť nastatia javu A v základnom súbore je π, potom náhodná premenná

$$\boldsymbol{P = \frac{X}{n}}$$

kde X je počet nastatí javu A má binomické rozdelenie, pričom platí $E(X) = n\pi$ a $D(X) = n\pi.(1 - \pi)$,ak platí $\boxed{n > \frac{9}{\pi \cdot (1-\pi)}}$, môžeme uvažovať *normálne* rozdelenie tejto náhodnej premennej a vytvoriť normovanú premennú v tvare

$$U = \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Keďže platí

$$P(-u_{1-\frac{\alpha}{2}} \leq U \leq u_{1-\frac{\alpha}{2}}) = 1-\alpha$$

potom interval spoľahlivosti pre relatívnu početnosť základného súboru má za podmienky

$$n > \frac{9}{p \cdot (1-p)}$$

nasledovný tvar:

$$\boxed{\left(p - u_{1-\alpha/2}\sqrt{\left(\frac{p \cdot (1-p)}{n}\right)}; p + u_{1-\alpha/2}\sqrt{\left(\frac{p \cdot (1-p)}{n}\right)}\right)}$$

resp. v prípade jednostranných intervalových odhadov

$$\left(-\infty; p + u_{1-\alpha}\sqrt{\left(\frac{p \cdot (1-p)}{n}\right)}\right) \quad \text{a} \quad \left(p - u_{1-\alpha}\sqrt{\left(\frac{p \cdot (1-p)}{n}\right)}; +\infty\right)$$

Prípustná chyba odhadu

v prípade obojstranného **intervalu spoľahlivosti pre relatívnu početnosť** má tvar:

$$\Delta = u_{1-\alpha/2}\sqrt{\left(\frac{p \cdot (1-p)}{n}\right)}$$