

Cvičenie 9 – Intervalový odhad strednej hodnoty a rozptylu

Základné pojmy

Stručný prehľad „nástrojov“, ktoré budeme používať. Určite si pozrite aj uvedené odkazy na wikipedii – aspoň obrázky (grafy).

Normálne rozdelenie

– už poznáme

Studentovo rozdelenie / t-rozdelenie

– v tabuľkách alebo cez online kalkulačky

https://en.wikipedia.org/wiki/Student%27s_t-distribution

Veľmi zjednodušene povedané, ide o zovšeobecnenie normálneho rozdelenia tak, aby sa dalo pracovať (a dopracovať sa k štatisticky hodnotným záverom) aj s menšími súbormi dát. Ľudová povest' hovorí, že t-rozdelenie vymyslel kontrolór kvality piva ako zlepšovák, aby sa pri testovaní vzoriek neotvorilo priveľa fliaš a aby kontrolóri ostali dlhšie triezvi.

Faktom je, že t-rozdelenie odvodil W.S. Gosset – naozaj zamestnanec pivovaru (Guinness), a asi kvôli utajeniu know-how pred konkurenciou publikoval výsledky pod pseudonymom Student.

Distribučná funkcia t-rozdelenia sa mení v závislosti od parametra $k=(n-1)$, kde n je počet vzoriek/dát v súbore. Pre rastúce n t-rozdelenie konverguje k normálnemu rozdeleniu $N(0,1)$. V starších tabuľkách sa preto uvádzali hodnoty len pre k do 30, prípadne 50. Pri vyšších hodnotách sa už prešlo na normálne rozdelenie.

chí-kvadrát / χ^2

– v tabuľkách alebo cez online kalkulačky

https://en.wikipedia.org/wiki/Chi-square_distribution

Rozdelenie chí-kvadrát má využitie pri odhadovaní rozptylu. Podobne ako pri t-rozdelení je závislé od parametra k (stupeň voľnosti), ktorý súvisí s počtom dát. V tabuľkách sa uvádzajú hodnoty pre k väčšinou do 30 alebo 50. Pre rastúce k rozdelenie χ^2 konverguje k normálnemu rozdeleniu $N(k,2k)$.

Bodový odhad

Bodový odhad je najlepší možný odhad neznámeho parametra rozdelenia. Zo štatistického hľadiska však nie je jasné, do akej miery sa mu dá veriť.

Intervalový odhad

Intervalový odhad vymedzuje interval, v ktorom sa neznámy parameter nachádza, a zároveň určuje spoľahlivosť, tj. pravdepodobnosť, nakoľko sa tomu odhadu dá veriť. Je vcelku pochopiteľné, že čím chceme byť presnejší (užší interval), tým menej sa nám dá veriť, a naopak, čím vyššiu spoľahlivosť odhadu chceme garantovať, o toľko širším intervalom sa poistíme.

Riešené príklady

Príklad 1:

Náhodná veličina má normálne rozdelenie $N(\mu, \sigma^2)$ s neznámymi parametrami. Pri 10 pokusoch dostaneme dátový vektor

$$\mathbf{x} = [3 \ 4 \ 7 \ 8 \ 9 \ 11 \ 12 \ 13 \ 16 \ 17]$$

Nájdite intervalové odhady a) strednej hodnoty a b) rozptylu so spoľahlivosťou 0.9 a 0.99.

Riešenie:

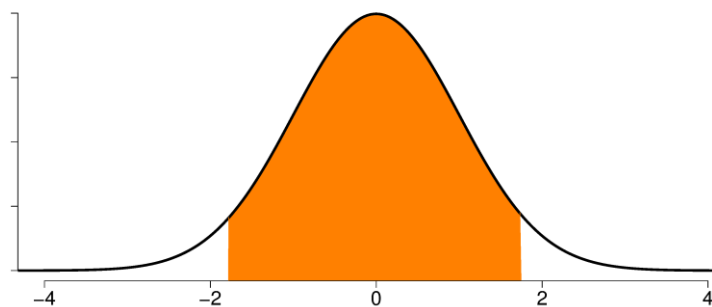
Aritmetický priemer čísel v $\mathbf{x}$ je $m=10$. Je to bodový odhad neznámej strednej hodnoty.

Bodový odhad rozptylu: $s^2 = 1/(n-1) (\sum_{i=1, \dots, 10} (x_i - m)^2) = 22$

Pri výpočte delíme číslom $(n-1)$, keďže dátový vektor je len výberom z celku, nie celok sám (preto tiež m nie je stredná hodnota, len jej odhad).

a) Spoľahlivosť 0.9

Vzhľadom na to, čo je k dispozícii, musíme použiť t-rozdelenie s parametrom 9 (lebo $10-1$). Hodnotu 0.9 ako plochu pod grafom funkcie hustoty vyberieme pekne súmerne zo stredu, od nejakého $-u$ po u :



Celá plocha pod krivkou je 1. Oba odignorované „chvosty“ musia mať po 0.05. Pre hľadané u musí preto platiť:

$$-u = t_{0.05, 9} \quad \text{a} \quad u = t_{0.95, 9}$$

V tabuľkách nájdeme $u = t_{0.95, 9} = 1.8331$. Faktor $t_{0.95, 9}$ je fakticky kvantilová funkcia (inverzná k distribučnej).

Ďalej pokračujeme už podľa vzorcov z prednášky.

Nájdene u je „šité na mieru“ pre rozptyl 1. Vypočítame preto hodnotu d , ktorá bude zodpovedať odhadovanému rozptylu v našej úlohe:

$$d = u * s / \sqrt{n} = 2.7189$$

A sme doma. Stredná hodnota veličiny realizovanej vo vektore $\mathbf{x}$ je na 90% v intervale 10 ± 2.7189 . Z posledného vzorca vidno, že menší rozptyl alebo viac pokusov vedia zúžiť interval odhadu.

Spoľahlivosť 0.99

Len zmeníme pár drobností .

$$u = t_{0.995, 9} = 3.2498$$

$$d = 4.82$$

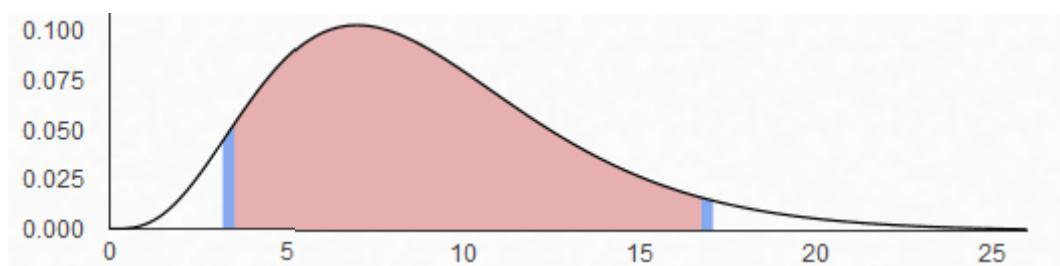
Stredná hodnota veličiny realizovanej vo vektore x je na 99% v intervale 10 ± 3.2498 .
Všimnime si, ako vyšší tlak na spoľahlivosť výsledku vedie nutne k širšiemu intervalu odhadu.

b) Spoľahlivosť 0.9

Bodový odhad rozptylu už poznáme, stačí si pripomenúť :

$$s^2 = 22$$

Budeme pracovať s rozdelením chí-kvadrát, stupeň voľnosti 9 (lebo 10-1).
Spoľahlivosť 0.9 musíme “vystrihnúť” ako plochu pod grafom funkcie hustoty, a to zase pekne zo stredu. Keďže graf nie je symetrický, “zo stredu” tu chápeme iba kvantitatívne, t.j. že oba odignorované chvosty musia mať po 0.05.



Vyfarbená plocha siaha od $u_1 = \chi^2_{0.05, 9} = 3.33$ po $u_2 = \chi^2_{0.95, 9} = 16.92$ (nájdeme v tabuľkách).

Pokračujeme dosadením do základných vzorcov:

$$\sigma^2 \in (L, P)$$

$$\begin{aligned} \text{kde } L &= (n-1) \cdot s^2 / u_2 = 11.7 \\ P &= (n-1) \cdot s^2 / u_1 = 59.46 \end{aligned}$$

Rozptyl skúmanej náhodnej veličiny sa na 90% nachádza v intervale (11.7 , 59.46) .

Spoľahlivosť 0.99

Zmenia sa iba hodnoty u , a to nasledovne (z tabuliek): $u_1 = \chi^2_{0.005, 9} = 1.73$ po $u_2 = \chi^2_{0.995, 9} = 23.59$.
Následne

$$\begin{aligned} L &= 8.39 \\ P &= 114.45 \end{aligned}$$

Rozptyl skúmanej náhodnej veličiny sa na 99% nachádza v intervale (8.39 , 114.45) .
Za vyššie istoty zaplatíme širším intervalom.

Príklad 2a:

Náhodná veličina má normálne rozdelenie $N(\mu, 22)$ s neznámym parametrom μ , ktorého bodový odhad na základe 10 dát je $m=10$.

Nájdite intervalový odhad strednej hodnoty so spoľahlivosťou 0.9.

Riešenie:

Príklad je určený iba na precvičenie vzorcov, v praxi sa ťažko stretnete so situáciou, že máte k dispozícii presnú hodnotu rozptylu, ale nepoznáte presne strednú hodnotu. A už vôbec nie pri 10 údajoch v dátovom vektore.

Ak poznáme presne σ^2 , použijeme normálne rozdelenie. Postup je rovnaký ako vyššie.

$$u = F^{-1}(0.95) = 1.645.$$

$$d = u * s / \sqrt{n} = 2.4399$$

Stredná hodnota veličiny je na 90% v intervale 10 ± 2.44 .

Príklad 2b:

Náhodná veličina má normálne rozdelenie $N(10, \sigma^2)$ s neznámym parametrom σ^2 , ktorého bodový odhad na základe 10 dát je $s^2=22$.

Nájdite intervalový odhad rozptylu so spoľahlivosťou 0.9.

Riešenie:

Budeme pracovať s rozdelením chí-kvadrát, stupeň voľnosti 10 (!! lebo μ máme presné).

Počítame

$$u_1 = \chi^2_{0.05, 10} = 3.94$$

$$u_2 = \chi^2_{0.95, 10} = 18.31$$

Pokračujeme dosadením do základných vzorcov:

$$\sigma^2 \in (L, P)$$

$$\begin{array}{lll} \text{kde} & L = n*s^2/u_2 & = 12.015 \\ & P = n*s^2/u_1 & = 55.838 \end{array}$$

Rozptyl skúmanej náhodnej veličiny sa na 90% nachádza v intervale (12.015 , 55.838) .

Poznámka.

V príklade 2 sme mali rovnaké hodnoty ako v príklade 1. Výsledky sme však dostali presnejšie (menšie intervaly). Je to vďaka tomu, že vstupné hodnoty boli “kvalitnejšie” (presné, nie odhady).

Príklad 3:

Náhodná veličina má normálne rozdelenie $N(\mu, \sigma^2)$ s neznámymi parametrami. Pri 1000 pokusoch dostaneme dátový vektor, z ktorého určíme bodové odhady

$$m = 11, \quad s^2 = 49$$

Nájdite intervalové odhady a) strednej hodnoty a b) rozptylu so spoľahlivosťou 0.95.

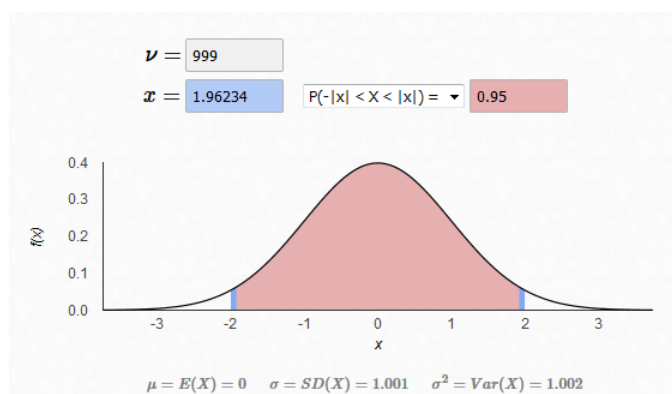
Riešenie:

Ak by sme chceli postupovať tak, ako v predošlom príklade, veľmi rýchlo narazíme na limity tabuliek. Tade cesta nevedie. Východiská sú dve.

I.

Ak máme k dispozícii dobrý softvér alebo pripojenie na internet, potrebné hodnoty si nájdeme tam, na internete je dnes skoro všetko.¹

a) Studentovo t-rozdelenie, stupeň voľnosti 999, 0.975-kvantil :



Máme teda $u = t_{0.975, 999} = 1.96234$.

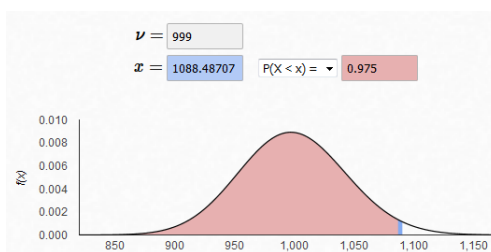
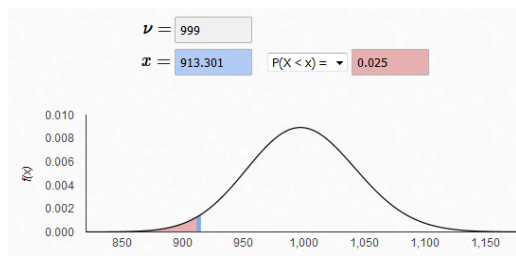
Odtiaľ dostávame $d = 1.96234 * 7 / \sqrt{999} = 0.4346$

Strednú hodnotu teda na 95% odhadujeme v intervale 11 ± 0.4346

b) Počítajme

$$u_1 = \chi^2_{0.025, 999} = 913.301$$

$$u_2 = \chi^2_{0.975, 999} = 1088.48707$$



Odtiaľ máme $L = 999 * 49 / 1088.48707 = 44.975$

$$P = 999 * 49 / 913.301 = 53.598$$

Rozptyl skúmanej náhodnej veličiny sa na 95% nachádza v intervale (44.975 , 53.598) .

¹ Využívam <https://homepage.divms.uiowa.edu/~mbognar/applets/>

II.

Ak nemáme k dispozícii dobrý softvér ani pripojenie na internet, potrebné hodnoty nájdeme aproximáciou pomocou normálneho rozdelenia.

a) Pri tisícke dát si môžeme dovoliť pracovať s $s^2=49$ ako s presnou hodnotou rozptylu.² Vtedy je namieste použiť normálne rozdelenie.

$$u = F^{-1}(0.975) = 1.96$$
$$d = 1.96 * 7 / \sqrt{999} = 0.4341$$

Strednú hodnotu teda na 95% odhadujeme v intervale 11 ± 0.4341

Ak porovnáme tento výsledok s tým presnejším v časti I., vidíme, že rozdiel je zanedbateľný.

a) Namiesto chí-kvadrátu použijeme normálne rozdelenie $N(1000, 2000)$.

Počítajme:

$$F(u_1) = 0.025$$
$$F_n((u_1 - 1000) / \sqrt{2000}) = 0.025 \quad / 0.025 \text{ je menej ako } 0.5, \text{ musíme aplikovať fintu}$$
$$1 - F_n(-(u_1 - 1000) / \sqrt{2000}) = 0.025$$
$$F_n(-(u_1 - 1000) / \sqrt{2000}) = 0.975$$
$$-(u_1 - 1000) / \sqrt{2000} = F_n^{-1}(0.975)$$
$$-(u_1 - 1000) / \sqrt{2000} = 1.96$$
$$u_1 = -1.96 * \sqrt{2000} + 1000 = 912.35$$

$$F(u_2) = 0.975$$
$$F_n((u_2 - 1000) / \sqrt{2000}) = 0.975$$
$$(u_2 - 1000) / \sqrt{2000} = F_n^{-1}(0.975) = 1.96$$
$$u_2 = 1.96 * \sqrt{2000} + 1000 = 1087.7$$

Získané hodnoty $u_{1,2}$ dosadíme:

$$L = 999 * 49 / 1087.7 = 45.004$$
$$P = 999 * 49 / 912.35 = 53.654$$

Rozptyl skúmanej náhodnej veličiny sa na 95% nachádza v intervale (45.004 , 53.654) .

Porovnanie s presným výsledkom z časti I ukazuje vskutku nepodstatné odchýlky.

Poznámka:

Obdivuhodne presné výsledky získané cestou aproximácie vďaka za svoju kvalitu hlavne veľkému množstvu dát (1000). Ak by ich bolo iba 100, odchýlky budú trochu väčšie, avšak stále to bude použiteľné.

² Podľa výpočtov vyššie tam síce je asi 8-percentná nepresnosť, ale na výsledku nižšie sa to nijak tragicky neprejaví.

Neriešené príklady

1. Je daná náhodná veličina $X \sim N(\mu, \sigma^2)$ s neznámymi parametrami. V 12 pokusoch sme získali ako jej konkrétnu realizáciu vektor hodnôt $x = [2 \ 2 \ 3 \ 3 \ 4 \ 4 \ 4 \ 4 \ 5 \ 5 \ 5 \ 7]$.

- a) Nájdite bodový odhad μ, σ^2 . (4, 2)
- b) Nájdite intervalový odhad μ, σ^2 so spoľahlivosťou 0.9 a 0.95.
(pre 0.9: 4 ± 0.7332 , [1.1179, 4.8140])
- c) Akú spoľahlivosť má intervalový odhad μ s dĺžkou intervalu 2.6 ? (0.99)
- d) Akú spoľahlivosť má intervalový odhad σ^2 , ak horná hranica intervalu je desaťnásobkom dolnej hranice? (0.99)

2. Je daná náhodná veličina $X \sim N(\mu, \sigma^2)$. V 12 pokusoch sme získali ako jej konkrétnu realizáciu vektor hodnôt x .

- a) Nájdite intervalový odhad μ so spoľahlivosťou 0.9 a 0.95, ak viete, že $\sigma^2 = 2$ (presne) a $m = 4$ (bodový odhad).
- b) Nájdite intervalový odhad σ^2 so spoľahlivosťou 0.9 a 0.95, ak viete, že $\mu = 4$ (presne) a $s^2 = 2$ (bodový odhad).

3. Je daná náhodná veličina $X \sim N(\mu, \sigma^2)$ s neznámymi parametrami. V 144 pokusoch sme získali ako jej konkrétnu realizáciu vektor hodnôt x s bodovým odhadom strednej hodnoty $m=5$ a variancie $s^2 = 9$

- a) Nájdite intervalový odhad μ, σ^2 so spoľahlivosťou 0.9, 0.95 a 0.99.
(pre 0.9: 5 ± 0.41125 , [7.487, 11.057])
- b) Akú spoľahlivosť má intervalový odhad μ s dĺžkou intervalu 0.4? (0.5762)
- c) Akú spoľahlivosť má intervalový odhad σ^2 , ak horná hranica intervalu je dvojnásobkom dolnej hranice? (0.995)