

Peatükk 5

Prognoosiintervall ja Usaldusintervall

5.1 Prognoosiintervall

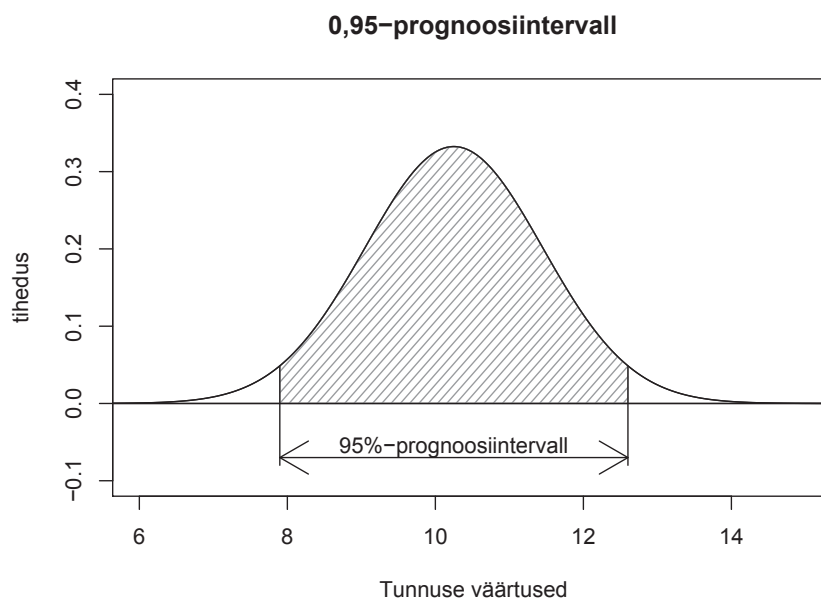
Unustame hetkeks populatsiooni parameetrite hindamise ja pöördume tagasi üksikvaatluste juurde.

On raske ennustada, milline on huvipakkuva tunnuse väärtus järgmisel juhuslikult populatsioonist valitud isendil. Samas on võimalik leida vahemik, millese järgmine vaatlus satub suure tõenäosusega. Näiteks, kui on teada uuritava tunnuse jaotus populatsioonis (teame sigade kaalu tihedusfunktsiooni), siis võib valida vahemiku, kuhu uuritava tunnuse väärtus (ühe juhuslikult valitud sea kaal) sattub mingi suure etteantud tõenäosusega. Alltoodud graafikul 5.1 on konstrueeritud selline prognoosintervall, kuhu järgmine üksikvaatlus sattub tõenäosusega 0,95 (95%-prognoosintervall).

Kuidas prognoosipiire reaalselt saaks leida? Üks võimalus on muidugi kasutada sobivalt valitud kvantiile — 0,025-kvantiili ja 0,975-kvantiili vahele jääb uuritava tunnuse väärtus tõenäosusega $0,975 - 0,025 = 0,95$, seega on nende kahe kvantiili poolt määratud vahemik 0,95-prognoosintervall. Üldjuhul läheks meil $(1 - \alpha)$ -prognoosintervalli leidmiseks vaja teada $\alpha/2$ ja $(1 - \alpha/2)$ -kvantiile — nende kahe kvantiili vahele jääb juhusliku suuruse väärtus täpselt tõenäosusega $(1 - \alpha)$.

Vaatame juhtu, mil uuritava tunnuse jaotuseks on standardne normaaljaotus (selline normaaljaotus, mille puhul $\sigma^2 = 1$ ja $\mu = 0$). Sellisel juhul võime tabelist 5.1 välja lugeda soovitud (standardse normaaljaotuse) kvantiilid ning leida soovitud prognoosintervalli. Märkus: standardse normaaljaotuse α -kvantiili tähistatakse traditsiooniliselt sümboliga z_α .

Joonis 5.1: 0,95-proгноосиинтервалл

Tabel 5.1: Standardse normaaljaotuse kvantiile z_α

α	0,005	0,025	0,05	0,5	0,95	0,975	0,995
z_α	-2,58	-1,96	-1,64	0	1,64	1,96	2,58

Seega standardse normaaljaotusega juhusliku suuruse korral oleks 0,95-proгноосиинтервалл $(-1,96 \dots 1,96)$; 0,9-proгноосиинтервалл $(-1,64 \dots 1,64)$ ja 0,99-proгноосиинтервалл $(-2,58 \dots 2,58)$.

Uuritavaid tunnuseid, mille jaotuseks oleks täpselt standardne normaaljaotus, esineb tavaelus haruharva. Küll aga esineb normaaljaotusega juhuslikke suuruseid, ja mitte harva.

Kuidas leida прогноосиинтервали normaaljaotusega juhuslikule suurusele X , mille keskvärtus $EX = \mu$ ja dispersioon on $DX = \sigma^2$, ehk teisisõnu, mille jaotuseks on $X \sim N(\mu, \sigma^2)$?

Esmalt märkame, et teisendatud juhusliku suuruse $X_{uus} := \frac{X - \mu}{\sigma}$ keskvärtuseks on 0 ja dispersiooniks 1:

$$EX_{uus} = \frac{1}{\sigma}(EX - \mu) = \frac{1}{\sigma}(\mu - \mu) = 0;$$

$$DX_{uus} = D\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma^2}D(X - \mu) = \frac{1}{\sigma^2}\sigma^2 = 1.$$

Kuna normaaljaotusega juhuslikule suurusele konstandi juurdeliitmisel või konstandiga korrutamisel saame ikka normaaljaotusega juhusliku suuruse, siis järelikult on uue, teisendatud juhusliku suuruse X_{uus} jaotuseks standardne normaaljaotus ja 95% tema väärtustest jääb vahemikku $-1,96..1,96$:

$$\begin{aligned} P(-1,96 \leq X_{uus} \leq 1,96) &= 0,95 \\ P(-1,96 \leq \frac{X - \mu}{\sigma} \leq 1,96) &= 0,95 \\ P(-1,96\sigma \leq X - \mu \leq 1,96\sigma) &= 0,95 \\ P(\mu - 1,96\sigma \leq X \leq \mu + 1,96\sigma) &= 0,95 \end{aligned}$$

Järeldus: Kui juhusliku suuruse X jaotuseks on normaaljaotus, $X \sim N(\mu, \sigma^2)$, siis tema 0,95-prognoosiintervall on leitav järgmise valemiga:

$$(\mu - 1,96\sigma \dots \mu + 1,96\sigma).$$

Loomulikult võime sama arutluskäiku korrata, otsides teisendatud juhuslikule suurusele X_{uus} mingit muud, näiteks $1 - \alpha$ -prognoosiintervalli. Tulemuseks saame:

$$(\mu + z_{\alpha/2}\sigma \dots \mu + z_{1-\alpha/2}\sigma). \quad (5.1)$$

Seega saame standardse normaaljaotuse kvantiile teades leida suvalist prognoosiintervalli suvalise normaaljaotusega juhuslikule suurusele — senikaua kuni teame meid huvitava juhusliku suuruse keskväärtust ja standardhälvet.

Märkus: Mis saab siis, kui me ei tea keskväärtust ja populatsiooni standardhälvet σ ? Suure valimi korral võime muidugi teha näo ja väita, et meie valimi keskmine ja standardhälve on väga-väga täpsed hinnangud populatsiooni parameetritele, peaaegu võrdsed nendega, ja seega võime kasutada ka valemit 5.1, asendades vaid μ valimi keskmisega $\bar{x}$ ja σ valimi standardhällbega s . Väiksemates valimites võib erinevus valimi põhjal saadud hinnanguite $(\bar{x}, s)$ ja populatsiooni väärtuste (μ, σ) vahel olla siiski märkimisväärne. Sellisel juhul tuleks $(1 - \alpha)$ -prognoosiintervall normaaljaotusega juhuslikule suurusele leida kasutades valemit:

$$\left(\bar{X} + t_{\alpha/2;n-1}S\sqrt{1 + \frac{1}{n}} \dots \bar{X} + t_{1-\alpha/2;n-1}S\sqrt{1 + \frac{1}{n}}\right). \quad (5.2)$$

Kus $t_{\alpha/2;n-1}$ ja $t_{1-\alpha/2;n-1}$ on vastavalt t-jaotuse $\alpha/2$ ja $1 - \alpha/2$ -kvantiilid. Tasub ehk ära märkida, et kui prognoosiintervall 5.1 on tõlgendatav kui vahemik, kuhu vahele jääb $(1 - \alpha) \cdot 100\%$ vaatlustest, siis valemiga 5.2 kirjeldatud prognoosivahemik näitab küll vahemikku, kuhu sattub järgmine vaatlus tõenäosusega $(1 - \alpha)$, aga kuhu ei pruugi jääda $(1 - \alpha)$ osa kõigist tulevastest vaatlustest (ligikaudu on väide siiski õige, aga mitte päris täpselt)!

5.2 Usaldusintervall

Ka punkthinnangud on juhuslikud suurused — sest valim on juhuslik. Iga uurija, kes üritab vastata samale (populatsiooni puudutavale) küsimusele, saab veidi teistest erineva vastuse. Kui lähemalt uurida selle juhusliku suuruse - punkthinnangu - jaotust, siis selgub üllatavalt sageli, et tegemist on normaaljaotusega. Näiteks on vähegi suurema valimi korral (kümnekond või enam vaatlust) valimi keskmise jaotuseks normaaljaotus:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Seega on teisendatud juhusliku suuruse $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ jaotuseks standardne normaaljaotus. Standardse normaaljaotuse korral aga oskame leida vahemikku, kuhu vahele standardse normaaljaotusega juhusliku suuruse väärtus peab sattuma tõenäosusega 0,95:

$$\begin{aligned} P\left(-1,96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1,96\right) &= 0,95 \\ P\left(-1,96 \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq 1,96 \frac{\sigma}{\sqrt{n}}\right) &= 0,95 \\ P\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right) &= 0,95 \end{aligned}$$

Seega 95% juhuslike valimite korral jääb populatsiooni tegelik keskväärtus vahemikku

$$\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} \dots \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right).$$

Antud juhul tasub tähele panna järgmist: juhuslik pole mitte populatsiooni tegelik keskväärtus μ , vaid leitud vahemik: iga uurija võib teisest saada veidi erineva vahemiku. Ühte konkreetset valimit (ja valimi keskmist) kasutades leitud vahemik kas sisaldab või ei sisalda populatsiooni keskväärtust.

Tõenäosusest konkreetse, väljaarvutatud intervalli kontekstis enam rääkida ei saa. Küll aga saab rääkida, et kasutasime arvutusmetoodikat, mis 95%-l juhtudest annab populatsiooni tegelikku keskväärtust sisaldava vahemiku ja seega võime 95%-se kindlusega (*confidence*) väita, et tegelik keskväärtus asub antud vahemikus. Vastavat vahemikku kutsutakse 95%-usaldusintervalliks või 95%-usaldusvahemikuks.

Hakates oma uuringu tarvis taolist 95%-st vahemikku leidma, põrkume aga raskuse otsa — me ei tea ju populatsiooni standardhälvet σ . Esimene mõte, mis pähe võiks tulla, oleks järgmine — asendame σ tema hinnanguga, valimi standardhälbe s -ga. Selgub, et väga suurte valimite korral (kus s ja σ nagunii üsna sarnased tulevad) võib seda tõepoolest teha. Väiksemate valimite korral ($n < 60$) paraku nii toimida ei tohi. Lahendus on siiski olemas.

Kui juhuslik suurus $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ oli standardse normaaljaotusega juhuslik suurus, siis asendades populatsiooni standardhälbe σ valimi standardhällbega S saame (Studenti) t -jaotusega juhusliku suuruse:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Hea sõnum seisneb selles, et t -jaotus on hästi teada, t -jaotuse kvantiilid on ära toodud igas endast lugupidavas statistikaalases raamatus ja seega kõik mis me ülaltoodud arutelus muutma peame, on see, et peame standardse normaaljaotuse kvantiilid asendama t -jaotuse vastavate kvantiilidega. Halvem on see, et t -jaotuseid on palju. Iga valimi suuruse korral on meil tegemist teistest veidi erineva t -jaotusega. T -jaotuse määrab üheselt ära t -jaotuse parameeter - vabadusastmete arv (degrees of freedom — d.f.). Kvantiilid erinevate t -jaotuste tarvis on ära toodud tabelis 5.2.

Peale standardse normaaljaotuse kvantiilide asendamist t -jaotuse kvantiilidega jõuame järgmise usaldusintervalli arvutamise valemieni. $(1 - \alpha)$ -usaldusintervall keskväärtusele on leitav järgmise valemiga:

$$\left(\bar{X} + t_{n-1;\alpha/2} \frac{S}{\sqrt{n}} \dots \bar{X} + t_{n-1;1-\alpha/2} \frac{S}{\sqrt{n}} \right).$$

Näide 5.1 Tuntud kanauurijat Hans Hane huvitab, mitu muna munevad keskmiselt eesti kanad ühe nädala jooksul. Kogumikus kanade kohta kavatses ta esitada 95%-lise usaldusintervalli munade arvu keskväärtusele. Usaldusintervalli arvutamiseks luges härra Hani nädala jooksul kokku 10 kana munad: 5 5 4 6 1 7 5 6 3 3.

Lahendus.

Kõigepealt leidis Hans valimi keskmise ja valimi standardhälbe: $\bar{x} = 4,5$; $s = 1,78$. Seejärel leidis ta tabelist 5.2 arvutustes vajaminevad t -jaotuse kvaantiilid: $t_{0,025;9} = -2,26$ ja $t_{0,975;9} = 2,26$. Asetades leitud arvud valemisse sai ta tulemuseks:

$$\begin{pmatrix} \bar{x} + t_{n-1;\alpha/2} \frac{s}{\sqrt{n}} & \dots & \bar{x} + t_{n-1;1-\alpha/2} \frac{s}{\sqrt{n}} \\ 4,5 + (-2,26) \frac{1,78}{\sqrt{10}} & \dots & 4,5 + 2,26 \frac{1,78}{\sqrt{10}} \\ (4,5 + (-2,26)0,56 & \dots & 4,5 + 2,26 \times 0,56) \\ (3,23 & \dots & 5,77) \end{pmatrix}$$

Vastus: 95%-usaldusintervall eesti kanade nädala jooksul munetud munade arvu keskväärtusele on $(3,23 \dots 5,77)$.

Kui soovime täpsemalt teada, milline tegelik keskväärtus olla võiks, tuleb suurendada valimi suurust. Mida suurem valim, seda kitsam tuleb ka usaldusintervall keskväärtusele.

5.3 Ülesanded

1. Uuriti sordi A saagikust. Keskmiseks saagikuseks saadi 1,2 tonni/ha; standardhälve oli 0,5. Milline on 95%-usaldusintervall sordi A saagikuse keskväärtusele, kui a) uuringu tulemused on saadud 8 põllu tulemuste põhjal ($n=8$) või b) uuringu tulemused on saadud 100 põllu tulemuste põhjal ($n=100$).
2. Kumb on laiem, kas 90% või 95%-usaldusintervall? Miks?
3. Kaks teadlast uurisid sordi B saagikust Eesti oludes. Nende kahe teadlase poolt leitud 95%-sed usaldusintervallid sordi B saagikuse keskväärtusele ei kattunud. Loetle põhjuseid, miks võidi samale küsimusele vastates saada mittekattuvad usaldusintervallid?
4. Üks teadlane mõõtis sordi C saagikust 8-l juhuslikult valitud põllul eestis. Teine teadlane mõõtis sordi C saagikust 8-l juhuslikult valitud katselapil ühel katsepõllul. Kumb teadlastest sai sordi C-saagikuse keskväärtusele laiem 95%-usaldusintervalli, miks? Kuidas tuleks kumagi teadlase poolt leitud usaldusintervalli interpreteerida?

Tabel 5.2: Studenti t-jaotuse kvantiilide t_α väärtused (nn kriitilised väärtused)

d.f. = n-1	$\alpha = 0,01$	$\alpha = 0,025$	$\alpha = 0,05$	$\alpha = 0,95$	$\alpha = 0,975$	$\alpha = 0,99$
1	-31,82	-12,71	-6,31	6,31	12,71	31,82
2	-6,97	-4,30	-2,92	2,92	4,30	6,97
3	-4,54	-3,18	-2,35	2,35	3,18	4,54
4	-3,75	-2,78	-2,13	2,13	2,78	3,75
5	-3,37	-2,57	-2,01	2,01	2,57	3,37
6	-3,14	-2,45	-1,94	1,94	2,45	3,14
7	-3,00	-2,36	-1,89	1,89	2,36	3,00
8	-2,90	-2,31	-1,86	1,86	2,31	2,90
9	-2,82	-2,26	-1,83	1,83	2,26	2,82
10	-2,76	-2,23	-1,81	1,81	2,23	2,76
12	-2,68	-2,18	-1,78	1,78	2,18	2,68
14	-2,62	-2,14	-1,76	1,76	2,14	2,62
16	-2,58	-2,12	-1,75	1,75	2,12	2,58
18	-2,55	-2,10	-1,73	1,73	2,10	2,55
20	-2,53	-2,09	-1,73	1,73	2,09	2,53
25	-2,49	-2,06	-1,71	1,71	2,06	2,49
30	-2,46	-2,04	-1,70	1,70	2,04	2,46
40	-2,42	-2,02	-1,68	1,68	2,02	2,42
60	-2,39	-2,00	-1,67	1,67	2,00	2,39
120	-2,36	-1,98	-1,66	1,66	1,98	2,36
∞	-2,33	-1,96	-1,64	1,64	1,96	2,33

5. 4-aastase kitse kaalu keskväärtus on 70kg, kaalu standardhälve on 3. Eeldades, et kitse kaal on normaaljaotusega juhuslik suurus (miks on see mõeldav eeldus?), leia 95%-prognoosintervall 4-aastaste kitsede kaalule.
6. uuritava tunnuse keskväärtus on 4, standardhälve on 1. Milline on 95%-usaldusintervall keskväärtusele?

