**Katseandmete analüüs**

**Loeng 2. Statistiline test. T-test.**

**Kirjutas Toomas Tammaru veebruaris 2002, viimati sätistatud oktoobris 2016; konsultant Ants Kaasik**

**Statistiline test**

Tuletame meelde, et kasutada on meil **valim**, järeldusi tahame teha **üldkogumi** kohta. Enamasti on nii, et näeme silmaga, et meid huvitav erinevus või seos on valimis olemas. **Tahame teada, kas oma valimi ja selles nähtud seose põhjal saame väita, et selline seos on olemas ka üldkogumis.** Sellisele küsimusele **statistiline test** vastabki.

*Natuke filosoofiat*. Üldkogum on kas reaalne (nt reaalne loomapopulatsioon) või mõnikord ka abstraktne (nt kõik hiired mis kunagi elanud on ja veel elama hakkavad, kõikvõimalikud evolutsioonilised muutused).On erandjuhtusid, kus me saame läbi uurida terve üldkogumi, nii näiteks võime küsida, kas saarepuu liikides on rohkem tanniine kui tammedes. Põhimõtteliselt on jõukohane läbi uurida kõik saareliigid ja kõik tammeliigid ja vaadata, kuidas asjaga lugu on. Nii jääbki. Statistilist testi pole niisuguse nn. kõikse valimi puhul läbi viia mõtet. *Aga see selleks, mis kohe järgneb, on hulka olulisem.*

*Saa sellest p tõlgendamise asjast korralikult aru – oluline (ja eksamil küsitakse ka kindlasti). Kui ei saanud, loe mitu korda, kui ikka ei saa, küsi nõu.*

Statistilise testi väljundina leitakse arvuline väärtus nimega **statistiline olulisus (**või **olulisustõenäosus, *significance*), mida tähistatakse** mõnikord suure, mõnikord väikese **p tähega**. Lihtne sissejuhatav seletus (*formaalsemate seletuste sissejuhatamiseks, aga ei piisa eksamil vastamiseks*): **p näitab, kui tõenäone oleks saada valimis nähtav olukord juhuslikult.**

Nüüd siis formaalsemalt: olgu meil konkreetne reaalne valim ja olgu selles tuvastatud miski tugevusega seos. **Statistiline olulisus väljendab tõenäosust saada** (valimi võtmise käigus üldkogumist) vähemalt nii suure erinevusega või vähemalt nii tugeva **seosega valim juhul, kui üldkogumis seda seost või erinevust tegelikult ei ole.** See ’nii suur’ ja ’nii tugev’ tähendab siis nii suurt või nii tugevat, nagu see uuritud valimis leiti olevat.

Ehk siis - kui võtame üldkogumist juhuslikult väikese valimi, on igati tõenäone juhtuma, et selles valimis esineb puhtjuhuslikult nn valimivea tõttu mingi seos, mida üldkogumis pole. Loogika on siis selline, et kui tõenäosus saada nii tugeva (nagu ta meil reaalses praegu uuritavas valimis on) seosega valim ilma seoseta üldkogumist on piisavalt väike, siis loetakse „statistika mõttes tõestatuks“ (mis ei tähenda siiski täit kindlust, vt veidi allpool), et üldkogumis meid huvitav erinevus või seos on. Igati mõistlik ju nii mõelda: kui tõenäosus saada see, mida me näeme, puhtalt juhuse läbi on väga väike, on ju mõistlik uskuda, et see, mida näeme, ei ole juhuse läbi saadud – ehk siis uskuda, et üldkogumis miski seos tõesti ka on!

Ehk siis “p=0,027” tähendab seda, et 2,7% on tõenäosus saada vähemalt nii tugev seos puhtjuhuslikult olukorras, kus üldkogumis mingit seost ei ole. Ehk siis tõenäosuse mõiste lahti seletatult: kui võtame pimesi tuhat valimit sellisest üldkogumist, kus seost tegelikult ei ole, siis keskmiselt 27-s neist saame nii tugeva (või tugevama) seose nagu see meie reaalses valimis seekord on.

**Ehk siis mida väiksem p, seda (statistiliselt) olulisem seos.** Siiski siiski - täit kindlust ei saavuta me kunagi, ehk siis päris null p olla ei saa, p võimalikud väärtused jäävad vahemikku 1p>0. Teisisõnu, alati on võimalik, et vägagi tugev seos meie valimis on lihtsalt juhuse tulemus ja teeme sisuliselt vale järelduse üldkogumi kohta (esimest liiki viga, statistika teooria terminoloogias, jutuks viimases loengus). Tuleb leppida teadmisega, et statistilist testi tehes võime ka eksida.

Tavaliselt loetakse ja nimetatakse olulisteks seoseid, mille puhul p<0,05. Enamasti siis nii, et **kui p<0,05, öeldakse, et valimis nähtud seos on (statistiliselt) oluline (*significant*)**, p0.05 korral seos ei ole (statistiliselt) oluline või siis võib öelda, et on mitteoluline (*non-significant,* mitte *~~insignifcant~~*).

See 0,05 on tegelikult puhtalt kokkuleppeline piir, sellel arvul ei ole mingit sügavamat filosoofilist põhjendust. Siiski on selle 0,05 kasutamine väga laialt juurdunud ja tulemust, kus meid huvitava seose p=0,051, on oluliselt raskem teadusajakirjas avaldada kui sellist, kus p=0,049 - kuigi sisuline vahe on tegelikult tühine. Juurdunud on praktika, et järelduste tegemisel kasutame vaid selliseid seoseid, mis tulid statistilistes testides olulisteks. Enamasti on siis asi teadustegevuse praktikas niipidi, et väga tahetakse, et uuritavaid seosed tuleksid testides olulisteks ja ollakse kurvad, kui nad mitte ei tule.

Pane veel väga tähele, et p on tõenäosus **saada** valimis nähtud seos juhuslikult; p ei ole tõenäosus, et seos **saadi** juhuslikult (või et seos **on** juhuslik), need on sisuliselt täiesti erinevad väited. Ehk kui näiteks p=0,027, siis sellised tõlgendused, et „2,7% tõenäosusega saadi erinevus juhuslikult“ või et „2,7% tõenäosusega on seos juhuslik“ on põhimõtteliselt vale. Väide „2,7% tõenäosusega saadi juhuslikult“ oleks ju samaväärne väitega, et „100-2,7=97,3% tõenäosusega ei saadud juhuslikult“, ja kui juhuslikult ei saadud, siis see ju tähendab, et „97,3% tõenäosusega on erinevus üldkogumis päriselt olemas“. Kuigi me sisuliselt ehk tahaksime teada eelkõige just seda, et millise tõenäosusega erinevus üldkogumis päriselt on või teda ei ole, ei saa me sellist asja kuidagimoodi arvutada, kui meil pole kasutada muud infot peale meie mõõdetud ja uuritud valimi (*kursuse lõpus mainime eriolukordi, kus lisaks veel ka muud infot kaasates saab, see on Bayesi statistika*).

Pane hästi tähele, et **p ei väljenda seose tugevust!** Sellist väärarusaamist tundub algajail tegijail sageli ette tulevat. Seose tugevust väljendavad muud näitajad, iga seose (analüüsi) puhul omad. Nii näiteks regressioonanalüüsi (tuleb varsti jutuks) puhul saab seose tugevust väljendada kahel eri moel (st kaks sisuliselt erinevat parameetrit) - sirge tõusuga ja determinatsioonikordajaga (näitab, kui suure osa varieeruvusest mudel seletab, selgitakse järgmistes loengutes), nendele asjadele pöörame iga analüüsi tutvustuse juures tähelepanu.

Pane veel tähele, et seose või erinevuse **puudumist** üldkogumis **ei saa tõestada**, st ei saa tõestada, et seose tugevust väljendav parameeter on täpselt null, nagu ei saa ka tõestada, et on täpselt 3,73. Kui me ei saanud seost oluliseks (**p>0,05**), siis tulemust tuleb tõlgendada nii, et **meie valimi põhjal ei õnnestunud seose olemasolu tõestada** – kui oleksime võtnud 10 korda suurema valimi, ehk oleksime siis saanud p alla 0,05: võib olla, ei tea. Küll on võimalik statistika meetoditega näidata, et “seos ei ole tugevam kui...”, selleks on parameetrite usalduspiirid arvutatavad ja testi võimsuse analüüs selleks (*kursuse lõpupoole tuleb põgusalt jutuks*).

Ja ikka veel pane tähele, et reaalsuses (üldkogumis, looduses) ole (statistika mõttes) olulisi ja mitteolulisi seoseid, looduses seos kas on või teda ei ole, **statistiline olulisus iseloomustab meie teadmist seosest, mitte seost ennast.**

**Statistiline olulisus sõltub:**

**- seose tugevusest -** st küll kahtlemata sõltub sellest *ceteris paribus*, aga ei mõõda seda, sest sõltub veel muudest asjadest ka; mida tugevam seos valimis, seda väiksem p;

**- valimi suurusest** (mida suurem valim, seda väiksem p)**.**

**- juhusliku varieeruvuse hulgast** (mida vähem varieeruvust, seda väiksem p – raske seletada üldjuhul, selgem konkreetsetes näidetes nagu t-test ja regressioonanalüüs, vt seal)**;**

Iga olukorra jaoks on siis omad meetodid, kuidas seda p-d arvutada, neist hakkamegi siis edaspidi rääkima.

**Vabadusaste** **(*degree of freedom*, *df*)** on statistiliste testide juures üldiselt ette tulev arusaamatu mõiste. Üldiselt näitab süsteemi vabadusaste, kui mitme sõltumatu arvuga on süsteem täielikult kirjeldatav. Nii on kolmnurgal kolm vabadusastet - kui teame kas kolme külge või kahte külge ja ühte nurka, siis on kõik kolmnurga parameetrid arvutatavad. Kolme nurga teadmisest seevastu ei piisa (teame küll kuju, aga mitte suurust), sest siin on tegelikult vaid kaks sõltumatut arvu - kolmas on arvutatav, kuna summa on konstantne.

Statistilise testi kontekstis on olukord täielikult kirjeldatav, kui teame **mudelit** (mida siis sinna sobitame, näiteks regressioonsirget koos oma parameetritega; ehk siis mudelit võib vaadelda kui üldistatud **kvantitatiivset kirjeldust meid huvitavast seosest,** enamasti on selleks võrrand) ja teame iga vaatluse hälvet mudeli poolt ennustatud väärtusest (saad paremini aru peale regressioonanalüüsi osa õppimist). Mudel on kirjeldatav miski hulga arvudega, ehk siis mudelil on vabadusastmed (*model df)*. Samamoodi on hälbed mudelist kirjeldatavad miski hulga arvudega (sõltuvalt sellest, palju noid hälbeid on ja hälbeid on ju samapalju kui üksikvaatlusi), ka nende puhul saab siis rääkida vabadusastmetest (*error df)*. Mõlemad vabadusastmed on tähtsad statistiliste testide tegemisel - mida rohkem on mudelil vabadusastmeid, seda tõenäosem on, et ta sobib andmetega juhuslikult. Seevastu mida rohkem on hälvetel vabadusastmeid (st mida rohkem on üksikvaatlusi), seda vähem tõenäone on, et miski mudel andmetega puhtjuhuslikult sobib. Seega on neid va vabadusastmeid vaja teada selle üle otsustamisel, kas mudeli kooskõla andmetega on puhta juhusega kirjeldatav või mitte - seega siis alati statistilise olulisuse määramisel ehk statistiliste testide tegemisel.

**Konkreetsete testide juurde**

**... ehk statistika valdamine rakendustasemel tähendabki oskust valida olukorrale vastav test (neid on väga palju erinevaid) ja selle tulemusi õigesti interpreteerida. Seda nüüd õpimegi.**

**Statistilise testi valimisel** pane kõigepealt erilise hoolega tähele seda, kas sõltuv ja sõltumatu muutuja on pidevad (väärtused saadud mõõtmise tulemusena, võivad saada mistahes väärtusi vähemalt miskis vahemikus) või diskreetsed (= kategoorilised) (saadud klassidesse jaotamise või loendamise tulemusena), vt esimest loengut. Sõltuv muutuja (enamasti üks) on selline, mille väärtusi tahame seletada (näiteks küüliku kaal), sõltumatud muutujad on sellised, millest me sõltuva muutuja sõltuvust uurime (et kas kaal sõltub 1) söödud toidu hulgast, b) joodud vee hulgast – need kaks on siis sõltumatud muutujad). Iga testi puhul ei ole jaotust sõltuvaks ja sõltumatuks muutujaks (sellest hiljem).

Vastavalt mainitud vahetegemisele jaotuvad testid nii:

|  |  |  |  |
| --- | --- | --- | --- |
| Sõltumatu muutuja | sõltuv muutuja | analüüsid | |
| diskreetne | pidev | t-test, ANOVA |
| diskreetne | diskreetne | sagedustabelite analüüsid: hii-ruut, log-lineaarsed |
| pidev | pidev | korrelatsioonid, regressioonid |
| pidev | diskreetne | logistiline regressioon |

Kõigepealt siis sellistest analüüsidest, kus sõltumatu muutuja on diskreetne ja sõltuv muutuja pidev, ehk siis kuidas sõltub miski mõõdetava tunnuse väärtus klassist, kuhu objekt kuulub.

T-test on kõige lihtsam meetod uurida pideva tunnuse väärtuste erinevust rühmiti, t-testi puhul on neid rühmi kaks. Pane tähele, et rühm võib olla meist sõltumatult olemas (nt kaks populatsiooni) või uurija tekitab need kaks rühma näiteks miskit manipulatsiooni (*treatment*) rakendades. Näiteks söödab loomi kahe erineva toiduga ja mõõdab kasvukiirusi, sellisel juhul on rühmade erinevuste uurimine võrdväärne manipulatsiooni mõju uurimisega. **T-testi tehes küsime siis** (vastava p-väärtuse leidmise teel)**, kas vaadeldav (valimis esinev) kahe rühma keskmiste erinevus on statistiliselt oluline ehk** siis ei ole juhusega seletatav. Ehk siis kui suure tõenäosusega võiksime saada oma valimisse nii suure erinevuse nagu ta meil parasjagu on (või veelgi suurema) sellises olukorras, kus üldkogumis erinevust ei ole.

Küsimusele vastamaks arvutatakse olemasoleva valimi põhjal **t-statistik,** mis siis hindab selle va erinevuse olemasolu usutavust; t-statistik arvutatakse vaat sellise valemi järgi:



milles x1 ja x2 (ülakriipsuga) on kahe valimi keskmised, n1 ja n2 vastavate valimite suurused (st mitu objekti on) ja



See on kahe valimi põhjal hinnatud dispersioonide (s21 ja s22) pealt kokku arvutatud üks ühine hinnang üldkogumi dispersioonile.

Valemist on lihtne näha, et t on seda suurem,

- mida suurem on valimite keskmiste vahe;

- mida suurem on valim;

- mida väiksem on dispersioon valimites.

Edasi käib analüüs nii, et statistiku t väärtust minnakse võrdlema asjakohase **tabeliga**, mille põhjal saab **leida** vastavad **p-väärtused**. Jällegi, selline statistiku arvutamine ja vastava tabeliga võrdlemine p-väärtuse leidmiseks on üldine statistilistes analüüsides (et miks tabelist? - tollepärast, et sõltuvus t ja p vahel ei ole selline, mida saaks tavalise matemaalise seosena esitada ja tavalise valemi abil välja arvutada). Tabeliga võrdlemist teostab tänapäeval muidugi arvuti, varem tehti seda käsitsi. Tabelist p leidmiseks on lisaks t väärtusele vaja teada ka **vabadusastmeid**. T-testi puhul on need vabadusastmed hälvete vabadusastmed, mudeli omi on t-testi puhul alati 1 (kahe rühma vaheline erinevus on väljendatav ühe arvuga) ja neid pole mõtet kirja panna. Hälvete vabadusastmeid on t-testi puhul df=n1+n2-2: kahe võrra vähem kui kahe võrreldava valimi suuruse summa.

*Kes loeb seda teksti muus kui kursusel osalemise kontekstis, pangu tähele, et jutt t-testi eeldustest on järgmise teema juures.*

Sisuliselt võrreldakse t-testis rühmadevahelisi erinevusi juhusliku varieeruvusega, see on väga üldine põhimõte statistilises testis. Küsitakse, kas juhuhälbed on nii suured, et nad oleksid võinud tekitada leitud erinevused eri rühmadest võetud valimite keskmiste vahel.

**Tulemuse esitamine**. Kirjutame nii:

“toidutaime liigil oli mõju rööviku kasvukiirusele (t=2,17; df=34; p=0,025)” .

... ja veel natuke illustreerimist:

Loengus saavad toodud illustreerivad näited veenmaks, et

- mida suurem on valimite keskmiste vahe, seda väiksem on p;

- mida suurem on juhuhajuvus, seda suurem on p;

- mida suurem on valim, seda väiksem on p; seda muidugi juhul kui seos erinevus ikka on olemas

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* jutu lõpp \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*