Statistiliste hüpoteeside kontrollimine

Statistiliseks hüpoteesiks nimetatakse teatud teineteist välistavate väidete paari üldkogumi või tema parameetrite kohta.

H_0 – nullhüpotees, mis tavaliselt väljendab uurijat mittehuvitavat juhtu (üldkogumi vastamine teatud standardile).
H_1 – sisukas e. **alternatiivne** e. **konkureeriv hüpotees**, mida uurija soovib tõestada (tavaliselt mingi erinevuse, mõju või seose olemasolu).
Statistilisteks hüpoteesideks võivad olla näiteks oletused:

- jaotusseaduse tüübi kohta;
- kahe jaotuse parameetrite võrdsusest või olulisest erinevusest;
- juhuslike suuruste vahelise seose olemasolust või puudumisest.

Näide 1. H_0: ei ole sõltuvust palgataseme ja elukoha vahel
H_1: eksisteerib sõltuvus palgataseme ja elukoha vahel
Näide 2. H_0: kahes sama profiiliga firmas töötajate keskmised palgad ei erine (on võrdsed)
H_1: kahes sama profiiliga firmas töötajate keskmised palgad on erinevad
Otsus nende hüpoteeside kohta langetatakse valimi põhjal. Otsuse langetamiseks valitakse teatav statistik, mille tõenäosusjaotuse saab nullhüpoteesi kehtivuse ja teatud eelduste täidetuse korral kindlaks määrata. Kui see statistik (juhuslik suurus, mille konkreetne väärtus arvutatakse kasutades valimi objektidel mõõdetud väärtusi) omandab väärtuse, mis nullhüpoteesi korral ei ole ootuspärane, võetakse vastu sisukas hüpotees, vastasel juhul jäädakse nullhüpoteesi juurde. **Valimi põhjal vastu võetud sisukat hüpoteesi loetakse tõestatuks. Valimi põhjal vastu võetud nullhüpoteesi ei loeta tõestatuks, vaid öeldakse, et meil ei õnnestunud nullhüpoteesi kummutada**. Niisugune sõnastus jätab avatuks võimaluse, et tehes rohkem mõõtmisi, omades rohkem informatsiooni üldkogumi kohta, võib nullhüpoteesi kummutamine osutuda võimalikuks.
Kuna statistiliste hüpoteeside kontrollimisel tehakse valimi põhjal järeldusi üldkogumi kohta, on võimatu vältida vigu. Vigu saab olla kaht liiki.

- **Esimest liiki viga** tekib siis, kui võetakse vastu sisukas hüpotees, aga tegelikult on õige nullhüpotees.

See on raske viga, mis tähendab, et uurija "tõestas" erinevuse, mõju või seose mida tegelikult ei ole, vaid mis juhuslikult ilmnes mõõdetud valimis.

- **Teist liiki viga** tekib siis, kui jäädakse nullhüpoteesi juurde, ehkki tegelikult on õige sisukas hüpotees.

See on kergem viga, mis enamasti tähendab, et soovitu tõestamiseks tuleb mõõtmisandmeid juurde koguda.
Vea tekkimise suurust mõõdetakse tõenäosusega. Esimest liiki vea tegemise suurimat lubatavat tõenäosust nimetatakse **olulisuse nivooks e. riskiprotsendiks**. Olulisuse nivood tähistatakse tähega α. Olulisuse nivooks valitakse mingi väike arv, sageli 0,1; 0,05; 0,01 (sõltuvalt selles, kui rasketele tagajärgedele võib 1. liiki vea tegemine viia).
Otsuse langetamise reegli moodustamiseks määratakse esmalt kindlaks olulisuse nivoo α. Seejärel leitakse kriitiline piirkond K_α, kuhu statistiku Z väärtus H_0 kehtides satub väiksema tõenäosusega kui α.

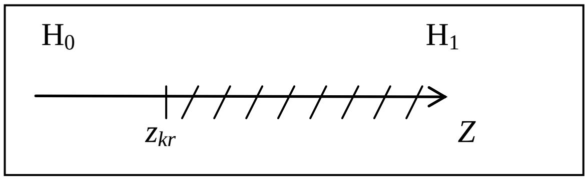
Otsuse langetamise reegel on alati ühesugune

- kui $Z \in K_\alpha$, siis võtta vastu sisukas hüpotees H_1;
- kui $Z \notin K_\alpha$, siis jääda nullhüpoteesi H_0 juurde.

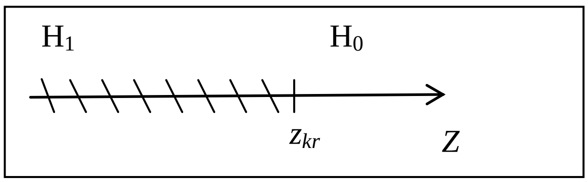
Kriitilised piirkonnad

Eristatakse järgnevaid kriitilisi piirkondi.

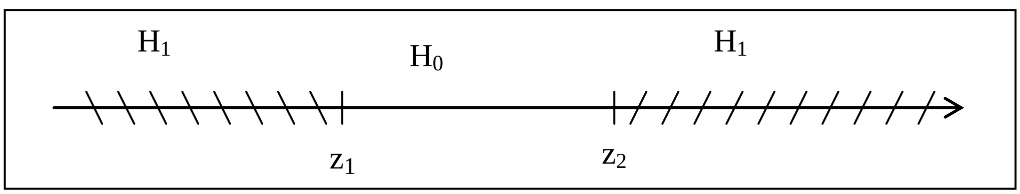
1. **parempoolne kriitiline piirkond**, mis jääb punktist z_{kr} paremale. Kriitiline punkt z_{kr} leitakse nii, et $P(Z > z_{kr}) = \alpha$



2. **vasakpoolne kriitiline piirkond**, mis jääb punktist z_{kr} vasakule. Kriitiline punkt z_{kr} leitakse nii, et $P(Z < z_{kr}) = \alpha$.



3. **kahepoolne kriitiline piirkond**, mis jääb ühest kriitilisest punktist z_1 vasakule ja teisest kriitilisest punktist z_2 paremale. Kriitilised punktid leitakse enamasti tingimustest $P(Z < z_1) = \alpha/2, \quad P(Z > z_2) = \alpha/2,$



Kahe normaaljaotuse keskväärtuse võrdlemine

Olgu juhuslikud suurused X ja Y normaaljaotusega ning sõltumatud. Olgu antud kaks valimit nende juhuslike suuruste väärtustest $x_1, x_2, \ldots x_n$ ja $y_1, y_2, \ldots y_m$. Vaatleme juhtu, kus nende valimite mahud on väikesed: $n < 30$ ja $m < 30$.

Leiame valimite aritmeetilised keskmised $\bar{x}$ ja $\bar{y}$. Üldiselt keskmised $\bar{x}$ ja $\bar{y}$ erinevad mõnevõrra teineteisest. Tekib küsimus, kas need keskmised erinevad sisuliselt või on see erinevus taandatav katsevigadele. Viimasel juhul peaks suuruste X ja Y keskväärtused EX ja EY olema võrdsed. Selle väite võtamegi nullhüpoteesiks, s.t. $H_0 : EX = EY$. Alternatiivse hüpoteesi formuleerime kujul $H_1 : EX \neq EY$.

Teststatistikuna kasutatakse sellisel juhul statistikut

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} . \qquad (1)$$

Kui valimite mahud on väikesed siis on see statistik ligikaudselt **Studenti jaotusega** vabadusastmete arvuga

$$k = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{\left(\frac{s_x^2}{n}\right)^2}{n-1} + \frac{\left(\frac{s_y^2}{m}\right)^2}{m-1}}.$$

Selle valemi abil saadud **tulemus ümardatakse vähima lähima täisarvuni**.
Olgu etteantud olulisuse nivoo α. Studenti jaotuse kvantiilide tabelist leiame kriitilise punkti t_{kr} kahepoolsele sümmeetrilisele kriitilisele piirkonnale (silmas pidades, et Studenti jaotusega juhusliku suuruse jaotustiheduse graafik on sümmeetriline vertikaaltelje suhtes)

$$t_{kr} = t(k;1-\frac{\alpha}{2})$$

Arvutame valemi (1) abil statistiku empiirilise väärtuse ja võrdleme seda tabelist leitud kriitilise väärtusega. Juhul kui $|t_{emp}| > t_{kr}$, siis loetakse tõestatuks alternatiivne hüpotees. Vastupidisel juhul jäädakse nullhüpoteesi juurde (öeldakse et meil ei õnnestunud nullhüpoteesi kummutada).

Näide. 1893. a. otsustas inglise füüsik J.W. Rayleigh mõõta võimalikult täpselt lämmastiku erikaalu. Ta teostas kaks seeriat katseid, kusjuures esimese seeria korral kasutati keemilistes reaktsioonides vabanenud lämmastikku, teise seeria korral aga õhulämmastikku. Esimene katseseeria koosnes kaheksat katsest, lämmastiku keskmine kaal teatud ruumalas oli $\bar{x} = 2{,}29947$ g, dispersioon oli $s^2 = 0{,}0000133152$. Teine katseseeria koosnes kümnest katsest ning andis keskmiseks kaaluks $\bar{y} = 2{,}31016$ g, dispersioon oli $s^2 = 0{,}0000001931$. Kas võib erinevust antud katseandmete vahel lugeda juhuslikuks?
Lahendus.
Antud on

$n = 8$	$\bar{x} = 2{,}29947$	$s_x^2 = 0{,}0000133152$
$m = 10$	$\bar{y} = 2{,}31016$	$s_y^2 = 0{,}0000001931$

Püstitame hüpoteesid
$H_0 : EX = EY$
$H_1 : EX \neq EY$
Kuna $n < 30$ ja $m < 30$, siis lähtume Studenti jaotusest ning arvutame vabadusastmete arvu

$$k = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{\left(\frac{s_x^2}{n}\right)^2}{n-1} + \frac{\left(\frac{s_y^2}{m}\right)^2}{m-1}} \approx 7 \text{ (ümardame lähima vähima täisarvuni).}$$

Olgu olulisuse nivoo $\alpha = 0{,}05$. Leiame kriitilise punkti

$$t_{kr} = t(k;1-\frac{\alpha}{2}) = t(7;1-\frac{0{,}05}{2}) = t(7;0{,}975) = 2{,}365.$$

Valimi andmetel arvutame teststatistiku empiirilise väärtuse

$$t_{emp} = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}} \approx -8{,}238$$

Kuna $|t_{emp}| > t_{kr}$, siis lükkame nullhüpoteesi tagasi ja loeme tõestatuks alternatiivse hüpoteesi. Järelikult lämmastiku erikaalud kahes katsereas on oluliselt erinevad. (Hiljem selgus, et erinevuse põhjustas õhulämmastikus esinev gaas argoon, mis on lämmastikust raskem. Need Rayleigh'i poolt teostatud katsed viisidki argooni avastamisele.)

Kahe normaaljaotuse dispersiooni võrdlemine. Fisheri jaotus

Olgu meil kaks katseseeriat valimi dispersioonidega s_1^2 ja s_2^2 ning vabadusastmete arvuga $k_1 = n-1$ ja $k_2 = m-1$. On võimalik näidata, et kui $DX = DY$ (vastavate üldkogumite dispersioonid on võrdsed), siis juhuslik suurus

$$F(k_1, k_2) = \frac{s_1^2}{s_2^2}$$

on **Fisheri jaotusega (*F*-jaotusega)**. Jaotuse nimetus on tulnud ameerika statistiku R. A. Fischeri (1890 – 1962) nimest. *F*-jaotus on ebasümmeetriline, selle kuju sõltub oluliselt vabadusastmete arvudest $k_1 = n-1$ ja $k_2 = m-1$.

Vaatleme järgmist ülesannet. Olgu antud kahe katseseeria tulemused. Esimene seeria koosnegu *n* katsest, selle valimi dispersioon olgu s_1^2, teine katseseeria koosnegu *m* katsest ja selle valimi dispersioon olgu s_2^2. Meid huvitab kas mõlemad katseseeriad on tehtud võrdse täpsusega või mitte. Et täpsust saab hinnata ülddispersiooniga, siis ühesuguse täpsuse korral peaksid mõlema katseseeria ülddispersioonid *DX* ja *DY* olema võrdsed.

Võtame nullhüpoteesiks väite, et mõlemad katseseeriad on läbi viidud sama täpsusega $H_0 : DX = DY$. Alternatiivseks hüpoteesiks olgu $H_1 : DX \neq DY$. Teststatistikuna kasutatakse sellisel juhul statistikut

$$F = \frac{s_1^2}{s_2^2}, \quad kus \quad s_1^2 > s_2^2. \tag{2}$$

Olgu etteantud olulisuse nivoo α. Fisheri jaotuse täiendkvantiilide tabelist leiame kriitilised punktid kahepoolsele kriitilisele piirkonnale

$$F_p = F(k_1, k_2, \alpha/2) \quad ja \quad F_v = \frac{1}{F(k_2, k_1, \alpha/2)},$$

seejuures k_1 on suurema dispersioonihinnanguga valimi vabadusastmete arv.

Arvutame valemi (2) abil teststatistiku empiirilise väärtuse ja võrdleme seda tabelist leitud kriitiliste väärtustega. Juhul kui $F_v < F_{emp} < F_p$, siis jäädakse nullhüpoteesi juurde. Kui üks võrratustest on rikutud, võetakse vastu alternatiivne hüpotees.

Näide. Kaks töötajat kaalusid mingit toiduainet ühesuguse kaaluga portsjoniteks. Esimene kaalus 13 portsjonit, nende põhjal leitud valimi dispersiooniks saadi $0{,}0139$. Teine kaalus 21 portsjonit dispersiooniga $0{,}0295$. Kas võime väita, et töötajad töötasid erineva täpsusega?

Lahendus.

Antud on

$n=13$	$s^2=0{,}0139$
$m=21$	$s^2=0{,}0295$

Püstitame hüpoteesid

$H_0: DX = DY$

$H_1: DX \neq DY$

Olgu $\alpha = 0{,}10$

Kuna $0{,}0295 > 0{,}0139$, siis $k_1 = 21-1 = 20$ (k_1 on suurema dispersioonihinnanguga valimi vabadusastmete arv) ja $k_2 = 13-1 = 12$.

F-jaotuse täiendkvantiilide tabelist leiame kaks kriitilist punkti

$$F_p = F(k_1, k_2, \alpha/2) = F(20;12;0{,}05) = 2{,}54$$

$$F_v = \frac{1}{F(k_2, k_1, \alpha/2)} = \frac{1}{F(12;\,20;\,0{,}05)} = \frac{1}{2{,}28} = 0{,}44.$$

Arvutame teststatistiku empiirilise väärtuse

$$F_{emp} = \frac{s_1^2}{s_2^2} = \frac{0{,}0295}{0{,}0139} \approx 2{,}12$$

Kuna $F_v < F_{emp} < F_p$, (0,44 < 2,12 < 2,54), siis jääme nullhüpoteesi juurde. Pole alust väita, et dispersioonid on erinevad: pole alust väita, et töötajad töötavad erineva täpsusega.