Moninkertaisen testaamisen ongelma

Oula Regina

LuK-tutkielma
Syyskuu 2025

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Oula Regina: Moninkertaisen testaamisen ongelma
LuK-tutkielma, 17 s.
Tilastotiede
Syyskuu 2025

Tutkielmassa käydään aluksi läpi tilastollisessa testaamisessa käytettäviä käsitteitä kuten p-arvo, merkitsevyystaso, nolla- ja vastahypoteesi, hylkäys- ja hyväksymisvirheet, eli tyypin I ja II virheet, FWER (*family-wise error rate*) eli perhekohtainen virhetodennäköisyys ja FDR (*false discovery rate*) eli väärien löydösten osuus. Esitellään lyhyesti Bonferronin korjaus, Holmin ja Hochbergin menetelmät perhekohtaisen virhetodennäköisyyden kontrolloimiseen sekä Benjaminin ja Hochbergin menetelmä ja Benjaminin ja Yekutielin menetelmä väärien löydösten osuuden kontrolloimiseen. Lopussa verrataan näitä menetelmiä toisiinsa.

Sisällys

1 Johdanto

Tilastotieteessä useimmiten käsitellään erilaisia aineistoja ja pyritään selvittämään, onko jollakin asialla vaikutusta johonkin toiseen asiaan. Aineistona voisi olla esimerkiksi muutaman tuhannen ihmisen kattava terveyshistoria ja mielenkiinnon kohteena tupakoinnin vaikutus keuhkosyövän riskiin. Erilaisilla testeillä voidaan todeta tupakoinnin positiivinen tai negatiivinen vaikutus ja tärkeimpänä asiana, onko vaikutus tilastollisesti merkitsevä. Tilastollista merkitsevyyttä havainnollistetaan p-arvolla.

P-arvo on helpointa selittää esimerkin avulla. Oletaan, että tutkimuksessa saatu tulos viittaa siihen, että tupakointi nostaisi keuhkosyövän riskiä ja p-arvoksi on saatu $p = 0.01$. Tämä tarkoittaa sitä, että jos alkuperäinen oletus on ollut, että tupakointi ei nosta keuhkosyövän riskiä, saatu tulos (tai huonompi) tapahtuisi vain 1 %:n todennäköisyydellä. Näin ollen voidaan olla melko varmoja siitä, että tupakointi oikeasti nostaa keuhkosyövän riskiä, eikä tulos ole ollut vain sattumaa.

Entä jos samalla aineistolla tehtäisiin useampia testejä, vaikka satoja tai tuhansia erilaisia testejä, eikö silloin p-arvoja ole niin paljon, että joukossa on sattuman kautta mukana tuloksia, jotka ovat tilastollisesti merkitseviä, mutta johtavat hylkäysvirheisiin eli totta olevien nollahypoteesien hylkäämiseen? Tässä tutkielmassa keskitytään juuri tähän ongelmaan ja käydään läpi millaisia ratkaisuja ongelmaan on kehitetty.

Tutkielmassa käydään aluksi läpi hieman tarkemmin, miten tilastollinen testaaminen käytännössä tapahtuu. Tämän jälkeen tarkastellaan, miten p-arvot ovat jakautuneet monen samanaikaisen testin tilanteessa ja millaisia päätelmiä voimme tehdä histogrammien avulla. Seuraavaksi esitellään perhekohtainen virhetodennäköisyys ja kolme erilaista menetelmää sen kontrolloimiseen. Sen jälkeen esitellään väärien löydösten osuus ja kaksi erilaista menetelmää sen kontrolloimiseen. Lopuksi verrataan kaikkia perhekohtaisen virhetodennäköisyyden ja väärien löydösten osuuden menetelmiä toisiinsa.

Lähteinä tutkielmassa käytettiin enimmäkseen Bradley Efronin ja Trevor Hastien kirjaa *Computer age statistical inference* (s. 271-297) [1] ja Wikipedian artikkeleita: perhekohtainen virhetodennäköisyys [9], Bonferronin korjaus [10], Holmin menetelmä [11] ja väärien löydösten osuus [12].

2 Tilastollinen testaaminen

Palautetaan mieleen, miten tilastollinen testaaminen käytännössä tapahtuu. Ennen testien tekemistä lähes aina tehdään nollahypoteesi H_0 ja vastahypoteesi H_1,

$$\begin{cases} H_0 : \text{säännöllinen liikunta ei vaikuta elinajanodotteeseen.} \\ H_1 : \text{säännöllinen liikunta nostaa elinajanodotetta.} \end{cases}$$

sekä valitaan testin merkitsevyystaso α. Merkitsevyystasolla tarkoitetaan sitä, että jos $p \leq \alpha$, niin hylätään H_0 ja hyväksytään H_1. Jos testejä tehdään enemmän kuin yksi, merkitään nollahypoteesejä H_{0i}, vastahypoteesejä H_{1i} ja niiden p-arvoja p_i ($i = 1, 2, 3, \ldots, N$). Yleisimpiä merkitsevyystasoja ovat $\alpha = 0.01$, $\alpha = 0.05$ ja $\alpha = 0.10$. P-arvo saadaan erilaisten testien tuloksena. Tässä (täysin keksityssä) esimerkissä käytetään yhden otoksen t-testiä. Entuudestaan on tiedossa, että ei-säännöllisesti

liikuntaa harrastavan ihmisen keskimääräinen elinajanodote on $\mu_0 = 78$ vuotta. Tutkimme tuhannen säännöllisesti liikuntaa harrastavan ihmisen terveyshistoriaa ja laskemme, että heidän keskimääräinen elinajanodote on $\bar{y} = 79$ vuotta. Otoksen otosvarianssi lasketaan seuraavasti:

$$s^2 = \frac{1}{1000 - 1} \sum_{i=1}^{1000} (y_i - \bar{y})^2$$

Kaavassa y_i tarkoittaa yksittäisen liikuntaa säännöllisesti harrastavan ihmisen elinajanodotetta ja $\bar{y}$ tarkoittaa kaikkien tuhannen liikuntaa säännöllisesti harrastavan ihmisen elinajanodotteen keskiarvoa. Oletetaan, että $s^2 = \hat{\sigma}^2 = 49$, jolloin voimme laskea t-testisuureen:

$$t = \frac{\bar{y} - \mu_0}{\hat{\sigma}/\sqrt{1000}} = \frac{79 - 78}{7/\sqrt{1000}} = 4.518$$

P-arvo saadaan tutkimalla t_{1000-1} jakauman oikean puolen häntätodennäköisyyttä pisteessä $t = 4.518$. P-arvo on suoraan jakauman t_{1000-1} pisteen t oikean puolen hännän pinta-ala. Tässä esimerkissä $p = 0.0000035$ ja voimme hylätä hypoteesin H_0 kaikilla yleisillä merkitsevyystasoilla α. Toinen tapa on valita merkitsevyystaso esim. $\alpha = 0.05$ ja laskea jakauman t_{1000-1} kriittinen piste $t_{1-\alpha,1000-1} = 1.962$. Jos $t \geq t_{1-\alpha,1000-1}$, niin hylätään H_0.

On kuitenkin mahdollista, että teemme hylkäysvirheen eli hylkäämme hypoteesin H_0, vaikka se todellisuudessa onkin totta. Vastaavasti hyväksymisvirhe tapahtuu silloin kun $p > \alpha$ ja hyväksymme hypoteesin H_0, vaikka oikeasti se ei pidä paikkaansa. Näitä nimitetään myös tyypin I ja II virheiksi.

3 Monta samanaikaista testiä

Palataan nyt siihen kysymykseen, mihin johdanto jäi, eli mitä jos teemme vaikkapa $N = 1000$ erilaista testiä jollakin merkitsevyystasolla α, minkälaisia tuloksia voimme odottaa? Erittäin suurella todennäköisyydellä osa p-arvoista on alle valitsemamme merkitsevyystason, mutta kuinka monta? Jos oletamme, että kaikki nollahypoteesit H_{0i} ovat oikeasti totta, jo pelkästään sattuman kautta odotusarvoisesti noin $\alpha \cdot N$ p-arvoa on alle merkitsevyystason. Tämä johtuu siitä, että jos oletamme, että kaikki nollahypoteesit H_{0i} ovat oikeasti totta, niin p-arvot ovat tasaisesti jakautuneet välille $(0, 1)$ [5]. Toisin sanoen jokaisella p-arvolla on yhtä suuri todennäköisyys osua mihin tahansa kohtaan väliä $(0, 1)$.

Kuvassa 1 on otos tasajakaumasta ($N = 1000$), joka kuvaa realistista tilannetta p-arvojen histogrammista, kun kaikki nollahypoteesit ovat oikeasti totta ja kuvassa 3 on p-arvojen odotusarvoinen histogrammi, kun kaikki nollahypoteesit ovat oikeasti totta. Kuvassa 2 on yhdistetty kaksi otosta, ensimmäinen otos ($N_0 = 900$) on otettu tasajakaumasta ja toinen otos ($N_1 = 100$) on otettu jakaumasta $Beta(0.1, 1)$, millä yritetään kuvata tilannetta, jossa nollahypoteeseistä 900 on totta ja 100 ei ole totta. Kuvassa 4 on kuvan 2 otoksen odotusarvoinen histogrammi.

Kun nollahypoteesit eivät ole totta, niiden p-arvot eivät ole yleensä jakautuneet jakauman $Beta(0.1, 1)$ mukaisesti, eikä niille luultavasti ole edes olemassa mitään

tiettyä jakaumaa, jakauma $Beta(0.1, 1)$ vain kuvastaa esimerkinomaisesti sitä, millä tavalla p-arvot suurinpiirtein käyttäytyvät, kun nollahypoteesit eivät ole totta.

Kuvien 3 ja 4 molemmissa histogrammeissa on yhteensä 1000 p-arvoa ja 20 pylvästä. Kuvassa 3 jokaisessa pylväässä on 50 p-arvoa, koska kaikki nollahypoteesit ovat totta ja kaikki p-arvot ovat tasaisesti jakautuneet. Kuvassa 3 punaisen viivan ($y = 50$) alapuolelle jäävät p-arvot kuvastavat niitä p-arvoja, jotka ovat tasaisesti jakautuneet eli tässä tapauksessa kaikki p-arvot. Kuvassa 4 100 nollahypoteesia ei ole oikeasti totta ja 900 nollahypoteesia on totta. Tällöin 900 p-arvoa on edelleen tasaisesti jakautuneet ja tätä kuvaa punainen viiva ($y = 45$). Punaisen viivan yläpuolella olevat 100 p-arvoa eivät ole tasaisesti jakautuneet.

Jos katsomme kuvia 3 ja 4, mitkä p-arvot jäävät merkitsevyystason alapuolelle? Jos valitsemme merkitsevyystason $\alpha = 0.05$, tällöin molempien histogrammien ensimmäinen pylväs kuvastaa niitä p-arvoja, jotka ovat alle α, eli löydöksiä. Kuvasta 3 nähdään, että 50 p-arvoa jää alle merkitsevyystason, punaisen viivan ($y = 50$) alapuolelle jäävät p-arvot histogrammin ensimmäisessä pylväässä kuvastavat niitä p-arvoja, jotka johtavat hylkäysvirheeseen, eli tässä tapauksessa kaikki 50 p-arvoa. Kuvan 4 perusteella noin 120 p-arvoa jää merkitsevyystason alapuolelle ja punaisen viivan ($y = 45$) alapuolelle jää 45 p-arvoa, jotka johtavat hylkäysvirheeseen. Huomataan, että hylkäysvirheitä tapahtuu vähemmän, kun totta olevien nollahypoteesien määrä vähenee olettaen, että testejä tehdään kuitenkin sama määrä.

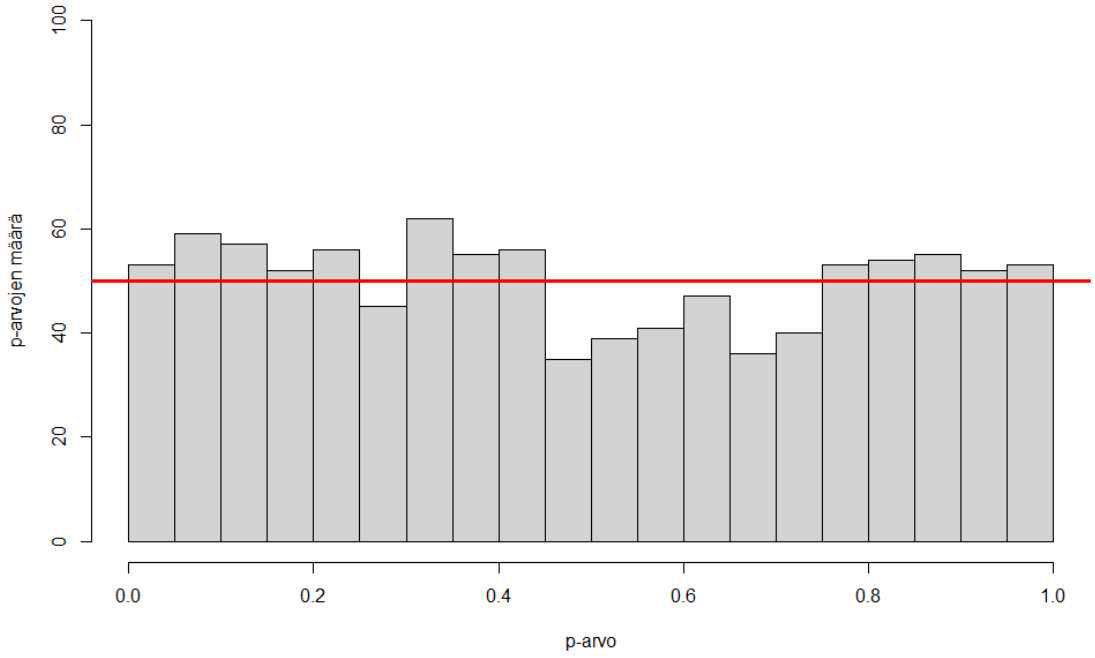
Entä jos emme tiedä, mitkä p-arvot ovat tasaisesti jakautuneet eli mitkä nollahypoteesit ovat oikeasti totta? Tämä on ehkä hieman tyhmä kysymys, koska jos tietäisimme mitkä nollahypoteesit ovat oikeasti totta niin miksi sitten edes testaisimme yhtään mitään. Eli lähes aina on niin, että emme tiedä mitkä nollahypoteesit ovat oikeasti totta. Näin ollen täytyy lähes aina olettaa pahinta, eli että kaikki p-arvot ovat tasaisesti jakautuneet. Tällöin meillä on odotusarvoinen yläraja $\alpha \cdot N$ hylkäysvirheiden määrälle jos testaisimme jokaista nollahypoteesiä tasolla α. Mainitaan vielä, että jos tehdään N määrä testejä ja tiedetään varmasti, että vähintään esimerkiksi 5 % nollahypoteeseistä ei pidä paikkaans, niin odotusarvoinen yläraja hylkäysvirheille on $\alpha \cdot N \cdot (1 - 0.05)$ jos testataan kaikkia nollahypoteesejä tasolla α.

3.1 Perhekohtainen virhetodennäköisyys (FWER)

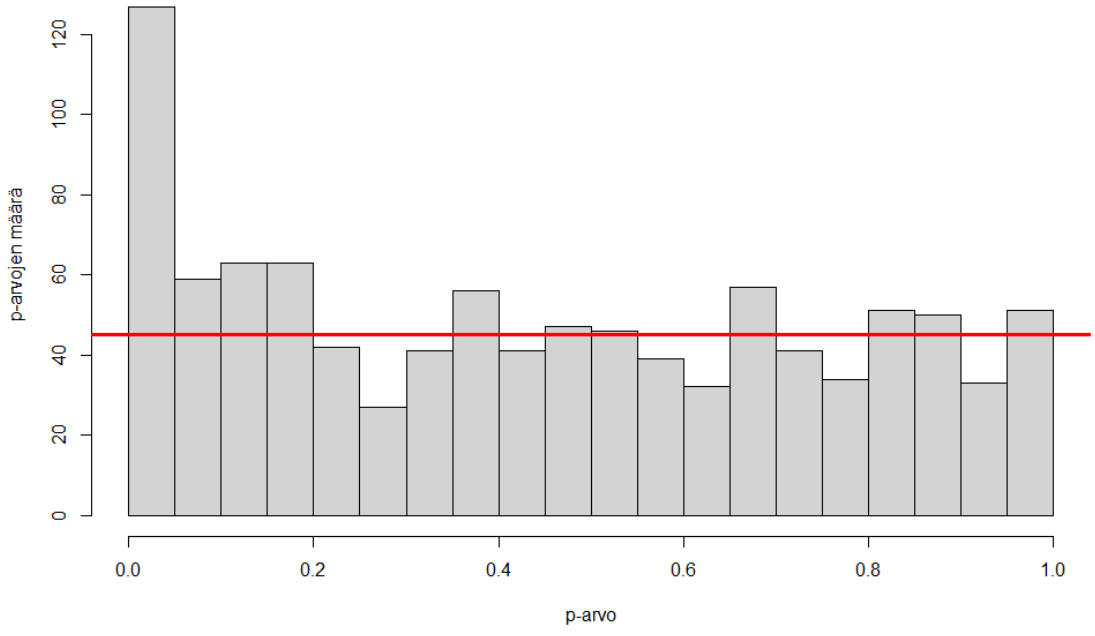
Perhekohtainen virhetodennäköisyys (eng. *family-wise error rate*) [9], eli yhden tai useamman hylkäysvirheen tekemisen todennäköisyys tulee ajankohtaiseksi silloin, kun teemme enemmän kuin yhden testin. Tämä määritellään kaavalla

$$FWER = P(V \geq 1), \text{ jossa } V \text{ on hylkäysvirheiden lkm.}$$

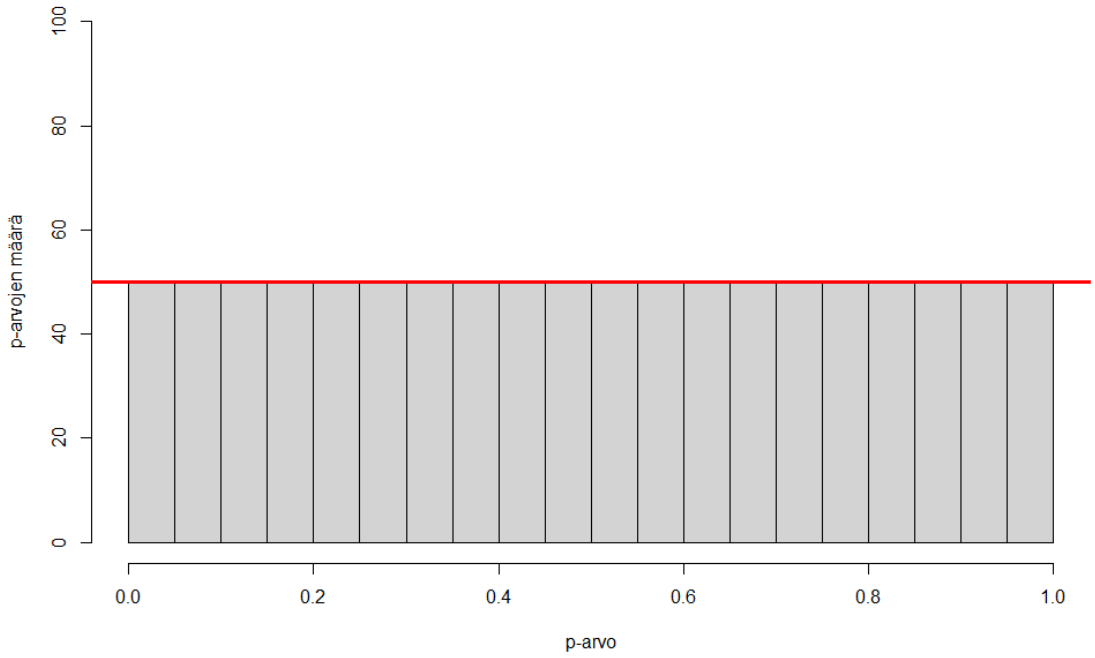
Jos teemme esimerkiksi 10 erilaista testiä tasolla $\alpha = 0.05$, todennäköisyys sille, että ainakin yksi testi johtaa hylkäämisvirheeseen on $1 - (1 - \alpha)^{10} = 0.401$, eli $FWER = 0.401$ edellyttäen, että testit ovat riippumattomia ja tässä tapauksessa myös, että kaikki nollahypoteesit ovat totta. Jos emme tiedä ovatko kaikki nollahypoteesit totta (mikä on lähtökohtainen tilanne, miksi muuten edes testaisimme mitään?), tällöin $FWER \leq 1 - (1 - \alpha)^{10} = 0.401$, koska on mahdollista, että jokin tai jotkin nollahypoteesit eivät ole totta, jolloin hylkäämisvirheiden tekemisen todennäköisyys laskee. Jos taas teemme sata erilaista testiä todennäköisyys yhdelle tai



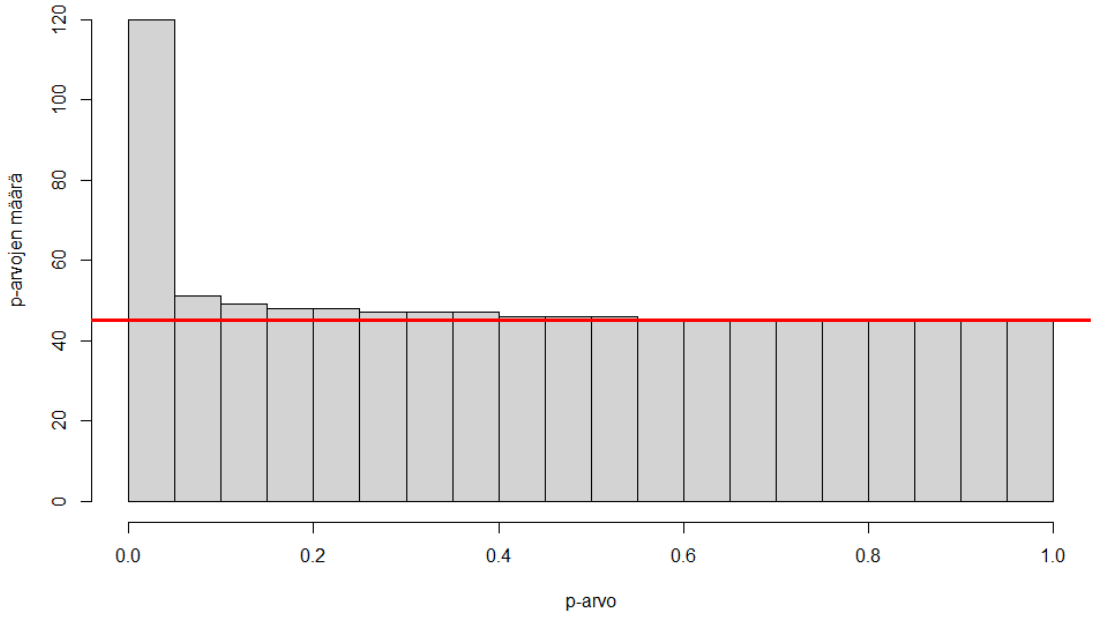
Kuva 1: Otos tasajakaumasta ($N = 1000$).



Kuva 2: Yhdistetty otos tasajakaumasta ($N_0 = 900$) ja jakaumasta $Beta(0.1, 1)$ ($N_1 = 100$).



Kuva 3: Tasajakaumasta otetun otoksen ($N = 1000$) odotusarvoinen histogrammi.



Kuva 4: Kuvan 2 otoksen odotusarvoinen histogrammi.

useammalle hylkäysvirheelle on jo 99.4 % tai vähemmän, toisin sanottuna $FWER \leq 1 - (1 - \alpha)^{100} = 0.994$. Jos haluamme, että perhekohtainen virhetodennäköisyys on alle α, yksittäisen testin merkitsevyystasoa täytyy muuttaa ja nimenomaan pienentää. Jos testejä tehdään esimerkiksi 10, niin voidaan intuitiivisesti kokeilla laskea yksittäisen testin merkitsevyystasoa 10 kertaa pienemmäksi, eli uusi merkitsevyystaso yksittäiselle testille olisi $\alpha/10$. Tällöin $FWER \leq 1 - (1 - \frac{\alpha}{10})^{10} = 0.0489$, joka on alle $\alpha = 0.05$, eli onnistuimme pitämään perhekohtaisen virhetodennäköisyyden alle α. Jos testejä tehdään 100 ja yksittäisen testin merkitsevyystasoksi valitaan $\alpha/100$, niin tällöin $FWER \leq 1 - (1 - \frac{\alpha}{100})^{100} = 0.0488 \leq \alpha = 0.05$, eli tämä menetelmä näyttäisi toimivan. Menetelmän nimi on itseasiassa Bonferronin korjaus (eng. *Bonferroni correction*) ja se on varmasti tunnetuin moninkertaisen testaamisen ongelmaan kehitetty ratkaisu ja se toimii ilman minkäänlaisia oletuksia testien riippumattomuudesta.

3.2 Bonferronin korjaus

Bonferronin korjauksessa [10] nollahypoteesi H_{0i} hylätään jos ja vain jos

$$p_i \leq \frac{\alpha}{N} \tag{1}$$

tai toisin ilmaistuna H_{0i} hylätään jos ja vain jos

$$p_i \cdot N \leq \alpha$$

Ehto (1) varmistaa sen, että perhekohtainen virhetodennäköisyys, eli todennäköisyys yhdenkin hylkäysvirheen tekemiselle on alle α eli $FWER \leq \alpha$. Todistetaan tämä väite:

Todistus. On tehty N testiä. Merkitään I:llä sitä indeksien i joukkoa, jota vastaavat nollahypoteesit H_{0i} ovat totta. Joukon I koko on $|I| = N_0 \leq N$, eli totta olevien nollahypoteesien H_{0i} lukumäärä. Olkoon V_i se tapahtuma, että H_{0i} tulee hylätyksi testin puitteissa. Tällöin

$$FWER = P\left(\bigcup_{i \in I} V_i\right) \leq \sum_{i \in I} P(V_i) \leq \sum_{i \in I} \frac{\alpha}{N} = N_0 \frac{\alpha}{N} \leq \alpha$$

□

Bonferronin korjaus kontrolloi perhekohtaista virhetodennäköisyyttä vahvassa mielessä, tarkoittaen sitä, että epäyhtälö $FWER \leq \alpha$ toteutuu kaikissa tapauksissa, huolimatta siitä kuinka moni nollahypoteeseistä on tai ei ole oikeasti totta. Vastaavasti ne menetelmät, jotka kontrolloivat perhekohtaista virhetodennäköisyyttä heikossa mielessä, epäyhtälö $FWER \leq \alpha$ toteutuu varmasti ainoastaan silloin, kun kaikki nollahypoteesit ovat oikeasti totta.

Käytännön esimerkki Bonferronin korjauksesta: oletetaan, että on tehty viisi erilaista testiä ($N = 5$) merkitsevyystasolla $\alpha = 0.05$. Testien nollahypoteesit ovat H_{01}, H_{02}, H_{03}, H_{04} ja H_{05}, ja testien p-arvot ovat $p_1 = 0.03$, $p_2 = 0.001$, $p_3 = 0.008$, $p_4 = 0.01$ ja $p_5 = 0.02$. Koska testejä on tehty viisi kappaletta, yksittäisen testin

merkitsevyystasoa on muutettava ehdon (1) mukaisesti. Testin hylkäysehto saa nyt muodon

$$p_i \leq \frac{\alpha}{N} = \frac{0.05}{5} = 0.01 \tag{2}$$

Epäyhtälö (2) totetuu ainoastaan p-arvoilla p_2, p_3 ja p_4, ja näin ollen hylkäämme vain nollahypoteesit H_{02}, H_{03} ja H_{04}.

Bonferronin korjaus on helppo menetelmä, mutta erittäin konservatiivinen tarkoittaen sitä, että hyväksymisvirheiden tekeminen kasvaa rajusti. Tämä on Bonferronin korjauksen iso haittapuoli ja tutkijan suuri harmi kun on syytä epäillä, että ei totta olevia nollahypoteesejä on luultavasti paljon enemmän.

3.3 Holmin menetelmä

Bonferronin korjauksesta on tehty hieman paranneltu versio nimeltään Holmin menetelmä (eng. *Holm's procedure*) [11]. Holmin menetelmässä p-arvot p_i järjestetään pienimmästä suurimpaan. Merkitään näitä p-arvoja $(p_{(1)}, p_{(2)}, ..., p_{(N)})$ ja järjestetään myös niiden nollahypoteesit samaan järjestykseen $(H_{0(1)}, H_{0(2)}, ..., H_{0(N)})$ merkinnällisistä syistä. Seuraavaksi käytetään algoritmia, jossa tarkastellaan p-arvoja pienimmästä lähtien

Jos $p_{(1)} \leq \alpha/N$, niin hylätään $H_{0(1)}$.
Jos $p_{(2)} \leq \alpha/(N-1)$, niin hylätään $H_{0(2)}$.
Jos $p_{(3)} \leq \alpha/(N-2)$, niin hylätään $H_{0(3)}$.
jne.

Ensimmäisen kerran kun epäyhtälö $p_{(i)} \leq \alpha/(N-i+1)$ ei pidä paikkaansa, algoritmi päättyy ja olemme hylänneet kaikki nollahypoteesit indeksiin $i-1$ asti.

Holmin menetelmän voi ilmaista myös hieman lyhyemmin. Etsitään pienin $p_{(i)}$ siten, että

$$p_{(i)} > \frac{\alpha}{N-i+1}$$

ja hylätään kaikki sitä edeltäneet $H_{0(1)}, ..., H_{0(i-1)}$ nollahypoteesit.

Holmin menetelmä kontrolloi perhekohtaista virhetodennäköisyyttä vahvassa mielessä samalla tasolla α kuin Bonferronin korjaus, eli Holminkin menetelmässä $FWER \leq \alpha$. Todistetaan tämä väite:

Todistus. On tehty N testiä. Merkitään I:llä sitä (alkuperäistä ja järjestämätöntä) joukkoa, joka sisältää kaikki oikeat nollahypoteesit ja $|I| = N_0 \leq N$. Järjestetään testien p-arvot pienimmästä suurimpaan $(p_{(1)}, p_{(2)}, ..., p_{(N)})$ ja niiden nollahypoteesit $(H_{0(1)}, H_{0(2)}, ..., H_{0(N)})$ samaan järjestykseen.

Väite: Jos teemme hylkäysvirheen jollekin nollahypoteesille, tällöin on olemassa nollahypoteesi $H_{0(l)}$, jonka p-arvo $p_{(l)}$ on korkeintaan $\frac{\alpha}{N_0}$.

Tässä tapauksessa $N_0 \geq 1$. Olkoon l sellainen, että $H_{0(l)}$ on Holmin algoritmin ensimmäinen totta oleva nollahypoteesi. Tällöin nollahypoteesit $H_{0(1)}, ..., H_{0(l-1)}$ on

hylätty ilman hylkäysvirheitä. Tästä seuraa, että $l - 1 \leq N - N_0$, mistä seuraa edelleen, että $N_0 \leq N - l + 1$, jolloin

$$\frac{\alpha}{N - l + 1} \leq \frac{\alpha}{N_0} \tag{3}$$

Koska $H_{0(l)}$ on hylätty, epäyhtälön $p_{(l)} \leq \frac{\alpha}{N-l+1}$ täytyy toteutua Holmin algoritmin mukaan. Epäyhtälön (3) avulla voimme todeta, että $p_{(l)} \leq \frac{\alpha}{N_0}$, kuten väite oli.

Määritellään satunnaistapahtuma $A = \bigcup_{i \in I} \left\{ p_i \leq \frac{\alpha}{N_0} \right\}$, jossa p_i on totta olevien nollahypoteesien H_{0i} p-arvo. Tällöin

$$P(A) \leq \sum_{i \in I} P\left(\left\{ p_i \leq \frac{\alpha}{N_0} \right\} \right) = \sum_{i \in I} \frac{\alpha}{N_0} = N_0 \frac{\alpha}{N_0} = \alpha$$

Tästä seuraa, että todennäköisyys yhden tai useamman hylkäysvirheen tekemiselle on korkeintaan α, eli

$$FWER \leq \alpha$$

□

Verrataan hieman Holmin menetelmää Bonferronin korjaukseen. Otetaan sama esimerkki, jota käytettiin Bonferronin korjauksessa, eli viisi testiä joiden p-arvot ovat $p_1 = 0.03$, $p_2 = 0.001$, $p_3 = 0.008$, $p_4 = 0.01$ ja $p_5 = 0.02$ samalla merkitsevyystasolla $\alpha = 0.05$. Järjestetään ensin p-arvot pienimmästä suurimpaan ($p_{(1)} = 0.001$, $p_{(2)} = 0.008$, $p_{(3)} = 0.01$, $p_{(4)} = 0.02$ ja $p_{(5)} = 0.03$). Käytetään nyt Holmin menetelmän algoritmia

$$\begin{gathered} p_{(1)} = 0.001 \leq \alpha/5 = 0.05/5 = 0.01 \\ p_{(2)} = 0.008 \leq \alpha/(5-1) = 0.05/4 = 0.0125 \\ p_{(3)} = 0.01 \leq \alpha/(5-2) = 0.05/3 = 0.0167 \\ p_{(4)} = 0.02 \leq \alpha/(5-3) = 0.05/2 = 0.025 \\ p_{(5)} = 0.03 \leq \alpha/(5-4) = 0.05/1 = 0.05 \end{gathered}$$

Jokainen epäyhtälö näyttäisi toteutuvan, joten tällä kertaa voimme hylätä kaikki viisi nollahypoteesiä.

Kuten esimerkistä ja yhtälöistä huomataan Holmin menetelmä ei ole aivan niin konservatiivinen kuin Bonferronin korjaus, eli Holmin menetelmällä voidaan tehdä enemmän löydöksiä (löydös tarkoittaa sitä, että p-arvo on alle merkitsevyystason, eli hylätään nollahypoteesi ja väärä löydös tarkoittaa, että tehdään hylkäysvirhe) kuin Bonferronin korjauksella ja tästä syystä se on kaikin puolin parempi menetelmä, jos Bonferronin helppoutta ei lasketa.

3.4 Hochbergin menetelmä

Mainitaan vielä Hochbergin menetelmä, jonka alkuperäinen lähde on [7]. Hochbergin menetelmässä käytetään muuten samaa algoritmia, jota käytettiin Holmin menetelmässä, mutta tällä kertaa algoritmi päättyy vasta, kun on löydetty suurin $p_{(i)}$, joka toteuttaa epäyhtälön

$$p_{(i)} \leq \frac{\alpha}{N - i + 1} \tag{4}$$

ja hylätään nollahypoteesit $H_{0(1)}, \ldots, H_{0(i)}$. Hochbergin menetelmä sallii siis tilanteen, jossa osa p-arvoista ei toteuta epäyhtälöä (4). Tämä edellyttää sitä, että algoritmin myöhemmissä askeleissa löydetään suurin p-arvo $p_{(i)}$, joka toteuttaa epäyhtälön $p_{(i)} \leq \alpha/(N-i+1)$ ja näin ollen kaikki järjestetyt nollahypoteesit $H_{0(1)}, ..., H_{0(i)}$ hylätään. Tämä käy selvemmäksi esimerkin kautta.

Oletetaan, että ollaan tehty viisi testiä ja testien p-arvot ovat $p_1 = 0.026$, $p_2 = 0.027$, $p_3 = 0.028$, $p_4 = 0.029$ ja $p_5 = 0.030$. Järjestetään p-arvot pienimmästä suurimpaan $p_{(1)} = 0.026$, $p_{(2)} = 0.027$, $p_{(3)} = 0.028$, $p_{(4)} = 0.029$ ja $p_{(5)} = 0.030$ Seuraavaksi etsitään suurin $p_{(i)}$, joka toteuttaa epäyhtälön (4). Suurin ja itseasiassa ainut p-arvo, joka toteuttaa epäyhtälön (4) on $p_{(5)}$ ja näin ollen voimme hylätä kaikki viisi nollahypoteesiä.

Jos näihin p-arvoihin olisi käytetty Holmin menetelmää, Holmin algoritmi olisi päättynyt jo ensimmäisellä askeleella ja emme olisi hylänneet yhtäkään nollahypoteesiä. Tästä syystä Hochbergin menetelmällä voidaan tehdä enemmän löydöksiä kuin Holmin menetelmällä. Hochbergin menetelmä ei kuitenkaan toimi joissakin tilanteissa, joissa p-arvot ovat negatiivisesti korreloituneita. Sen sijaan Holmin menetelmä toimii kaikissa tapauksissa.

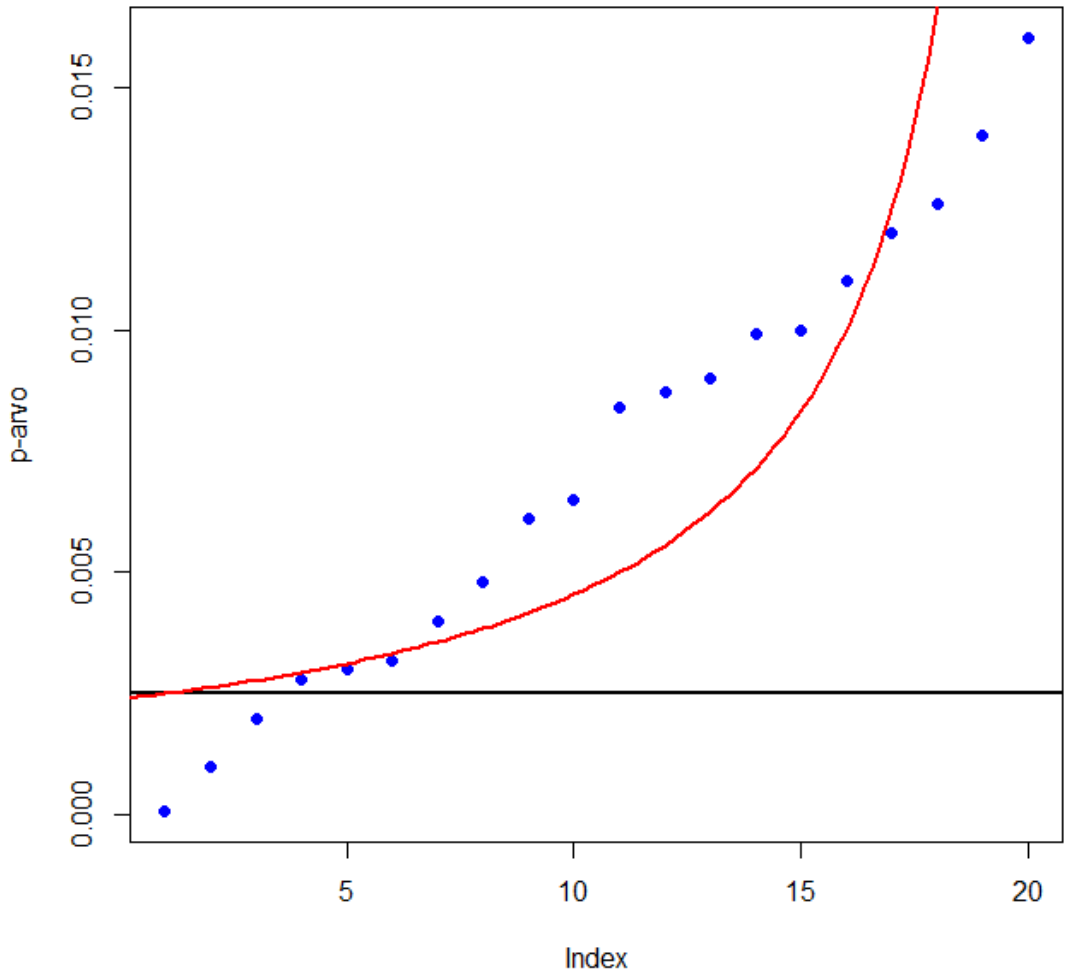
Hochbergin menetelmän algoritmin voisi rakentaa myös siten, että lähdetäänkin liikkeelle suurimmasta p-arvosta $p_{(N)}$,

$$\begin{gathered}\text{Jos } p_{(N)} \leq \alpha, \text{ niin hylätään } H_{0(1)}, ..., H_{0(N)}.\\ \text{Jos } p_{(N-1)} \leq \alpha/2, \text{ niin hylätään } H_{0(1)}, ..., H_{0(N-1)}.\\ \text{Jos } p_{(N-2)} \leq \alpha/3, \text{ niin hylätään } H_{0(1)}, ..., H_{0(N-2)}.\\ \text{jne.}\end{gathered}$$

Algoritmi päättyy heti, kun saavutetaan askel, jossa epäyhtälö $p_{(i)} \leq \alpha/(N-i+1)$ toteutuu. Jos siis algoritmin ensimmäisellä askeleella epäyhtälö $p_{(N)} \leq \alpha$ toteutuu, hylätään kaikki nollahypoteesit. Jos samoihin p-arvoihin käytettäisiinkin Holmin menetelmää, on mahdollista, että yksikään nollahypoteesi ei tule hylätyksi. Esimerkkinä tilanne jossa kaikki p-arvot $p_{(1)} = \cdots = p_{(N)} = \alpha$. Tämä kuulostaa oudolta, mutta myöhemmin nähdään, että Holmin ja Hochbergin menetelmät tekevät käytännössä yhtä paljon löydöksiä.

Havainnollistetaan vielä hieman Bonferronin korjauksen sekä Holmin ja Hochbergin menetelmien eroja. Oletetaan, että ollaan tehty 20 testiä merkitsevyystasolla $\alpha = 0.05$. Kuvassa 5 nämä 20 p-arvoa on järjestetty pienimmästä suurimpaan (siniset pisteet). Kuvaan 5 on myös piirretty mustalla funktion $p = \alpha/20$ kuvaaja, joka kuvaa Bonferronin korjausta, sekä punaisella funktion $p = \alpha/(20-i+1)$ kuvaaja, joka kuvaa Holmin ja Hochbergin menetelmiä. Kuvasta 5 nähdään, että Bonferronin korjaus tekee 3 löydöstä, Holmin menetelmä tekee 6 löydöstä ja Hochbergin menetelmä tekee 20 löydöstä.

Bonferronin korjaus sekä Holmin ja Hochbergin menetelmät kontrolloivat kaikki perhekohtaista virhetodennäköisyyttä, eli edes yhden hylkäysvirheen tekemistä tasolla α. Tästä syystä hyväksymisvirheiden määrä kasvaa rajusti. Toisin sanoen voi olla mahdollista, että iso osa löydöksistä jää tekemättä sen takia, että halutaan pitää perhekohtainen virhetodennäköisyys eli yhden tai useamman väärän löydöksen todennäköisyys tasolla α. Emme toki tee kovinkaan usein hylkäysvirheitä, mutta sen sijaan teemme todennäköisesti paljon hyväksymisvirheitä. Voisiko hylkäys- ja



Kuva 5: P-arvot (sininen), Bonferronin korjaus (musta), Holmin ja Hochbergin menetelmät (punainen).

hyväksymisvirheiden suhteellista osuutta jotenkin kontrolloida sen sijaan, että kontrolloitaisiin edes yhden hylkäysvirheen tekemistä ja onko siitä jotakin hyötyä?

4 Väärien löydösten osuus (FDR)

Väärien löydösten osuudella (eng. *false-discovery rate*) [12] tarkoitetaan sitä, kuinka suuri osa löydöksistä odotusarvoisesti johtaa hylkäysvirheeseen. Esimerkiksi jos olemme tehneet 20 löydöstä ja tiedämme, että viisi niistä johtaa hylkäysvirheeseen, tällöin aineistosta realisoitunut väärien löydösten osuus $FDP = 5/20 = 0.25$ (eng. *false-discovery proportion*). Mistä sitten voimme tietää mitkä löydöksistä ovat oikeita ja mitkä vääriä? Oletetaan, että olemme tehneet N eri testiä. Palautetaan mieleen, että p-arvot ovat tasaisesti jakautuneet, jos nollahypoteesit ovat oikeasti totta [5]. Tämä tarkoittaa sitä, että merkitsevyystason α alapuolella on pahimmassa tapauksessa odotusarvoisesti $\alpha \cdot N$ p-arvoa, jotka johtavat hylkäysvirheeseen.

Esimerkiksi jos olemme tehneet 1000 testiä merkitsevyystasolla $\alpha = 0.05$, odotusarvoisesti vähintään $\alpha \cdot 1000 = 50$ p-arvoa on alle merkitsevyystason ja odotusarvoisesti 50 p-arvoa johtaa hylkäysvirheeseen, jos kaikki nollahypoteesit ovat oikeasti totta.

Oletetaan nyt, että olemme tehneet 1000 testiä merkitsevyystasolla $\alpha = 0.05$ ja testien p-arvoista 120 on alle merkitsevyystason. Jos odotusarvoisesti näistä 120:stä p-arvosta $\alpha \cdot 1000 = 50$ johtaa hylkäysvirheeseen, tällöin jäljelle jäävät 70 p-arvoa ovat odotusarvoisesti oikeita löydöksiä. Väärien löydösten osuuden kaava on yksinkertaisesti:

$$FDR = E[FDP] = E\left[\frac{\text{väärät löydökset}}{\text{väärät löydökset + oikeat löydökset}}\right]$$

Pyrkimyksenä on pitää väärien löydösten osuus alle jonkin valitun tason q, esimerkiksi $FDR \leq q = 0.10$, jolloin vääriä löydöksiä olisi 10 % ja oikeita löydöksiä 90 % kaikista löydöksistä.

Väärien löydösten osuutta on helppo kontrolloida, kun pidetään mielessä, että merkitsevyystason α alapuolelle jää aina odotusarvoisesti vähintään $\alpha \cdot N$ p-arvoa. Voimme siis valita jonkin pisteen c väliltä $(0, 1)$, tällöin välillä $(0, c)$ on odotusarvoisesti vähintään $c \cdot N$ p-arvoa ja jos väliltä $(0, c)$ löytyy esimerkiksi d p-arvoa ja $c \cdot N < d$, tällöin odotusarvoisesti $c \cdot N$ p-arvoa johtaa hylkäysvirheeseen ja $d - c \cdot N$ p-arvoa ovat oikeita löydöksiä. Jos $d \leq c \cdot N$, niin odotusarvoisesti kaikki löydökset ovat vääriä.

Oletetaan, että olemme tehneet 1000 testiä ja saaneet 1000 p-arvoa ja oletetaan myös, että nyt jotkut nollahypoteesit eivät välttämättä ole totta. Jaetaan aluksi väli $(0, 1)$ tuhanteen yhtäsuureen osaan, eli asetetaan $c = 1/1000 = 0.001$ ja tarkastellaan seuraavia välejä $(0, c), (0, 2c), \ldots, (0, 10c)$. Ensimmäiseltä väliltä $(0, c)$ löytyy 20 p-arvoa, joista odotusarvoisesti $c \cdot 1000 = 1$ on väärä löydös ja $20 - 1 = 19$ ovat oikeita löydöksiä. Tarkastellaan sitten väliä $(0, 2c)$, josta löytyy yhteensä 45 p-arvoa, näistä p-arvoista odotusarvoisesti $2c \cdot 1000 = 2$ on vääriä löydöksiä ja $45 - 2 = 43$ on oikeita löydöksiä. Taulukossa 1 käydään läpi välit $(0, c), (0, 2c), ..., (0, 10c)$ ja lasketaan FDR. Taulukosta 1 voi huomata, että väärien löydösten osuutta pystyy

Väli	E[väärät löydökset]	E[oikeat löydökset]	FDR
$(0, c)$	1	19	0.050
$(0, 2c)$	2	43	0.044
$(0, 3c)$	3	57	0.050
$(0, 4c)$	4	71	0.057
$(0, 5c)$	5	72	0.065
$(0, 6c)$	6	71	0.078
$(0, 7c)$	7	74	0.086
$(0, 8c)$	8	75	0.096
$(0, 9c)$	9	74	0.108
$(0, 10c)$	10	76	0.116

Taulukko 1: Esimerkki väärien löydösten osuuden kontrolloimisesta.

kontrolloimaan valitsemalla eri pituisia välejä, esimerkiksi jos haluamme toteuttaa epäyhtälön $FDR \leq q = 0.10$, voimme valita välin $(0, 8c)$, jolloin

$$FDR = \frac{8}{8 + 75} = 0.096 \leq q = 0.10$$

Tämä on ihan toimiva tapa, mutta väärien löydösten osuuden kontrolloimiseen on keksitty parempiakin menetelmiä.

4.1 Benjaminin ja Hochbergin menetelmä

Palautetaan hetkeksi mieleen Hochbergin menetelmä 3.4, jossa kontrolloitiin perhekohtaista virhetodennäköisyyttä (FWER) tasolla α. Testien p-arvot järjestettiin pienimmästä suurimpaan $p_{(1)}, ..., p_{(N)}$ ja nollahypoteesit samaan järjestykseen $H_{0(1)}, ..., H_{0(N)}$. Tämän jälkeen etsittiin suurin p-arvo, joka toteuttaa epäyhtälön

$$p_{(i)} \leq \frac{\alpha}{N - i + 1}$$

ja hylättiin kaikki nollahypoteesit $H_{0(1)}, ..., H_{0(i)}$. Benjaminin ja Hochbergin menetelmä ([1, s. 276]) [12] toimii samalla tavalla, mutta nyt kontrolloidaan väärien löydösten osuutta tasolla q ja etsitään suurinta p-arvoa, joka toteuttaa epäyhtälön

$$p_{(i)} \leq \frac{i}{N} q \tag{5}$$

Tässäkin menetelmässä hylätään kaikki nollahypoteesit $H_{0(1)}, ..., H_{0(i)}$, mutta tällä kertaa teemme odotusarvoisesti $q \cdot i$ hylkäysvirhettä.

Esimerkiksi jos olemme tehneet $N = 1000$ erilaista testiä, asettaneet tason $q = 0.10$ ja suurin p-arvo, joka toteuttaa epäyhtälön 5 on $p_{(30)}$, tällöin hylkäämme nollahypoteesit $H_{0(1)}, ..., H_{0(30)}$, joista hylkäysvirheitä on odotusarvoisesti $q \cdot 30 = 3$. Benjaminin ja Hochbergin menetelmän todistus on esitetty lähteissä [2] ja [3]. Valitettavasti Hochbergin menetelmän tavoin Benjaminin ja Hochbergin menetelmäkään ei toimi niissä tapauksissa, joissa p-arvot ovat negatiivisesti korreloituneita. On kuitenkin olemassa menetelmä, joka toimii kaikissa tapauksissa, ja se esitellään seuraavaksi.

N	$c(N)$
100	5.19
1000	7.49
10000	9.79
100000	12.09
1000000	14.39

Taulukko 2: $c(N) = \sum_{i=1}^{N} \frac{1}{i}$

4.2 Benjaminin ja Yekutielin menetelmä

Benjaminin ja Yekutielin menetelmässä [12] p-arvot järjestetään pienimmästä suurimpaan $p_{(1)}, ..., p_{(N)}$ ja niiden nollahypoteesit samaan järjestykseen $H_{0(1)}, ..., H_{0(N)}$. Tämän jälkeen etsitään suurin p-arvo, joka toteuttaa epäyhtälön

$$p_{(i)} \leq \frac{i}{N \cdot c(N)} q,$$

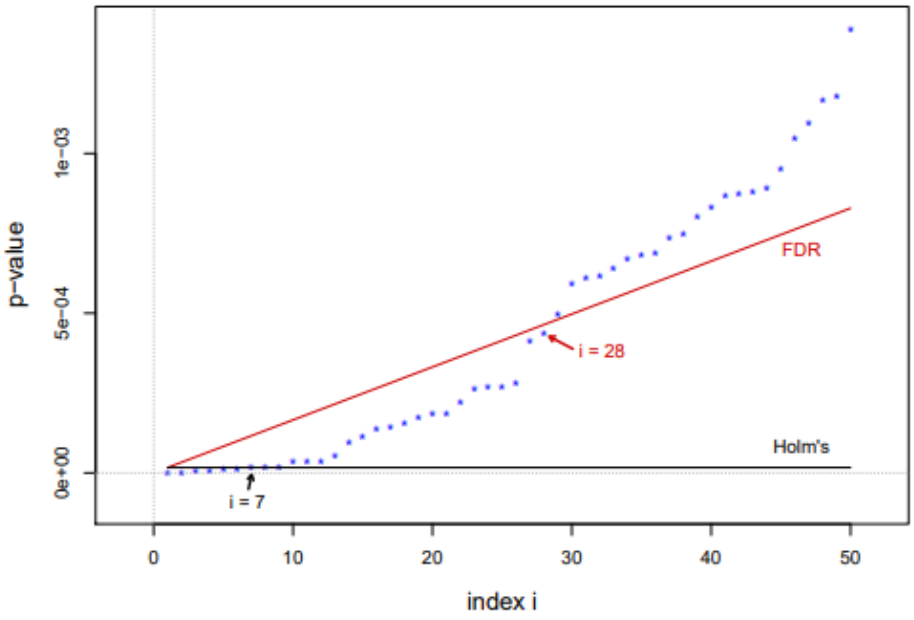
jossa $c(N) = \sum_{i=1}^{N} \frac{1}{i}$ ja hylätään kaikki nollahypoteesit $H_{0(1)}, ..., H_{0(i)}$. Benjaminin ja Yekutielin menetelmä toimii siis täysin samalla tavalla kuin Benjaminin ja Hochbergin menetelmä, mutta epäyhtälön 5 oikean puolen jakajaan on lisätty kerroin $c(N)$. Benjaminin ja Yekutielin menetelmä toimii siis kaikissa tapauksissa riippumatta siitä, millä tavalla p-arvot ovat korreloituneita. Kertoimen $c(N)$ lisäämisellä on kuitenkin hintansa, mikä nähdään myöhemmin. Taulukkoon 2 on koottu joitain kertoimen $c(N)$ arvoja.

Benjaminin ja Yekutielin menetelmän todistus on esitetty lähteessä [4, s. 6-7].

5 Menetelmien vertailua

Verrataan nyt hieman edellä mainittuja menetelmiä toisiinsa. Ainoa oikean elämän esimerkki, jonka löysin on Bradley Efronin ja Trevor Hastien kirjasta [1, s. 271-277]. Tutkimuksessa oli 102 miestä, joista 52 oli eturauhassyöpäpotilaita ja 50 tervettä verrokkia. Tutkimus pyrki löytämään poikkeamia eri geenien ($N = 6033$) aktiivisuustasoissa miehillä. Tutkimuksen p-arvot saatiin tekemällä kahden otoksen t-testejä vertaamalla eturauhassyöpäpotilaiden geenien aktiivisuustasoja normaalien kontrollien aktiivisuustasoihin. Holmin menetelmällä tehtiin seitsemän löydöstä tasolla $\alpha = 0.10$ ja Benjaminin ja Hochbergin menetelmällä tehtiin 28 löydöstä tasolla $q = 0.10$ (kuva 6).

Muistetaan, että Holmin menetelmällä kontrolloidaan perhekohtaista virhetodennäköisyyttä eli edes yhden hylkäysvirheen tekemistä tasolla α, kun taas Benjaminin ja Hochbergin menetelmällä kontrolloidaan väärien löydösten osuutta tasolla q. Tässä tapauksessa Benjaminin ja Hochbergin menetelmä tekee siis odotusarvoisesti $q \cdot 28 = 2.8$ hylkäysvirhettä.



Kuva 6: Esimerkki Bradley Efronin ja Trevor Hastien kirjasta.

5.1 Keksittyjä esimerkkejä

Oikean elämän esimerkkejä oli vaikea löytää, joten generoin itse p-arvoja ottamalla satunnaisia otoksia ($N = 10000$) erilaisista beta-jakaumista. Näissä beta-jakaumissa parametri β on aina 1 ja parametri α vaihtelee välillä $(0, 1)$, koska kyseiset beta-jakaumat näyttävät jäljittelevän kohtalaisen hyvin p-arvojen käyttäytymistä. Esimerkiksi $Beta(1, 1)$ on sama kuin tasajakauma ja kuvaa p-arvojen jakaumaa tilanteessa, jossa kaikki nollahypoteesit $H_{01}, ..., H_{0N}$ ovat oikeasti totta, kun beta-jakauman parametrin α arvoa lasketaan alle yhden, beta-jakauman vasen häntä alkaa kasvaa ja oikea häntä pienentyä ja tämä taas kuvaa tilannetta, jossa osa nollahypoteeseista $H_{01}, ..., H_{0N}$ ei ole totta. Beta-jakauman käyttäytymistä voi käydä kokeilemassa esimerkiksi geogebrassa [8].

Menetelmiä vertailtiin siis siten, että erilaisista beta-jakaumista generoitiin 10000 keinotekoista p-arvoa ja katsottiin kuinka monta löydöstä kukin menetelmä tekee. Tämä toistettiin 10000 kertaa, jotta voitiin laskea kuinka monta löydöstä menetelmät tekivät keskimäärin. Tulokset ovat taulukossa 3.

Taulukosta 3 nähdään, että Bonferronin korjaus tekee odotetusti vähiten löydöksiä kaikista menetelmistä, mutta ero Holmin ja Hochbergin menetelmiin ei ole suuri. Tästä syystä on ehkä suositeltavampaa käyttää Bonferronin korjausta sen helppouden takia, Holmin ja Hochbergin menetelmien sijaan. Muistetaan kuitenkin, että Holmin ja Hochbergin menetelmät kontrolloivat perhekohtaista virhetodennäköisyyttä samalla tasolla α kuin Bonferronin korjaus, mutta tekevät aina vähintään saman verran löydöksiä kuin Bonferronin korjaus ja joissain tapauksissa enemmän.

Holmin ja Hochbering menetelmät näyttävät käytännössä tekevän identtisen määrän löydöksiä, taulukkoon 3 on merkitty tähdellä (*) kohdat, joissa Hochbergin menetelmä teki marginaalisesti enemmän löydöksiä kuin Holmin menetelmä. Jakaumasta $Beta(0.2, 1)$ generoiduista p-arvoista Hochbergin menetelmä teki 0.00008 %

Jakauma	Bonf	Holm	Hoch	B-H	B-Y
$Beta(1.0, 1)$	0.100	0.100	0.100	0.124	0.010
$Beta(0.9, 1)$	0.316	0.316	0.316	0.551	0.038
$Beta(0.8, 1)$	0.997	0.997	0.997	4.182	0.201
$Beta(0.7, 1)$	3.132	3.133	3.133	50.40	1.202
$Beta(0.6, 1)$	10.03	10.03	10.03	319.0	12.04
$Beta(0.5, 1)$	31.65	31.69	31.69	1001	103.0
$Beta(0.4, 1)$	99.96	100.4	100.4	2155	471.1
$Beta(0.3, 1)$	316.3	319.4	319.4	3728	1402
$Beta(0.2, 1)$	1001	1022	1022*	5624	3180
$Beta(0.1, 1)$	3162	3291	3291*	7743	6009

Taulukko 3: Kyseisistä beta-jakaumista on generoitu 10000 keinotekoista p-arvoa ja laskettu kuinka monta löydöstä kukin menetelmä keskimäärin tekee tasolla $\alpha = q = 0.10$.

enemmän löydöksiä kuin Holmin menetelmä ja jakaumasta $Beta(0.1, 1)$ Hochbergin menetelmä teki 0.0001 % enemmän löydöksiä kuin Holmin menetelmä. Käytännössä ei siis minkäänlaista eroa Holmin ja Hochbergin menetelmien välillä. Muistetaan vielä, että Holmin menetelmä toimii kaikissa tapauksissa p-arvojen korreloituneisuudesta huolimatta, kun taas Hochbergin menetelmä ei toimi niissä tapauksissa, joissa p-arvot ovat negatiivisesti korreloituneita. Näin ollen jos halutaan kontrolloida perhekohtaista virhetodennäiköisyyttä kannattaa ensisijaisesti käyttää Bonferronin korjausta sen helppouden takia, mutta jos halutaan tehdä enemmän löydöksiä mitä Bonferronin korjaus teki voi kokeilla Holmin menetelmää. Taulukon 3 simulaation perusteella Hochbergin menetelmää ei kannata ikinä käyttää, ellei ole varmaa tietoa siitä, että p-arvot eivät ole negatiivisesti korreloituneita.

Mielenkiintoisinta ainakin itselleni oli verrata perhekohtaisen virhetodennäköisyyden ja väärien löydösten osuuden menetelmiä toisiinsa. Harmillisesti Benjaminin ja Yekutielin menetelmä (B-Y) ei näytä toimivan kovinkaan hyvin, kun perhekohtaista virhetodennäköisyyttä kontrolloivat menetelmät ovat tehneet vähemmän kuin 10 löydöstä. Tämä on harmi siitä syystä, että Benjaminin ja Yekutielin menetelmä toimii kaikissa tapauksissa p-arvojen korreloituneisuudesta riippumatta. Sen sijaan Benjaminin ja Hochbergin menetelmän (B-H) hyödyllisyys tulee esiin taulukon 3 simulaatiossa jo, kun perhekohtaista virhetodennäköisyyttä kontrolloivat menetelmät ovat tehneet keskimäärin 0.997 löydöstä tasolla $\alpha = 0.10$. Tällöin Benjaminin ja Hochbergin menetelmä tekee keskimäärin 4.182 löydöstä tasolla $q = 0.10$, joista keskimäärin $4.182 \cdot q = 0.418$ löydöstä johtaa hylkäysvirheeseen. Taulukon 3 seuraavilla kahdella rivillä Benjaminin ja Hochbergin menetelmän hyödyllisyys korostuu entisestään, kun perhekohtaista virhetodennäköisyyttä kontrolloivat menetelmät tekevät keskimäärin 3.132-3.133 ja 10.03 löydöstä, Benjaminin ja Hochbergin menetelmä tekee keskimäärin 50.40 ja 319.0 löydöstä, joista keskimäärin $50.40 \cdot q = 5.04$ ja $319.0 \cdot q = 31.9$ löydöstä johtaa hylkäysvirheeseen. Muistetaan, että Benjaminin ja Hochbergin menetelmä ei toimi niissä tilanteissa, joissa p-arvot ovat negatiivisesti korreloituneita. Mainitaan kuitenkin, että taustatietoa hakiessani en löytänyt yhtäkään oikean elämän esimerkkiä, missä olisi näytetty, että Benjaminin ja Hochber-

gin menetelmä ei olisi toiminut. Yhdessäkään testissä ei myöskään koskaan käytetty Benjaminin ja Yekutielin menetelmää eikä Hochbergin menetelmää, vaan aina Bonferronin korjausta, Holmin menetelmää tai Benjaminin ja Hochbergin menetelmää.

Lopuksi voidaan vielä miettiä milloin kutakin menetelmää kannattaa käyttää. Perhekohtaista virhetodennäköisyyttä kontrolloivista menetelmistä Holmin menetelmä on paras, koska se toimii kaikissa tapauksissa p-arvojen korreloituneisuudesta riippumatta ja menetelmällä tehdään aina vähintään yhtä monta löydöstä kuin Bonferronin korjauksella, mutta koska ero ei ole kovin suuri ainakaan taulukon 3 simulaation perusteella, Bonferronin korjaus on ehkä suositeltavampi menetelmä sen helppouden vuoksi. Hochbergin menetelmällä voidaan tehdä enemmän löydöksiä kuin Holmin menetelmällä, mutta taulukon 3 simulaation perusteella eroa ei käytännössä ole ja Hochbergin menetelmä ei toimi niissä tilanteissa, joissa p-arvot ovat negatiivisesti korreloituneita, näistä syistä Hochbering menetelmää ei kannata luultavasti ikinä käyttää. Toki jos tiedetään varmasti, että p-arvot eivät ole negatiivisesti korreloituneita, voihan menetelmää kokeilla siinä toivossa, että se tekee enemmän löydöksiä kuin Holmin menetelmä.

Väärien löydösten osuutta kontrolloivia menetelmiä kannattaa käyttää silloin, kun perhekohtaista virhetodennäköisyyttä kontrolloivat menetelmät eivät ole tehneet tarpeeksi löydöksiä. Jos tiedetään, että p-arvot eivät ole negatiivisesti korreloituneita, siinä tapauksessa kannattaa aina käyttää Benjaminin ja Hochbergin menetelmää, koska se tekee aina vähintään yhtä monta löydöstä kuin Benjaminin ja Yekutielin menetelmä, mutta ainakin taulukon 3 simulaation perusteella Benjaminin ja Hochbergin menetelmä tekee suhteessa paljon enemmän löydöksiä kuin Benjaminin ja Yekutielin menetelmä. Jos taas ei ole tietoa p-arvojen korreloituneisuudesta on pakko käyttää Benjaminin ja Yekutielin menetelmää. Taulukosta 3 nähdään, että voi käydä niinkin, että Benjaminin ja Yekutielin menetelmä tekee itseasiassa vähemmän löydöksiä kuin mitä perhekohtaista virhetodennäköisyyttä kontrolloivilla menetelmillä on tehty. Tällöin tietysti valitaan ne löydökset, mitkä perhekohtaista virhetodennäköisyyttä kontrolloivat menetelmät tekivät.

Viitteet

[1] Bradley Efron, Trevor Hastie: *Computer age statistical inference*, Cambridge University Press, 2016, verkko-osoite:

https://hastie.su.domains/CASI_files/PDF/casi.pdf

[2] Xiaodong Li, *Multiple Testing*, verkko-osoite:

https://www.stat.ucdavis.edu/~xdgli/Xiaodong_Li_Teaching_files/multiple_testing.pdf

[3] Spätzle, käyttäjäprofiili: (https://stats.stackexchange.com/users/144600/sp%c3%a4tzle), Todistus väärien löydösten osuudelle Benjaminin ja Hochbergin menetelmässä, verkko-osoite (versio: 25.11.2021):

https://stats.stackexchange.com/q/553535

[4] Ruodu Wang, *Elementary proofs of several results on false discovery rate*, 27.7.2022, verkko-osoite:

https://arxiv.org/pdf/2201.09350

[5] jII, käyttäjäprofiili: (https://stats.stackexchange.com/users/49690/jii), Miksi totta olevien nollahypoteesien p-arvot ovat tasaisesti jakautuneet?, verkko-osoite (versio: 11.5.2018):

https://stats.stackexchange.com/q/345763

[6] Sanat K. Sarkar: *Generalizing Simes' test and Hochberg's stepup procedure*, The Annals of Statistics, Ann. Statist. 36(1), 2008, 337–363. verkko-osoite:

https://doi.org/10.1214/009053607000000550

https://arxiv.org/pdf/0803.1961

[7] Yosef Hochberg: *A sharper Bonferroni procedure for multiple tests of significance*, Biometrika, lehtinumero 75, neljäs julkaisu, joulukuu 1988, sivut 800–802, verkko-osoite:

https://doi.org/10.1093/biomet/75.4.800

http://www-stat.wharton.upenn.edu/~steele/Courses/956/Resource/MultipleComparision/Hochberg88.pdf

[8] https://www.geogebra.org/calculator

[9] "Family-wise error rate." *Wikipedia*, haettu 17.2.2025, verkko-osoite:

https://en.wikipedia.org/wiki/Family-wise_error_rate

[10] "Bonferroni correction." *Wikipedia*, haettu 4.3.2025, verkko-osoite:

https://en.wikipedia.org/wiki/Bonferroni_correction

[11] "Holm–Bonferroni method." *Wikipedia*, haettu 17.3.2025, verkko-osoite:

https://en.wikipedia.org/wiki/Holm%E2%80%93Bonferroni_method

[12] "False discovery rate." *Wikipedia*, haettu 21.2.2025, verkko-osoite:

https://en.wikipedia.org/wiki/False_discovery_rate